

Utilização do algoritmo Apriori para traçar o perfil sociodemográfico do homem brasileiro com câncer de próstata

Use of the Apriori algorithm to trace the sociodemographic profile of Brazilian men with prostate cancer

Uso del algoritmo Apriori para trazar el perfil sociodemográfico de hombres brasileños con cáncer de próstata

Recebido: 13/04/2022 | Revisado: 21/04/2022 | Aceito: 27/04/2022 | Publicado: 29/04/2022

Gustavo Dias da Silva

ORCID: <https://orcid.org/0000-0001-5752-6693>

Universidade Estadual da Paraíba, Brasil

E-mail: gustavouepb@gmail.com

Wellington Candeia de Araújo

ORCID: <https://orcid.org/0000-0003-2102-7993>

Universidade Estadual da Paraíba, Brasil

E-mail: wcandeia@uepb.edu.br

Resumo

Entre as doenças que acometem a população masculina o câncer de próstata tem aumentado a taxa de mortalidade entre eles, onde no mundo é a sexta neoplasia maligna e no Brasil a primeira. Apesar das iniciativas de ajuda a população masculina contra a neoplasia prostática, ainda falta um direcionamento quanto ao diagnóstico e tratamento. Mas as iniciativas seriam mais bem direcionadas se tivessem os perfis dos pacientes assistidos por elas, porém ainda é um campo de pesquisa com lacunas. Além disso, dados que possam ajudar se encontram armazenados em grandes bases de dados com muitas informações, principalmente devido ao processo de informatização do setor de saúde, que dificulta uma análise manual desses dados. Este trabalho tem o objetivo de determinar o perfil sociodemográfico do brasileiro com o câncer de próstata por meio do algoritmo Apriori com dados de 2010 a 2019. Com isso, aplicamos na base de dados do INCA o algoritmo Apriori com a finalidade de termos as regras de associação. Ao final percebe-se que os fatores de tabagismo, alcoolismo, raça e estado conjugal são os fatores que mais destacaram por aparecerem nas regras com os maiores índices de confiança. Porém, depreendemos que a raça parda é de maior incidência do câncer de próstata no Brasil. Apesar da incompletude dos dados opcionais na base do INCA, destaca-se a análise realizada a nível nacional e a possibilidade de utilização para nortear campanhas no contexto da saúde do homem.

Palavras-chave: Câncer de próstata; Mineração de dados; Regras de associação; Algoritmo apriori; Ensino em saúde.

Abstract

Among the diseases that affect the male population, prostate cancer has increased the mortality rate among them, where it is the sixth malignant neoplasm in the world and in Brazil the first. Despite the initiatives to help the male population against prostate cancer, there is still a lack of guidance regarding diagnosis and treatment. However, the initiatives would be better targeted if they had the profiles of patients assisted by them, but it is still a field of research with gaps. In addition, data that can help are stored in large databases with a lot of information, mainly due to the computerization process of the health sector, which makes manual analysis of this data difficult. This work aims to determine the sociodemographic profile of Brazilians with prostate cancer through the Apriori algorithm with data from 2010 to 2019. With this, we applied the Apriori algorithm to the INCA database in order to have the rules of Association. In the end, it is clear that the factors of smoking, alcoholism, race and marital status are the factors that stood out the most as they appear in the rules with the highest levels of confidence. However, we infer that the brown race has a higher incidence of prostate cancer in Brazil. Despite the incompleteness of the optional data in the INCA database, the analysis carried out at the national level and the possibility of using it to guide campaigns in the context of men's health stands.

Keywords: Prostate cancer; Data mining; Association rules; Apriori algorithm; Teaching in health.

Resumen

Entre las enfermedades que afectan a la población masculina, el cáncer de próstata ha aumentado la tasa de mortalidad entre ellos, donde es la sexta neoplasia maligna en el mundo y en Brasil la primera. A pesar de las iniciativas para ayudar a la población masculina contra el cáncer de próstata, aún falta orientación en cuanto al diagnóstico y tratamiento. Pero las iniciativas estarían mejor dirigidas si tuvieran los perfiles de los pacientes atendidos por ellas, pero aún es un campo de investigación con lagunas. Además, los datos que pueden ayudar se almacenan en grandes bases de datos con mucha información, principalmente debido al proceso de informatización del sector salud, lo que dificulta el análisis manual

de estos datos. Este trabajo tiene como objetivo determinar el perfil sociodemográfico de los brasileños con cáncer de próstata a través del algoritmo Apriori con datos de 2010 a 2019. Con eso, aplicamos el algoritmo Apriori a la base de datos INCA para tener las reglas de Asociación. Al final, se puede apreciar que los factores tabaquismo, alcoholismo, raza y estado civil son los factores que más se destacaron por aparecer en las reglas con mayores niveles de confianza. Sin embargo, inferimos que la raza parda tiene mayor incidencia de cáncer de próstata en Brasil. A pesar de lo incompleto de los datos opcionales en la base de datos del INCA, se destaca el análisis realizado a nivel nacional y la posibilidad de utilizarlo para orientar campañas en el contexto de la salud del hombre.

Palabras clave: Cáncer de próstata; Procesamiento de datos; Reglas de asociación; Algoritmo apriori; Enseñanza en salud.

1. Introdução

A saúde é um bem que todos têm direito e a Carta Magna Brasileira garante isso. Porém, muitas vezes, as políticas assistencialistas que objetivam levar a uma melhor qualidade de vida não chegam às classes menos favorecidas de conhecimento e recursos financeiros, precisando inúmeras vezes de que o governo procure meios alternativos para isso. Contudo, o pouco conhecimento da população pode dificultar a difusão da assistência. Além disso, os tratamentos e diagnósticos de doenças ficam prejudicados quando o cenário do sistema de saúde é desigual em se tratando do acesso (Sacramento et al., 2019).

Dentre tantas doenças que acomete a população, destaca-se o câncer, já que esse atinge o indivíduo tanto em termos físicos quanto em termos psicológicos. E os números sobre ele tendem a aumentos consideráveis nos anos seguintes (Aguiar et al., 2017; Ferrão et al., 2017; Bettinelli & Portela, 2017; Santos, 2018).

Ao considerar a população masculina, o câncer de próstata tem se apresentado como destaque tanto a nível mundial quanto brasileiro, haja vista que ele tem um índice de mortalidade alto, sendo considerado o segundo câncer que mais mata entre os homens, ficando atrás apenas do câncer de pulmão (Mota & Barros, 2018; Rego et al., 2020).

Apesar do alto índice, no Brasil há um preconceito na realização de exames para diagnóstico, devido às questões culturais. Segundo Menezes et al. (2019), a questão cultural barra o homem na procura da sua saúde, dificultando assim a identificação de pacientes com este tipo de câncer. Além disso, o Ministério da Saúde (doravante MS) brasileiro não recomenda o rastreamento populacional para o câncer de próstata, pois há riscos e limitações dos exames (Rego et al., 2020).

No caso do câncer de próstata, de acordo com Mota e Barros (2019), é de suma importância a definição do perfil do paciente. Essa definição do perfil do paciente que possui o câncer pode facilitar o diagnóstico, o encaminhamento de exames e o tratamento adequado. Ademais, segundo Araújo et al. (2015), o campo da pesquisa que define o perfil dos pacientes ainda precisa ser desenvolvido.

Atualmente as informações que podem ser utilizadas para formar o perfil dos pacientes estão sendo armazenada em base de dados. E as novas tecnologias têm produzido grande quantidade de dados nas mais diversas áreas, no campo da saúde, não é diferente (Oliveira, et al., 2007). Porém, se for necessária uma análise das informações armazenadas, de forma manual, é impossível, pois esbarra no cálculo e associação de informação de grandes quantidades de dados (Preissler, 2016).

No campo brasileiro, ainda há muitas restrições e poucas ações. E, quando ocorrem ações, há o destaque do uso da estatística descritiva, sendo impossível aplicar tal tática em uma base com milhares de dados para serem levantados e analisados (Araújo et al. 2015; Mota & Barros, 2018; Rego et al., 2020).

As táticas presentes na mineração de dados, a associação, será a tarefa adotada por facilitar o entendimento dos dados presentes em grandes bases de dados, pois conforme Silva et al. (2016). Portanto, este trabalho tem o objetivo de determinar o perfil sociodemográfico do brasileiro com o câncer de próstata por meio do algoritmo Apriori aplicado na base de dados do Instituto Nacional do Câncer José Alencar Gomes da Silva (INCA) com dados de 2010 a 2019.

2. Metodologia

Nessa seção, encontra-se todos os procedimentos para a realização do trabalho, a fim de tornar possível sua reprodução, como sugere o método científico, pois, conforme Wazlawick (2009), o não seguimento do método científico pode levar a conclusões erradas.

Esta pesquisa se caracterizará de acordo com os critérios definidos por Prodanov e Freitas (2013). Desse modo, ela pode ser considerada como pesquisa de natureza aplicada, de um método científico dedutivo, objetivos descritivos, um procedimento de um estudo de caso e uma abordagem quantitativa.

A caracterização da metodologia da pesquisa permite a sua reprodução em meio acadêmico. Assim, pode-se dizer que a pesquisa é aplicada em sua natureza, por tentar definir um perfil específico do homem com o câncer de próstata.

Quanto ao método científico que será utilizado, ela se apresenta, de forma dedutiva, uma vez que será a partir da realidade nacional brasileira da população masculina que será construído o perfil sociodemográfico dos homens com câncer de próstata.

Apresenta um objetivo descritivo, haja vista a análise dos dados presentes nas bases de dados do INCA, para expor as características dos indivíduos com câncer de próstata. Isto será possível através da técnica de associação da mineração de dados, pois serão criadas várias regras de associação a partir do algoritmo Apriori, que coletará os dados dessa base de dados.

No caso do procedimento adotado, a presente pesquisa é caracterizada como estudo de caso, pois o propósito é analisar uma situação específica da população masculina brasileira para a formação de um perfil, que descreva uma parte dela que esteja propensa ao câncer de próstata.

E por fim a sua abordagem é definida como quantitativa, pois os dados serão analisados considerando sua frequência relativa e a confiabilidade das regras de associação encontradas nos dados da base de dados do INCA com as características dos homens com câncer de próstata.

Porém, antes de qualquer coisa para o embasamento da pesquisa em primeiro plano foi realizada uma revisão sistemática sobre o tema do presente trabalho, com o propósito de verificar o estado da arte.

Nas seções a seguir, é detalhada a descrição de todo o procedimento metodológico adotado nesta pesquisa com os seus detalhes.

2.1 Fonte de Dados

A primeira etapa do processo de adquirir conhecimento em base de dados é a aquisição dos dados que serão analisados durante todo o processo. Para o estudo que se segue foi adquirida a base de dados do INCA.

Essa base de dados surge dos registros hospitalares do câncer que devido à sua importância, o Ministério da Saúde, em 1998, publicou a Portaria Nº 3535/GM que estabeleceu requisitos para cadastramento de centros de atendimentos em oncologia, e tornou obrigatório o funcionamento do Registro Hospitalar de Câncer (RHC) (Ministério da Saúde, 1998).

Além disso, em 2005, o Ministério da Saúde, com o propósito de organizar a rede atenção oncológica, instituiu a Política Nacional de Atenção Oncológica, que tornou obrigatório os registros de câncer e define parâmetros para sua implantação e funcionamento.

Ademais, a Portaria MS/SAS Nº 741 de 2005, em seu parágrafo único, determina que “Arquivos eletrônicos dos dados anuais consolidados deverão, no mês de setembro de cada ano, a partir de 2007, ser encaminhados para o INCA, que deverá publicá-los e divulgá-los de forma organizada e analítica” (Ministério da Saúde, 2005). Isso faz do INCA o concentrador de das informações dos tratamentos de câncer do país.

Portanto, é do INCA o local onde estão os dados de todos os cânceres e estão disponíveis para serem consultados. A aquisição dos dados para a análise foi possível pelo portal do Integrador RHC, Figura 1, que está disponível no sítio

<https://irhc.inca.gov.br/RHCNet/>, e através da opção tabular dados que abrirá uma janela onde se encontra a opção de *download*. Logo em seguida, temos a opção de fazer o *download* dos dados através de três modos: Base de dados de todos os estados, exceto SP; Base de dados de todos os estados; e Base de dados do estado.

Figura 1 - Portal integrador RHC.

Fonte: Autores (2021).

Para este estudo, foi realizado o *download* da base de dados de todos os estados, haja vista necessidade para formação do perfil nacional. Ao selecionar essa opção de *download*, há o encaminhamento para terceira janela, ver Figura 2, em que foi selecionado o período de 2010 a 2019, por ser a última década disponível para *download* no sistema.

Figura 2 - Download de dados do portal integrador RHC.

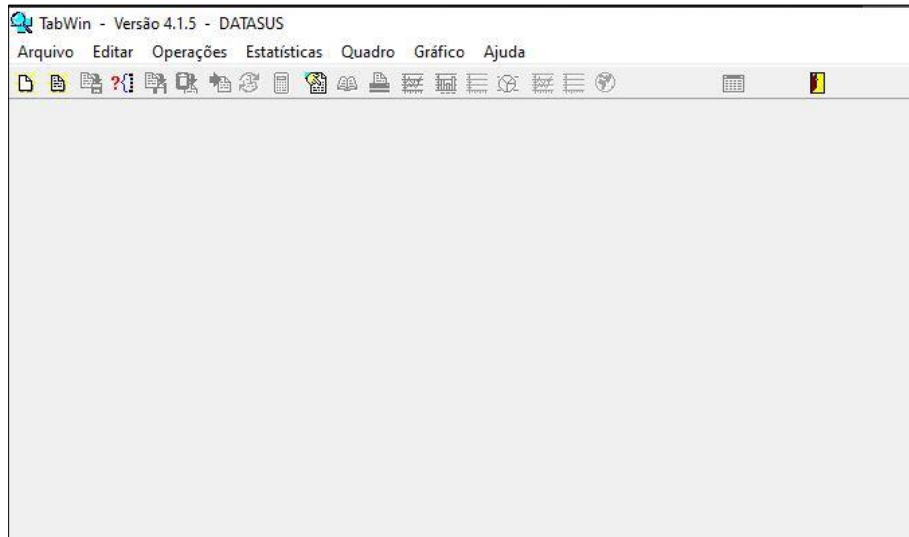
Fonte: Autores (2021).

Além dos dados em si, para realizar o estudo, foi necessário o *download* do dicionário da base de dados, onde se encontra disponível o significado de cada código que é salvo na base de dados do portal do Integrador RHC. Ele pode ser acessado no mesmo portal.

2.2 Recursos de Softwares

Para analisar os dados, foram utilizadas as ferramentas: *Tab* para *Windows* (*TabWin*) e *Spyder*. O *TabWin*, Figura 3, é o sistema para facilitar a tabulação e tratar os dados gerados pelo aplicativo TABNET¹, que foi desenvolvido pelo DATASUS. Ele se encontra disponível para *download* e com o manual de instalação no sítio do portal da saúde do DATASUS. Além disso, o *TabWin* permite operações aritméticas e estatísticas nos dados da tabela gerada ou importada, elaborar gráficos e mapas, e efetuar operações nas tabelas importadas, por exemplo, exportar para arquivos como *Comma-separated values* (CSV).

Figura 3 – TabWin.



Fonte: Autores (2021).

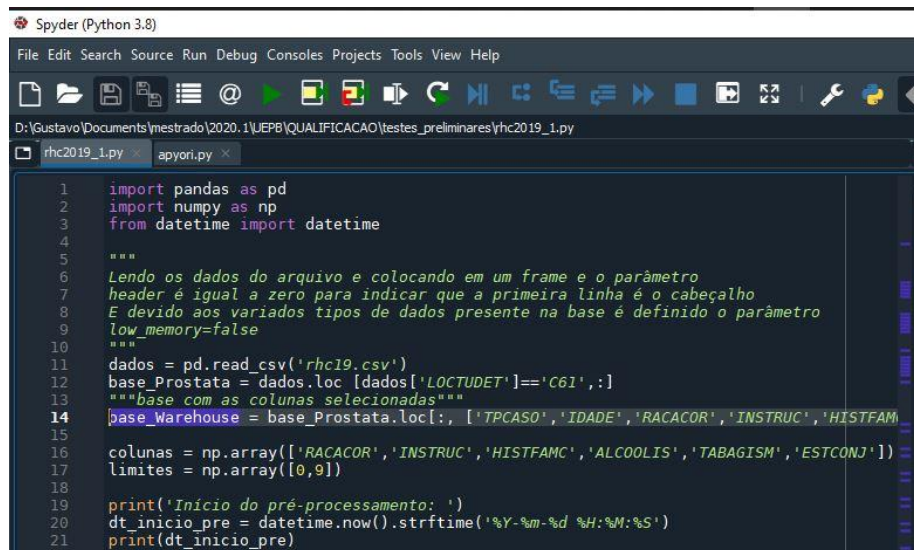
Após o *download* do arquivo exportado pelo portal Integrado do RHC, foram verificados que os arquivos com os dados de cada ano vinham na extensão *dbf*, que, para ser trabalhado na linguagem *Python*, foram convertidos para arquivos de extensão *csv*, com o auxílio do *TabWin* através da opção disponível no menu arquivo.

Com os arquivos convertidos para o formato *csv*, foi utilizado o *Spyder* versão 4.1.4 para análise dos dados, ver Figura 4, que segundo Sharma e Bansal (2020), essa é uma ferramenta técnica poderosa de *Python*. Ela se encontra no ambiente de desenvolvimento integrado de *Python*, chamado de *Anaconda*, ou pode ser feito o *download* no seu sítio², que é um ambiente *open source*, ou seja, *software* livre e de código aberto e está disponível no sítio www.anaconda.com/products/individual. Para o projeto, utilizamos a versão para *Windows*.

¹ Tabnet: Tabulador genérico de domínio público que permite organizar dados de forma rápida e se encontra disponível no sítio: <https://datasus.saude.gov.br/informacoes-de-saude-tabnet/>

² <https://www.spyder-ide.org>

Figura 4 – Spyder.



```
1 import pandas as pd
2 import numpy as np
3 from datetime import datetime
4
5 """
6 Lendo os dados do arquivo e colocando em um frame e o parâmetro
7 header é igual a zero para indicar que a primeira linha é o cabeçalho
8 E devido aos variados tipos de dados presente na base é definido o parâmetro
9 low_memory=false
10 """
11 dados = pd.read_csv('rhc19.csv')
12 base_Prostata = dados.loc [dados['LOCTUDET']=='C61',:]
13 """base com as colunas selecionadas"""
14 base_Warehouse = base_Prostata.loc[:, ['TPCASO','IDADE','RACACOR','INSTRUC','HISTFAM
15
16 colunas = np.array(['RACACOR','INSTRUC','HISTFAMC','ALCOOLIS','TABAGISM','ESTCONJ'])
17 limites = np.array([0,9])
18
19 print('Inicio do pré-processamento: ')
20 dt_inicio_pre = datetime.now().strftime('%Y-%m-%d %H:%M:%S')
21 print(dt_inicio_pre)
```

Fonte: Autores (2021).

A escolha do ambiente Anaconda foi realizada por ser uma ferramenta *open source*, a qual, além de ser de código aberto, não exige pagamento de licenças, por ser um ambiente versátil, e por possuir vários pacotes pré-instalados como *numpy*, *matplotlib*, *scipy*, *seaborn* e etc (Sharma & Bansal, 2020).

2.3 Transformação dos Dados

Para possibilitar um resultado satisfatório e de confiança, a fase de preparação dos dados, é importante que seja executada a seleção, integração e limpezas dos dados (Campbell, 2021).

O processo de transformação consiste em preparar e organizar os dados para realização da mineração de dados, que pode ser realizada por meio da criação de consultas com a linguagem *Structured Query Language* (SQL), aplicações personalizadas, ou ferramentas de terceiros (Jamsa, 2021).

Para a aplicação da transformação neste projeto, primeiro foi realizado o processo de seleção dos dados, integração e limpeza, pois os dados brutos apresentavam dados de outros cânceres e com ruídos, que poderiam prejudicar a mineração de dados. Para o desenvolvimento deste projeto, houve o uso da linguagem de programação *Python*, e, com o auxílio da biblioteca *Pandas* foi possível a conversão de dados numéricos brutos da base para dados categóricos, como pode ser visto na Quadro 1 o trecho do código onde é realizada a passagem do atributo raça numérico para categórico.

Quadro 1 - Código de conversão do atributo numérico raça para categórico

```
raca = ['BRANCA', 'PRETA', 'AMARELA', 'PARDA', 'INDIGENA', 'S/I-RA']
classes_raca= [0,1,2,3,4,8,9]
column_raca = pd.cut(x=base_Warehouse['RACACOR'], bins=classes_raca,
labels=raca)
```

Fonte: Autores (2021).

2.4 Regras de Associação

Ao final do pré-processamento, os dados ficaram organizados de forma tabular que permitiu a aplicação do algoritmo Apriori, que é explicado na seção 2.4.1.

2.4.1 Definição de parâmetros do algoritmo Apriori

Para a aplicação do algoritmo Apriori na base em estudo, definimos os índices suporte, confiança e lift. O refinamento dos índices ocorreu com testes realizados em dados que representam dois anos da amostra estudada, e seguindo os procedimentos abaixo.

Para a definição do valor do suporte realizou-se o cálculo através da Equação I.

$$Suporte_{regra}(CP) = \frac{qtde(CP)_{período}}{qtde C_{período}} \quad (I)$$

Onde, temos:

- $Qtde(CP)_{período}$ para a quantidade de casos de câncer de próstata ocorrida naquele período;
- $Qtde C_{período}$ para a quantidade de casos de câncer naquele período.

Ademais, a definição do índice de confiança, devido à ausência de critérios específicos para o caso, que indicasse o valor, foram realizados testes onde seu valor modifica a partir de 0,3 até o valor de 0,8 variando em 0,1, fazendo a seleção dos resultados únicos, ou seja, removendo os repetidos, nesse intervalo foi o momento em que apareceu a maior quantidade de resultados, haja vista que, valores altos de suporte podem deixar de lado regras interessantes. Mas, para a confiança valores altos trazem a tendência de retornar as melhores regras (Baldomir, 2017).

Portanto, foram realizados testes com esses valores para filtrar os melhores resultados e assim, conseguir extrair a maior quantidade de conhecimento possível que os dados da base de dados estudada pode oferecer.

Para a definição do *lift*, após realização de testes nos dois primeiros anos, a quantidade maior de resultados se deu quando o *lift* foi definido como 1,2, para confiança mínima da regra, ou seja, para que cada regra fosse considerada forte, teria de se indicar que no mínimo ocorreria um inteiro e dois décimos do consequente da regra dados que ocorreu o antecessor da regra. Pois, de acordo Jamsa (2021), valores perto de um indicariam que a regra seria uma mera coincidência, e acima disso seria realmente uma associação.

2.5 Seleção dos Atributos

Para análise, foram obtidos os dados do período de 2010 a 2019, pois era a última década disponível no portal do Integrador RHC do INCA, até o presente estudo. Para cada ano, desta década a ferramenta de exportação do portal do Integrador do INCA criou um arquivo dbf com colunas e linhas, em que as colunas representam os atributos e as linhas os casos.

Para a seleção dos fatores que ajudassem a desenvolver o perfil sociodemográfico do homem com câncer de próstata, foi levado em consideração os trabalhos analisados da literatura e os fatores presentes na base de dados do INCA. Assim, filtrou-se, entre todas as colunas presentes na base, as seguintes variáveis: o ano de diagnóstico, data do primeiro diagnóstico, estado de residência, idade, grau de instrução, raça/cor, e tipo de caso. Esses atributos são obrigatórios, ou seja, sempre terão algum valor armazenado.

Mas, foram selecionados alguns dados opcionais, como consumo de álcool, tabagismo, estado conjugal atual e histórico familiar. Porém, segundo INCA (2010), a partir do momento em que a unidade hospitalar resolver preencher os dados opcionais, o preenchimento se tornará obrigatório para todo o registro que for cadastrado.

Como procedimento para selecionar os dados, foi consultado o dicionário da base de dados do SisRHC, atualizado e disponível para download a partir do dia 03 de março de 2020. Em seguida, foram aplicadas as etapas para o pré-processamento dos dados analisados:

1. Filtragem dos casos de câncer de próstata através do código de Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID), que nesse caso é C61, e o atributo fica na coluna LOCTUDET;
2. Utilizando a biblioteca Pandas do Python, foram selecionadas as colunas: TPCASO, IDADE, RACACOR, INSTRUC, HISTFAM, ALCOOLIS, TABAGISM, ESTADRES, ANOPRIDI, ESTCONJ, DTDIAGNO;
3. Seleção e remoção de registros com raça, escolaridade, histórico familiar, alcoolismo e tabagismo fora das opções 0 e 1, de acordo com o dicionário da base;
4. Seleção e remoção de registros com idades fora do intervalo de 0 a 150 anos;
5. Seleção e remoção de registros com ano de diagnóstico fora do intervalo de 1900 a 2021;
6. Conversão dos campos numéricos (RACACOR, ALCOOLIS, ESTCONJ, HISTFAM, INSTRUC, TABAGISM e TPCASO) para campos categóricos.

Com os dados selecionados para o próximo passo, foi desenvolvido uma aplicação na linguagem *Python* e aplicado o algoritmo Apriori implementado por Yu Mochizuki denominado *apyori*³ em *Python*, que se encontra na versão 1.1.1. Essa aplicação proporcionou a geração das regras de associação, e, deste modo, a formação do perfil sociodemográfico do homem brasileiro com o câncer de próstata.

2.6 Pós-processamento dos Dados

Com as regras de associação geradas, os dados foram analisados de acordo com fundamentação teórica já exposta para verificar se a realidade brasileira segue o mesmo perfil ou apresenta peculiaridades com relação ao encontrado na literatura consultada.

Ademais, para permitir um maior entendimento dos leitores e gestores que possam utilizar este trabalho para tomar decisão no campo das campanhas sobre o câncer de próstata, será feito o uso de gráficos, histogramas e planilhas com as regras de associação encontrada com a análise da base do INCA. Afinal, segundo Campbell (2021) as pessoas absorvem melhor a informação quando é representada visualmente.

3. Resultados e Discussão

O propósito de analisar uma década de dados para definir o perfil sociodemográfico do homem brasileiro com o câncer de próstata, o período de 2010 a 2019 de dados do INCA foi explorado. Para embasamento da discussão, gráficos foram produzidos com uma aplicação desenvolvida em *Python*. Essa aplicação se encontra disponível no repositório do GitHub regras de associação⁴.

A partir dos dados selecionados foi possível montar uma amostra de 1.844.810 registros contidos na base de dados do estudo. Mas, depois do filtro C61, que corresponde ao CID da doença, realizado na coluna LOCTUDET para filtrar os dados sobre o câncer de próstata restou um total de 222.951 registros. Nesses dados em seguidas foram aplicados os filtros para excluir dados ruidosos e ausentes, que poderiam estar presentes.

No pré-processamento foram removidos 144.237 registros, que equivale a 64,69% da amostra, que são dados ruidosos, como por exemplo, valores de idade com a atribuição de 9999 e Estados de residência declarados como 99 e 77, que não fazem parte do dicionário da base. Além disso, foram removidos os registros, que apresentam a declaração de “sem informação”, que

³ <https://pypi.org/project/apyori/>

⁴ <https://github.com/DiasGustavo/regrasdeassociacao>

é selecionada nos quesitos do formulário de cadastro do paciente, a qual pode influenciar nas regras, haja vista que em dados opcionais (Alcoolismo, Tabagismo, Histórico familiar e Estado Conjugal) muitos escolheram essa opção ou o atendente a selecionou na hora do preenchimento do formulário. Como pode ser visto nos Gráficos 1, 2, 3 e 4.

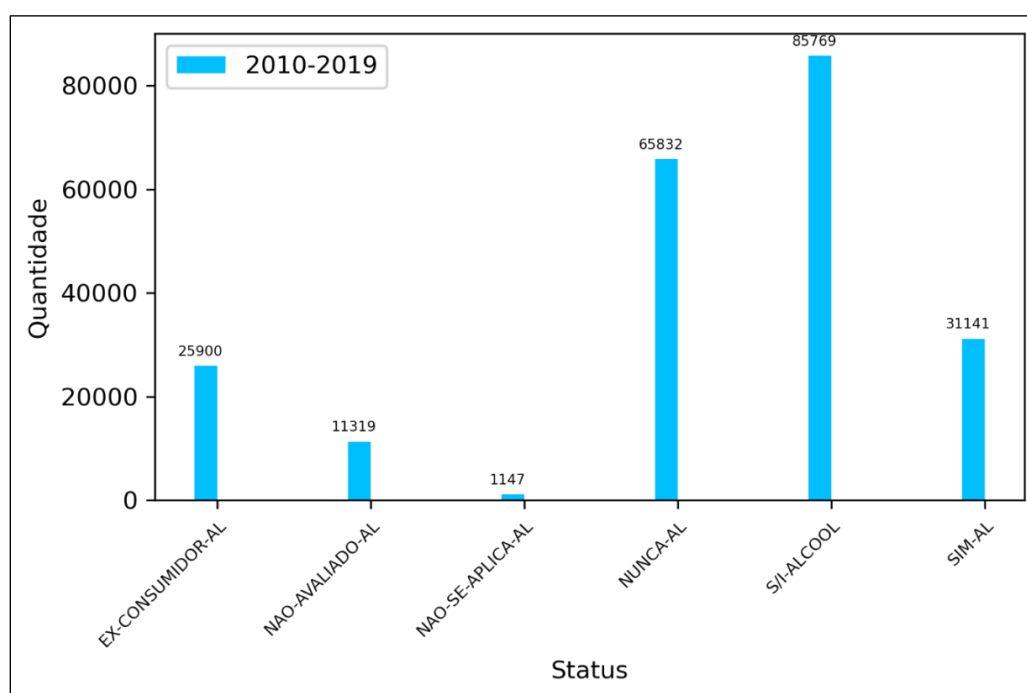
Ademais, os dados ausentes na base não ultrapassaram o valor de 119 registros, que corresponde ao percentual de 0,05% de todos os dados analisados. Porém, os valores referentes aos dados ruidosos de idade e opções 77 e 99 para o estado de residência do paciente chegaram ao valor de 1843 registros, isso corresponde a 0,8% da amostra. E por último, os valores referentes as opções que se declarou não possuir informações (S/I) sobre o quesito, a quantidade removida equivale a 142.394 registros que no geral corresponde a 64,4%, aproximadamente, dos dados restantes depois da aplicação dos filtros anteriores.

Contudo, o INCA (2020) em sua análise destacou que no período de 2012 a 2016 esses dados opcionais são coletados pelas instituições informantes do RHC de forma inconsistente, e não tem como distinguir qual instituição os itens passaram ser obrigatórios, haja vista que, a partir do momento que a instituição coleta esses dados pela primeira vez, para os demais casos que venham a ser registrados, esses itens passam a ser obrigatórios. Ou seja, muito dos dados opcionais não foram preenchidos. Ademais, segundo INCA (2010) quando é selecionado a opção “sem informação” ou o dado foi coletado incorretamente ou não há informações.

Esses tipos de informações podem criar tendências para resultados incoerentes tanto com a realidade quanto com os dados presentes na base. Nesse sentido, segundo Frank, Hall e Witten (2016, p.63) os erros e omissões em dados podem adquirir importância dentro da base de dados. Deste modo foi optado pela exclusão dos dados incoerentes ou ausentes, pois pode gerar resultados equivocados.

Contudo, inicialmente, foi analisado os campos opcionais, pois trazem informações que caracterizam a população masculina em fatores de risco. O primeiro ponto analisado será referente ao consumo de álcool que ficou distribuído de acordo com o Gráfico 1.

Gráfico 1 – Quantitativo de respostas para o quesito sobre o consumo de álcool.



Fonte: Autores (2021).

Como pode ser apreendido do Gráfico 1 o total de registros declarados “sem informação” é a opção mais encontrada dentro dos prontuários dos pacientes. Assim, traduzindo em números, os registros chegam ao valor de 85.769, que corresponde ao percentual de 38,8 % da amostra.

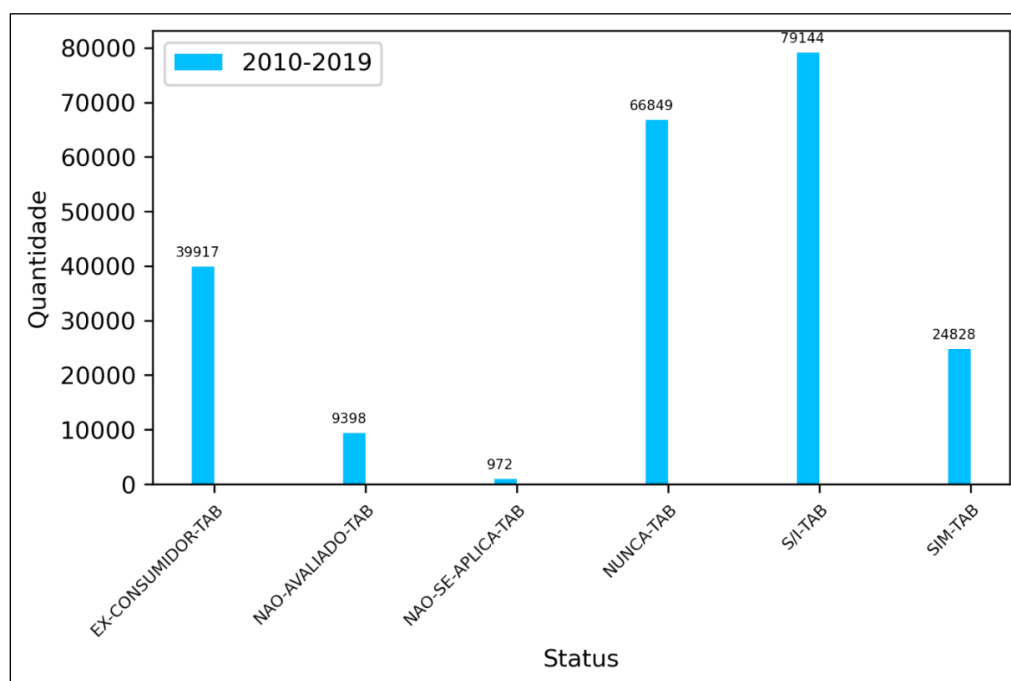
Ao eliminar esses dados inconsistentes encontra-se que a maioria declarou nunca ter bebido, isso chegou ao valor de 65.832 registros. Assim, o resultado aproxima-se ao da pesquisa de Rego et al. (2020) que apresentaram índices maiores para os homens que possuem câncer de próstata e não bebem.

De acordo com a Pesquisa Nacional de Saúde (PNS) realizada em 2019 pelo Instituto Brasileiro de Geografia e Estatísticas (IBGE) a população masculina acima de 18 anos, formada, aproximadamente, 76,69 milhões, o índice de pessoas que bebem pelo menos uma vez ou mais por semana é 37,1%, apesar da maioria não beber, o consumo excessivo de álcool é causa de 5,3% das mortes daquele ano no mundo (IBGE, 2020a). E em comparação aos dados do IBGE, os consumidores e ex-consumidores, encontrados na base do INCA, representam apenas 0,2% da população masculina.

Assim, depreende-se que apesar o consumo de álcool ser considerado como um fator de risco para a saúde a população masculina brasileira em sua maioria não faz uso regular de álcool de acordo com os dados do INCA e do IBGE.

Não só o alcoolismo é um fator impactante no câncer, como também o uso do tabaco contribui para o desenvolvimento do câncer (Hussain, et al., 2019). Com o auxílio do Gráfico 2 foram analisadas as respostas referentes ao consumo do tabaco por pacientes com câncer de próstata.

Gráfico 2 – Quantitativo de respostas para o quesito sobre o consumo de tabaco.



Fonte: Autores (2021).

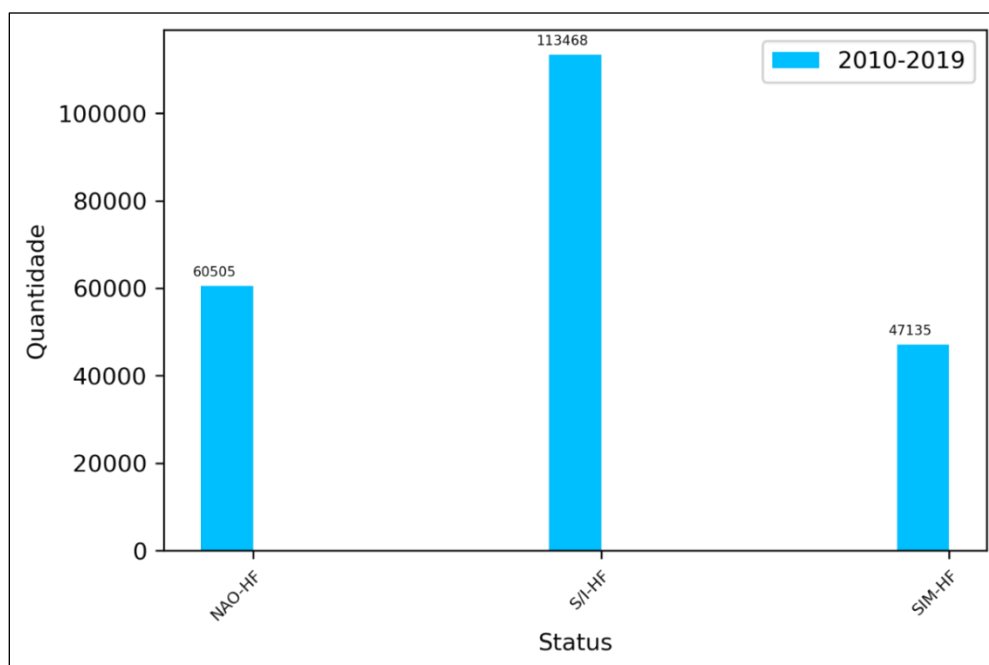
Neste Gráfico 2 também se identifica um valor maior para opção “sem informação”, que contabilizou o número de 79.144 registro, assim temos um percentual de 35,8% da amostra. Contudo, ao eliminarmos esses registros a opção que ficou em destaque é referente aos pacientes que nunca fumaram, chegando a um total de 66.849 registros.

Porém, não se pode deixar de realçar o número de consumidores e ex-consumidores de tabaco, que apresentam-se em 64.745 registros (29,28%). Isso caracterizando, que o brasileiro com câncer de próstata tem ou teve contato com o tabaco, que é um fator de agravo ao câncer.

No cenário brasileiro segundo o IBGE (2020a) 16,2% da população masculina faz uso de produtos derivados do tabaco, totalizando, aproximadamente, 12 milhões, sendo a minoria. Mas, em proporção ao resultado PNS 2019, a amostra do INCA representa apenas 0,5%.

Em seguida no Gráfico 3 se verifica o índice de casos de câncer de próstata na família, ou seja, quanto o fator de hereditariedade aparece nos casos registrados pelo INCA.

Gráfico 3 – Quantitativo de respostas para o quesito sobre caso de câncer na família.



Fonte: Autores (2021).

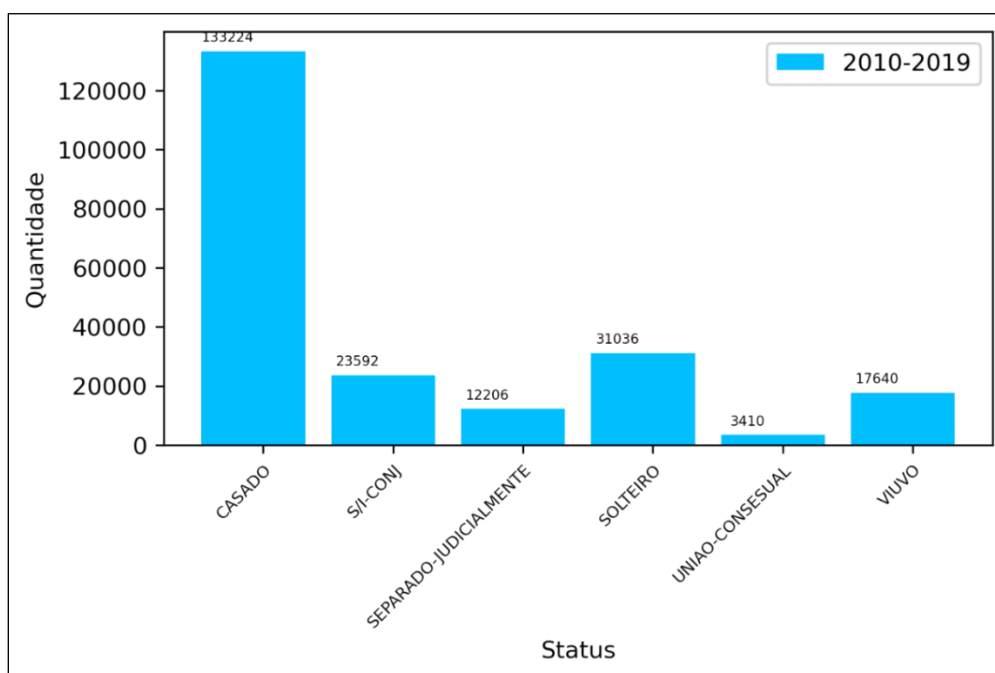
O Gráfico 3 representa os valores referente ao quesito da existência de caso de câncer de próstata na família. Como observa-se, não foi diferente dos que já foi exposto, a opção que mais se destaca é referente a “sem informações”, que contabiliza 113.468 registros, registrando um percentual de 51,3% da amostra.

No entanto, quando excluídos os dados sobre a opção “sem informação”, as opções ficaram niveladas, havendo 60.505 (27,3%) registros para não ocorrência de câncer e 47.135 (21,3%) para ocorrência de casos na família. Assim, foi percebido o percentual a mais para a não ocorrência de câncer de próstata na família.

Apesar de Silva e Nascimento (2017) em sua pesquisa, na cidade de Parintins, no Amazonas, ter chegado ao entendimento que o histórico familiar pode aumentar de 3 a 10 vezes a possibilidade de câncer, e Rego et al. (2020) na sua pesquisa, mais de 60% dos participantes do evento avaliado possuía histórico familiar de câncer, a realidade nacional de um país com dimensões continentais, se encontra quase que dividida segundo os registros do INCA para o período de 2010 a 2019.

Mas também, entre outros fatores, que foram analisados, tem-se o estado conjugal de cada indivíduo, ver Gráfico 4, que se torna importante no tratamento da doença, uma vez que, segundo Ferrão, Bettinelli e Portella (2017) a esposa possui um papel fundamental no apoio ao paciente, fazendo deste modo o papel de cuidadora e companheira do paciente.

Gráfico 4 – Quantitativo de respostas para o quesito sobre estado conjugal.

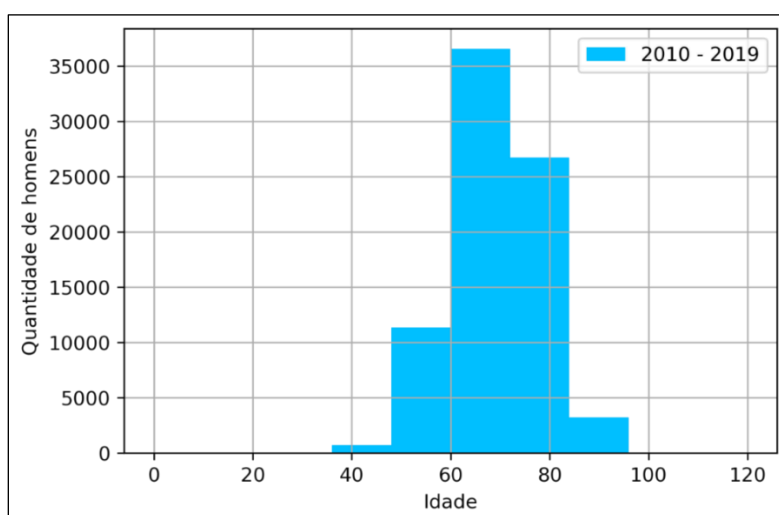


Fonte: Autores (2021).

No Gráfico 4 é exposta a situação com relação ao estado conjugal dos pacientes. Entretanto, foi percebido que entre os campos opcionais é o único em que a opção de “sem informação” foi menor em relação uma parte dos dados. Esta opção ficou com um total de 23.592 registros, que representa 10,7% da amostra, enquanto a opção de casado que apresenta o maior resultado ficou com um valor de 133.224 registros (60,3%) da amostra. Isso só vem corroborar o resultado encontrado nos trabalhos de Araújo et al. (2015), Menezes et al. (2019), e Rego et al. (2020).

Ao considerarmos o atributo idade, no cenário brasileiro, os homens com câncer de próstata, neste período, apresentaram uma média de idade em torno de 68,66 anos, e desvio-padrão de 9,04, demonstrando um certo nível de variância desse conjunto de registros. A distribuição dos pacientes que tiveram seus dados cadastrados na base de dados do INCA se encontram exposta no histograma do Gráfico 5.

Gráfico 5 - Histograma das idades dos homens propensos ao câncer de próstata no período de 2010 e 2019.



Fonte: Autores (2021).

O resultado do Gráfico 5 só vem confirmar que o câncer de próstata atinge os homens com idade avançada, ou seja, acima de 60 anos, e, muitas das vezes, raro em homens abaixo dos 50 (Cassell et al., 2019; Mota & Barros, 2019; Rego et al., 2020). Assim, percebe-se que no período de 2010 a 2019 a idade ficou concentrada entre 60 e 80 anos.

De acordo com as projeções do IBGE (2020b) para a população brasileira masculina no ano de 2019 disponível no sítio de projeções⁵ foi montada a Tabela 1 onde é disposta a proporcionalidade entre os números de casos de câncer de próstata e a população.

Tabela 1 – Proporcionalidade dos casos de câncer de próstata e a população masculina por faixa etária.

Faixa Etária (ANOS)	População 2019 (IBGE, 2020b)	Casos de Câncer de Próstata (INCA)	Proporcionalidade (%)
40 – 49	13.872.925	1260	0,0091
50 – 59	11.152.139	10.775	0,0966
60 – 69	7.431.729	29.971	0,4033
70 – 79	3.781.955	28.497	0,7533
80 – 89	1.371.137	7.572	0,5522
90 +	252.916	515	0,2036

Fonte: Autores (2021).

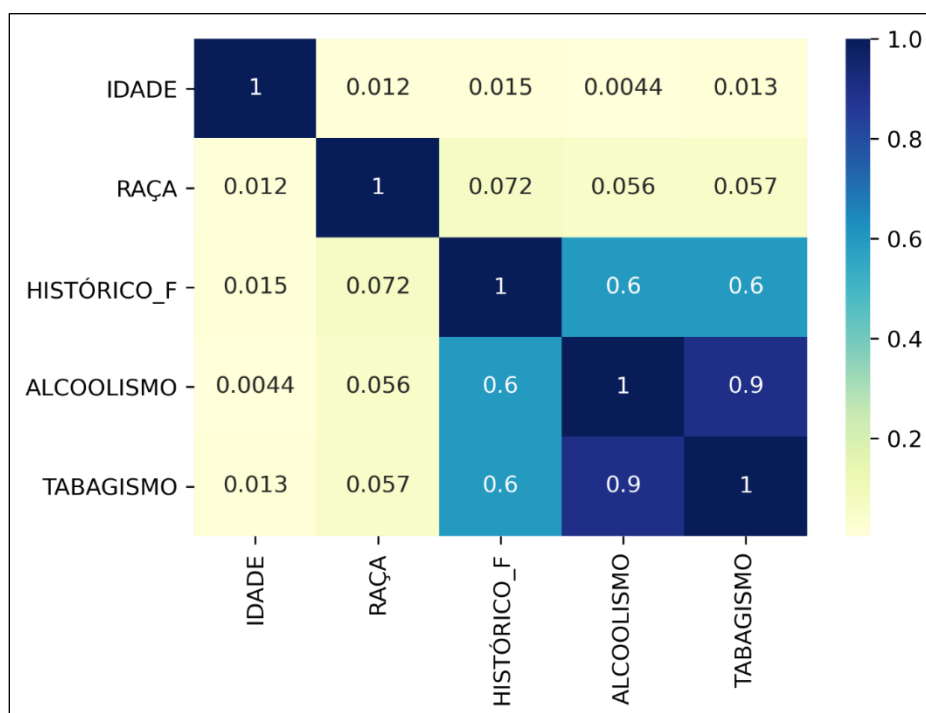
Assim, a Tabela 1 confirma que quanto maior a idade do homem, maior a possibilidade de câncer de próstata. Em conformidade com a observação feita por Mota e Barros (2019) os casos de próstata ocorrem mais nos homens acima de 60 anos.

Esses últimos cinco atributos analisados da base do INCA formam os principais fatores de risco para os homens que possuem câncer de próstata, pois podem contribuir para o surgimento e aumento das possibilidades do homem possuir esse tipo de doença durante a sua vida (Silva & Nascimento, 2017; Cavalcanti & Kruger, 2018).

Assim, para verificar o nível de relação entre esses fatores na base de dados do estudo desenvolvemos o mapa de calor das correlações, e o resultado ficou disposto no Gráfico 6.

⁵<https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?edicao=21830&t=resultados>

Gráfico 6 – Mapa de calor das correlações entre os atributos idade, raça, histórico familiar, alcoolismo e tabagismo.



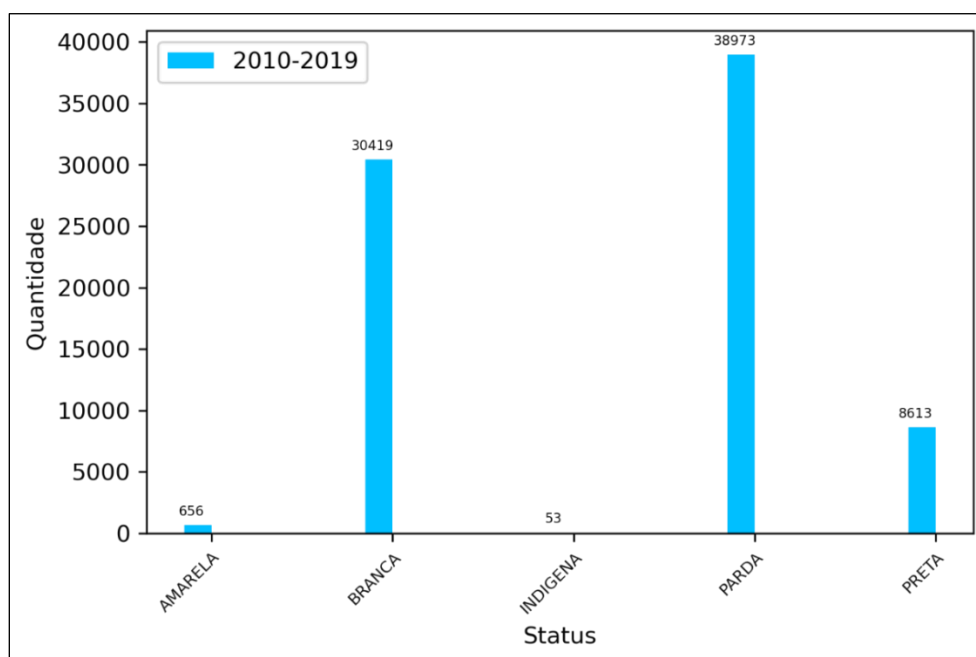
Fonte: Autores (2021).

Assim, como se ver no Gráfico 6 é possível verificar que o histórico familiar, o alcoolismo, e o tabagismo apresentam uma correlação moderada positiva, no valor de 0,6. Desse modo, pode-se depreender, que esses fatores se relacionam dentro dos dados do INCA, principalmente, do tabagismo e alcoolismo, que ao serem analisados juntos apresentam uma correlação forte.

Além disso, Silva e Nascimento (2017), Junior et al. (2015), e Mota e Barros (2019) em suas pesquisas em três cidades diferentes do Brasil, Boa Vista-RR, Parintins-AM e Recife-PE, respectivamente, confirmam que a idade, raça, etilismo e tabagismo são fatores que influenciam no câncer de próstata.

Apesar de a raça não apareça como fator influenciador nos dados de idade, alcoolismo e tabagismo, mas é um fator de risco (Araújo et al., 2015). Desse modo, em seguida foram analisadas as raças dos homens encontradas na amostra, e os resultados colocados no Gráfico 7. Contudo, o resultado demonstra que a realidade do homem brasileiro apresenta um destaque para a raça branca nos casos, isso visualizado nos trabalhos de Araújo et al. (2015) e Menezes et al. (2019). Contudo, a raça parda foi a que mais se destacou com 38.973, isso corresponde ao percentual de 49,5% da amostra. Esse resultado sendo acompanhado de perto pela raça branca com o valor de 30.419 registros, percentual de 38,6%, como pode ser percebido através da leitura do Gráfico 7.

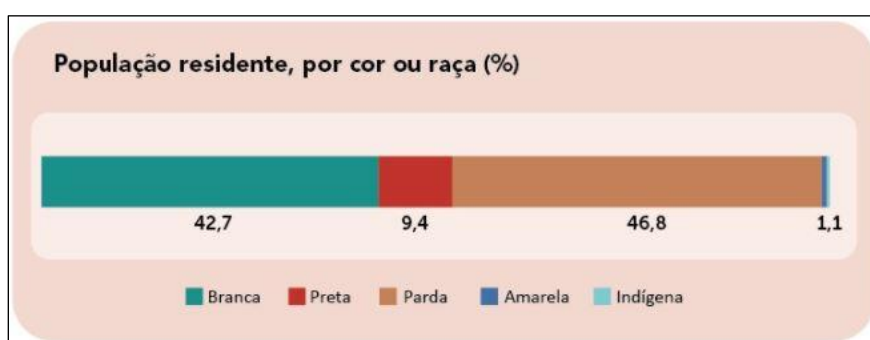
Gráfico 7 - Comparativo entre as raças dos homens com câncer de próstata.



Fonte: Autores (2021).

Deste modo, de acordo com o Gráfico 8, percebe-se que a realidade brasileira foge dos estudos que aponta a raça negra como maior incidência do câncer de próstata, uma vez que a mesma é considerada um fator de risco (Mota & Barros, 2019; Cassell et al., 2019). Esse fato é compreensível, pois a maior parte da população brasileira é composta por pardos, a qual corresponde a 46,8% da população no ano de 2019, como pode ser visto no Gráfico 8 (IBGE, 2019).

Gráfico 8 - População residente, por cor ou raça.



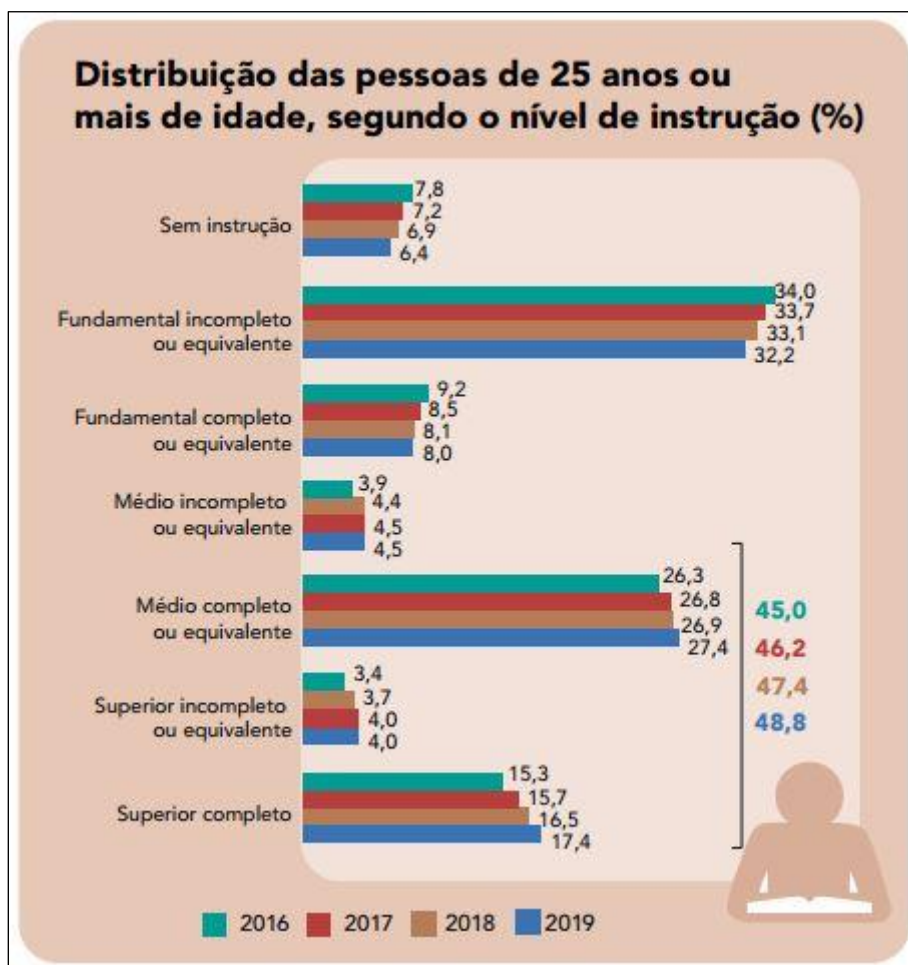
Fonte: IBGE (2019).

No Gráfico 10, encontra-se o estudo do grau de instrução da população masculina brasileira, que possuiu câncer de próstata no período estudado. Como é notório o número maior se deu ao fundamental incompleto, chegando ao número de 41.741 registros, que forma o percentual de 53% da amostra. Porém, um destaque tem que ser feito com relação a taxa de homens que declararam não ter formação nenhuma, 11.623 registros (14,8%), pois entende-se deste modo que a população masculina brasileira não possui um grau de instrução elevado. Para este cenário, de acordo com Panis et al. (2018) a baixa escolaridade é um fator socioambiental associado aos casos de câncer.

Apesar desse nível de instrução encontrado, de acordo com a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD) de 2019 com dados da educação o brasileiro fica em média 9,4 anos estudando, a taxa de analfabetismo para as pessoas

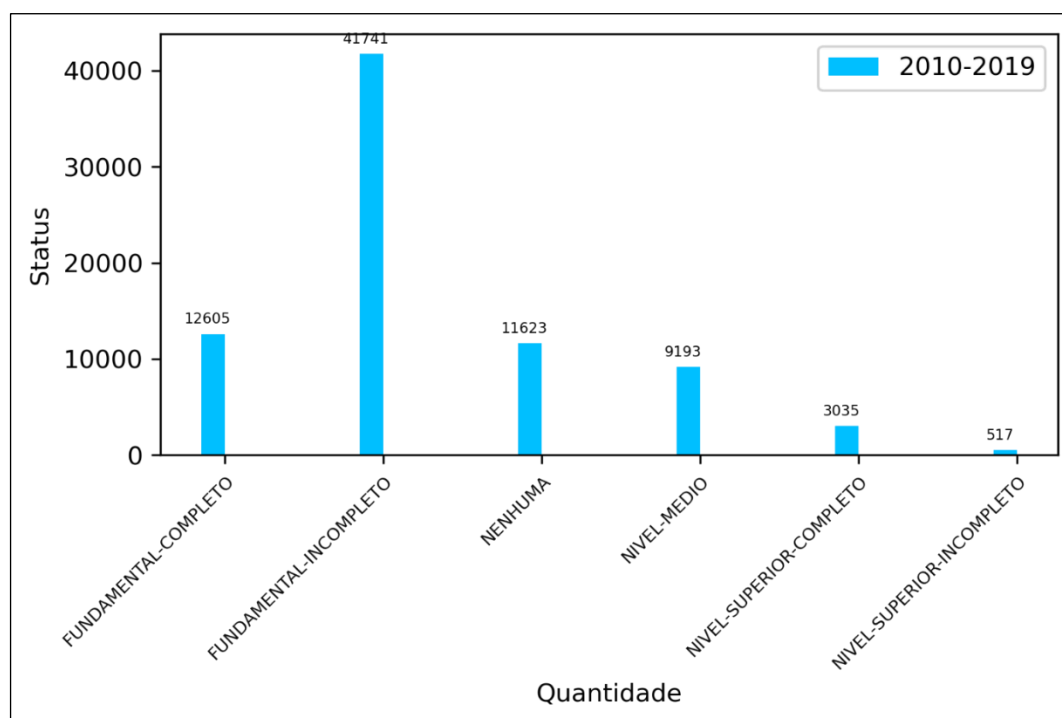
15 anos ou mais 6,6% e sem instrução de 6,4% da população (IBGE, 2020c). Essa pesquisa apresenta dados que demonstra a diminuição nos últimos anos dos índices, como pode ser visto no Gráfico 9, contudo ainda existe um número considerável de pessoas analfabetas ou sem instrução.

Gráfico 9 – Distribuição das pessoas de 25 anos ou mais de idade, segundo o nível de instrução.



Fonte: IBGE (2020b).

Gráfico 10 - Comparativo entre os graus de instrução dos homens com câncer de próstata.



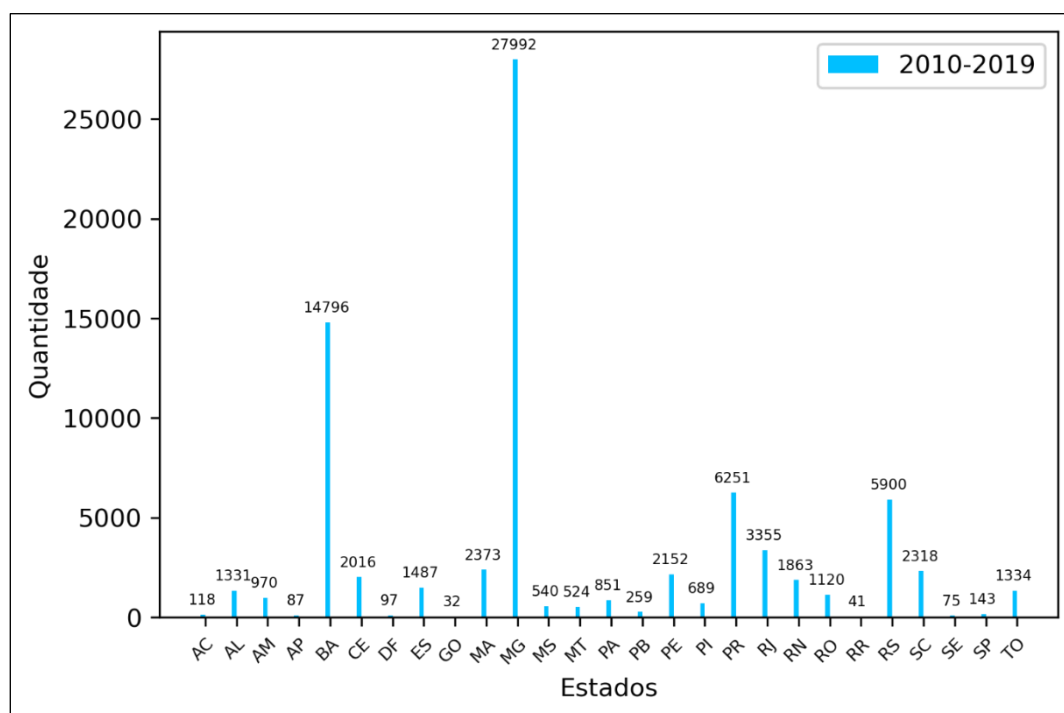
Fonte: Autores (2021).

Contudo, entendam que apesar deste trabalho ser uma visão geral do Brasil, estudos de locais específicos do país já refletem este perfil como é apresentado no trabalho de Ferrão, Bettinelli e Portella (2017) e Rego et al. (2020), o qual o primeiro analisou o atendimento de homens com câncer de próstata em um hospital, do Rio Grande do Sul, e o segundo analisou pacientes atendidos em um evento de prevenção, no estado de Minas Gerais.

Os casos de câncer de próstata vêm crescendo ano a ano no Brasil, assim se torna um caso de saúde pública, que precisa ser cuidado (Cavalcanti & Kruger, 2018). No Gráfico 11, pode-se notar que o Estado de Minas Gerais se destaca entre todos nessa década avaliada, chegando a 27.992 registros, que equivale a 35,6% de toda a amostra. E em segundo lugar temos o estado da Bahia com 14.796 registros (18,8%).

No âmbito regional, a região que se destaca é a região sudeste com 32.997 registros, que equivale a 41,9% de todos os dados, em segundo lugar temos a região nordeste com 25.554 registros ou 32,5% dos casos, em terceiro temos a região sul com 14.469 registros ou 18,4%, logo em seguida em quarto lugar temos a região norte com 4.521 registros, ou 5,7% da amostra e por último a região centro-oeste com 1.193 registros, ou 1,5%. Essa disposição já vem acontecendo e prevista pela análise do período de 2012 a 2016 do INCA (2020).

Gráfico 11 - Distribuição das ocorrências de casos de próstata no Brasil no período de 2010 a 2019.



Fonte: Autores (2021).

Além disso, as maiores proporcionalidades entre a população e os casos de câncer de próstata ficam ainda destacando os Estados de Minas Gerais e Bahia conforme a Tabela 2. Todavia, é possível ainda perceber que os Estados do Tocantins, Rondônia, Paraná, Rio Grande do Norte, e Rio Grande do Sul estão entre as maiores proporcionalidades.

Tabela 2 - Proporcionalidade dos casos de câncer de próstata e a população masculina por estado.

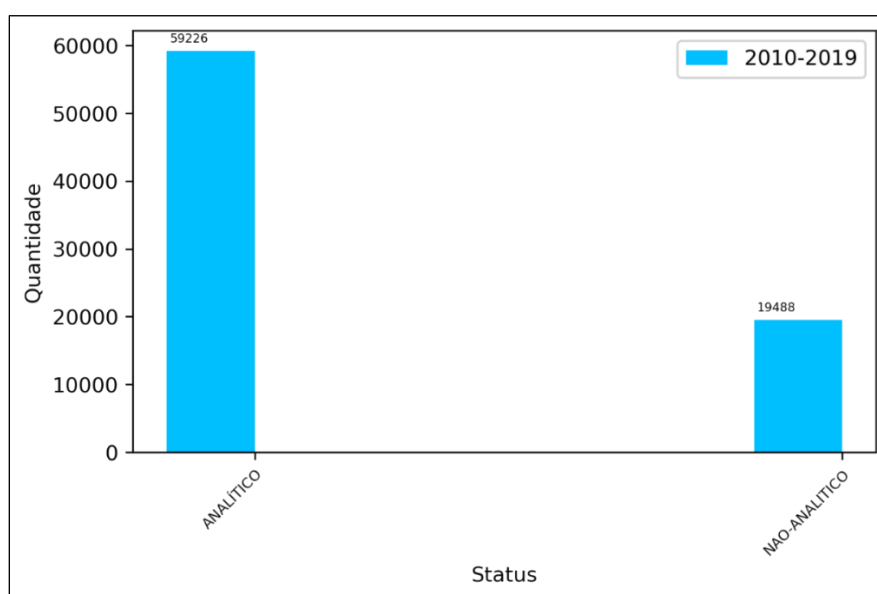
Estado	Projeção de População de 2019 (IBGE, 2020b)	Casos de Câncer de Próstata (INCA)	Proporcionalidade (População/Casos) (%)
AC	441.216	118	0,0267
AL	1.601.112	1331	0,0831
AM	2.081.262	970	0,0466
AP	423.498	87	0,0205
BA	7.233.509	14.796	0,2045
CE	4.432.035	2016	0,0455
DF	1.447.284	97	0,0032
ES	1.978.483	1487	0,0752
GO	3.481.598	32	0,0009
MA	3.479.859	2373	0,0682
MG	10.422.468	27.992	0,2686
MS	1.379.290	540	0,0391
MT	1.767.091	524	0,0296
PA	4.315.587	851	0,0197
PB	1.939.480	259	0,0133
PE	4.588.206	2152	0,0469
PI	1.586.538	689	0,0434
PR	5.602.812	6.251	0,1116
RJ	8.255.502	3.355	0,0406
RN	1.709.856	1.863	0,1089
RO	899.108	1.120	0,1246
RR	312.110	41	0,0131
RS	5.536.738	5.900	0,1065
SC	3.554.814	2.318	0,0652
SE	1.110.281	75	0,0067
Estado	Projeção de População de 2019 (IBGE, 2020b)	Casos de Câncer de Próstata (INCA)	Proporcionalidade (População/Casos) (%)
SP	22.388.135	143	0,0006
TO	792.423	1.334	0,1683

Fonte: Autores (2021).

Quando os pacientes procuram um hospital ligado ao RHC, o seu caso pode ser classificado como caso analítico ou não analítico. No caso analítico, o hospital vinculado ao RHC diagnosticou e realizou o tratamento, ou acompanhou o tratamento do paciente, mas no caso não analítico o diagnóstico e o tratamento muitas vezes não são realizados pelo hospital vinculado ao RHC ou o paciente faleceu nas primeiras 48 após a matrícula no hospital (INCA, 2010).

No Gráfico 12 tem-se que os casos analíticos se destacam mais do que os casos não analíticos na rede brasileira. Mas, deve-se da ênfase que a classificação depende do atendente, que realiza a mesma na primeira entrevista do paciente (INCA, 2010).

Gráfico 12 - Classificação dos tipos de casos dos pacientes no período de 2010 e 2019.



Fonte: Autores (2021).

Apesar desses dados iniciais já retornarem informações, que passa construir o perfil do brasileiro com o câncer de próstata, ainda existem muitos dados ainda para serem analisados, uma vez que segundo Frank et al. (2016) é no meio dos dados que se encontram informações importantes que raramente são reveladas.

Desse modo, com o objetivo de revelar o conhecimento intrínseco dentro da base do INCA, logo após o levantamento dos dados anteriormente analisados, foi executado o algoritmo Apriori para a formação das regras de associação, e descobrir todo o conhecimento que possa ainda estar faltando ser demonstrado.

O suporte foi o primeiro índice definido com o valor de 0,129, definido de acordo com a Equação I, uma vez que o total de casos de câncer no período de 2010 a 2019 foram 1.844.810 registros, e entre eles os referentes ao câncer de próstata ficaram com 238.795 registros. Além disso, a confiança ficou variando de 0,3 a 0,8, e por último o lift mínimo de 1,2 definido de acordo com a metodologia. Na implementação *Apyori* o total de elementos mínimos na regra também podem ser definidos, assim foi definido que as regras deveriam ser formadas por no mínimo 2 itens, pois as regras são formadas no padrão *X* então *Y*, logo se garante que sempre as regras serão formadas por antecedentes e consequentes.

Outrossim, destaca-se os principais resultados encontrados com as regras que mais se repetiram, apresentando um grau de confiança acima de 30%, e não apresenta a ausência de informação no quesito, ou seja, não foi envolvido a opção onde o atendente ou o paciente declarou “sem informação”. Logo, a Tabela 3 apresenta os resultados do período estudado, onde tem-se representado por “c” o índice de confiança e o *lift* da regra encontrada.

Tabela 3 - Regras de associação com dados do período de 2010 a 2019.

1.	Mora na Bahia, então nunca fumou, c=0,71, lift= 1,367
2.	Mora na Bahia, então é da raça parda, c=0,78, lift= 1,587
3.	Ex-consumidor de álcool, então é ex-consumidor de tabaco, c=0,68, lift=2,476
4.	Ex-consumidor de tabaco, então é ex-consumidor de álcool, c=0,49, lift=2,476
5.	Nunca bebeu, então nunca fumou, c=0,74, lift=1,430
6.	Nunca fumou, então nunca bebeu, c=0,75, lift=1,430
7.	Nunca bebeu e o caso é analítico, então nunca fumou, c=0,74, lift=1,427
8.	Caso analítico e nunca fumou, então nunca bebeu, c=0,75, lift=1,424
9.	Nunca bebeu e a raça é branca, então nunca fumou, c=0,70, lift=1,342
10.	Raça branca e nunca fumou, então nunca bebeu, c=0,77, lift=1,456
11.	Nunca bebeu e casado, então nunca fumou, c=0,74, lift=1,431
12.	Casado e nunca fumou, então nunca bebeu, c=0,76, lift=1,440
13.	Possui fundamental incompleto e nunca bebeu, então nunca fumou, c=0,73, lift=1,413
14.	Possui fundamental incompleto e nunca fumou, então nunca bebeu, c=0,77, lift=1,457
15.	Nunca bebeu e não possui histórico familiar, então nunca fumou, c=0,76, lift=1,459
16.	Nunca fumou e não possui histórico familiar, então nunca bebeu, c=0,77, lift=1,478
17.	Nunca bebeu e a raça é parda, então nunca fumou, c=0,77, lift=1,493
18.	Raça parda e nunca fumou, então nunca bebeu, c=0,75, lift=1,419
19.	Nunca bebeu e possui histórico familiar, então nunca fumou, c=0,72, lift=1,390
20.	Possui histórico familiar e nunca fumou, então nunca bebeu, c=0,72, lift=1,364
21.	Nunca bebeu, casado e o caso é analítico, então nunca fumou, c=0,74, lift=1,425
22.	Casado, o caso analítico e nunca fumou, então nunca bebeu, c=0,75, lift=1,433
23.	Possui fundamental incompleto, nunca bebeu e o caso é analítico, então nunca fumou, c=0,73, lift = 1,410
24.	Possui fundamental incompleto, caso analítico e nunca fumou, então nunca bebeu, c=0,76, lift=1,445
25.	Nunca bebeu, caso analítico, e não possui histórico familiar, então nunca fumou, c=0,75, lift=1,452
26.	Nunca fumou, caso analítico e não possui histórico familiar, então nunca bebeu, c=0,77, lift=1,467
27.	Nunca bebeu, caso analítico e raça parda, então nunca fumou, c=0,77, lift=1,485
28.	Raça parda, caso analítico e nunca fumou, então nunca bebeu, c=0,74, lift=1,406
29.	Possui fundamental incompleto, nunca bebeu e casado, então nunca fumou, c=0,73, lift=1,412
30.	Possui fundamental incompleto, casado e nunca fumou, então nunca bebeu, c=0,77, lift=1,473
31.	Nunca bebeu, casado, não possui histórico familiar, então nunca fumou, c=0,76, lift=1,467
32.	Nunca fumou, casado e não possui histórico familiar, então nunca bebeu, c=0,78, lift=1,493

Fonte: Autores (2021).

Além do que foi caracterizado nos parágrafos iniciais, as regras de associação apresentaram dados envolvendo Estado de residência, estado conjugal, grau de instrução, raça, alcoolismo, tabagismo, presença de histórico familiar e tipo de caso, o qual a variação da confiança alterou apenas na exclusão das regras 3, 4 e 9 presentes na Tabela 3 à medida que definimos um nível maior.

Ademais, pode-se destacar alguns pontos dentre as regras, o primeiro ponto, foi visualizado quanto ao termo de consumo de álcool, onde a maioria declara que nunca bebeu, desse modo as regras geradas com relação à esta opção houve uma maior significância na regra 32 da Tabela 3. Essa regra apresenta uma confiança de 78% e um lift de 1,493, esses dados garante que esse consequente (nunca bebeu) é mais frequente em regras contendo os antecedentes de nunca ter fumado, ser casado e não apresentar histórico familiar dentro da base de dados estudada.

Esses dados só confirmam mais ainda a pesquisa realizada por Menezes et al. (2019) onde eles aplicaram um questionário e um dos quesitos traz a afirmativa, que é falsa segundo os autores, “As bebidas alcoólicas contribuem para o desenvolvimento do câncer de próstata.” (Menezes et al., 2019, p. 1176).

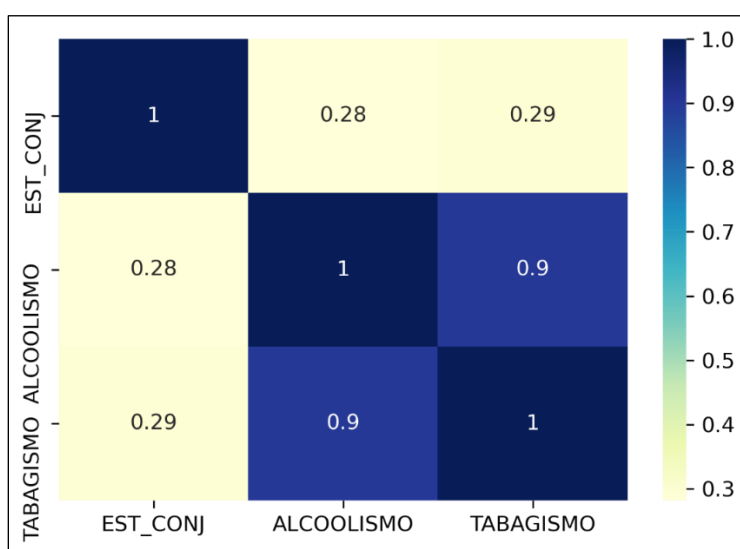
Além disso, outro fator, que apareceu várias vezes no consequente das regras, o fator de consumo de tabaco, onde os pacientes em sua maioria declararam que nunca usou, esse consequente sofreu mais influência, de acordo com as regras geradas, quando os antecedentes da regra eram nunca bebeu e a raça é parda. Essa regra apresentou o índice de confiança de 77% e lift de confiança de 1,493, como pode ser visto na regra 17. Isso no cenário de Montes Claros, em Minas Gerais, já ocorria, como é exposto na pesquisa feita por Rego et al. (2020), o qual 52,6% dos participantes do estudo declararam ser não fumantes.

Esses dois últimos fatores analisados estão, frequentemente, associados quando ocorrem, assim podemos dizer que a ausência de consumo também estaria (Stone et al., 2019).

Quando se considera o histórico familiar, pode-se concluir de acordo com as regras geradas, que a maior parte é influenciada por não possuir histórico familiar, cenário que foi encontrado também nos trabalhos de Rego et al. (2020) e Stone et al. (2019). Mas, nota-se que ainda no meio dos dados estudados, há influência do histórico familiar, haja vista que apareceu duas regras ligadas ao consumo de álcool e tabaco, como pode ser visto nas regras 19 e 20. Isso é validado pela correlação positiva moderada, que ficou demonstrada no Gráfico 6.

Não só esses últimos fatores influenciaram nos dados, mas também o estado conjugal dos pacientes interferiu no resultado das regras de associação. Esse item ratificou que a maioria dos homens são casados, e se relaciona na maioria das regras com o consumo de álcool e tabaco na formação das mesmas. Porém, como podemos ver no Gráfico 13, a correlação entre esses três fatores é fraca.

Gráfico 13 – Correlação dos atributos estado conjugal, alcoolismo e tabagismo.



Fonte: Autores (2021).

De acordo com as regras geradas, a raça parda como antecedente, ocorre entre as regras que possui um dos maiores lift, como pode ser visto na regra 17, ou seja, essa raça possui um nível de influência maior. Assim, depreende-se que a raça parda é a que mais se destaca entre os pacientes com câncer de próstata. Além do mais, destaca-se que os pacientes da Bahia eram pardos, isso com uma confiança de 78%.

Esse resultado para a raça diverge do que ocorre no mundo, onde se dá destaque de ocorrência de câncer de próstata em indivíduos da raça negra, fato esse ocorrido também nas pesquisas de Araújo et al. (2015) e Stone et al. (2019).

Seguindo a análise detalhada das regras, a única característica quanto ao grau de instrução, que se destaca entre os antecedentes é a opção de fundamental incompleto, que entre as regras se encontra em uma que apresenta um dos maiores índices de lift, como pode ser visto na regra 30. Haja vista que, pesquisas no território nacional, já demonstrava que os brasileiros com câncer de próstata não possuíam nível de escolaridade maior do que o fundamental completo (Ferrão et al., 2017; Menezes et al., 2019; Rego et al., 2020)

Apesar de que os dados brutos destacaram o Estado de Minas Gerais como residência da maioria dos pacientes, mas o Estado que apareceu dentro das regras foi o Estado da Bahia, onde depreende-se que os baianos que possuíam câncer de próstata eram pardos e declararam que nunca fumaram.

Portanto, a partir das regras obtidas e da moda das confianças que formam as regras, pode-se formar o perfil sociodemográfico do homem brasileiro com câncer de próstata no período de 2010 a 2019. Este perfil ficaria como sendo homens

que nunca fumaram, com confiança de 74%, nunca beberam, com confiança de 77%, de raça parda, em sua maioria, com a confiança de 77%, casados, o qual destacamos a confiança de 76%, de fundamental incompleto, cujo dado aparece com confiança de 73%, e não possuem histórico familiar da doença, com confiança de 77%.

Além disso, como a maioria dos casos foram considerados analíticos, assim entende-se que o atendimento na rede pública de hospitais de oncologia prevalece em relação a rede privada no tratamento de câncer de próstata.

4. Considerações Finais

A ausência de um perfil do homem com o câncer de próstata dificulta o diagnóstico, pois o câncer deste tipo é assintomático no começo (Mota & Barros, 2019). Além disso, é importante a definição desse perfil, haja vista a possibilidade de facilitar o diagnóstico e o encaminhamento do tratamento.

Não só a escassez de literatura que trate do desenvolvimento de técnicas que possam definir esse perfil, mas também há o agravante da grande quantidade de dados armazenada em bases de dados que dificulta o estudo e a análise por pessoas (Oliveira, et al., 2007; Araujo et al., 2015).

Deste modo, este trabalho uniu os registros armazenados na base de dados do INCA com homens diagnosticados com câncer de próstata, para poder traçar o perfil sociodemográfico do homem brasileiro no período de 2010 a 2019 através do uso do algoritmo Apriori. Assim, através da análise dos dados formou-se as regras de associação para ser projetado o possível perfil do homem propenso a esse tipo de câncer.

Com os gráficos e regras geradas ao final do processo de aplicação do algoritmo Apriori, foi percebido que os fatores de tabagismo, alcoolismo, raça e estado conjugal são os que mais se destacaram por aparecerem nas regras com os maiores índices de confiança. Porém, fugindo um pouco da literatura encontrada, que dá o destaque a raça negra como maior incidência de casos, a realidade brasileira é destacada pelos casos na raça parda.

Ademais, o destaque dos fatores de tabagismo e de alcoolismo declarados no preenchimento das fichas dos pacientes, em sua maioria, como nunca beberam ou nunca fumaram apresentam uma forte correlação. Assim, resultados demonstram que esses dois tópicos são bastantes interligados.

Apesar da incompletude dos dados opcionais, na base de registros do INCA, é importante destacar que a análise foi feita a nível nacional e pode ser utilizada para definição de campanhas de informação e acompanhamento dos homens com câncer de próstata. Assim, se tornar um instrumento norteador no contexto da saúde do homem no Brasil.

Com o propósito de um aprofundamento sobre o câncer de próstata no Brasil, pretende-se, para trabalhos futuros, desenvolver pesquisas que possam garantir a análise de dados em comparativo com a base de registro de câncer de base populacional – RCBP, também mantido pelo INCA; Aplicar análise de Componentes Principais (ACP) para constatar o nível de influência entre os fatores de risco cor/raça, tabagismo, alcoolismo e histórico familiar; e implementar algoritmos de previsão nos dados da base do INCA para assim auxiliar os profissionais da saúde no diagnóstico e tratamento da doença.

Agradecimentos

O presente trabalho foi realizado com apoio da Universidade Estadual da Paraíba, edital PRPGP 003/2022.

Referências

Aguiar, J. S., Pereira, L. D. A. & Thomas, C. A. B. (2017). Assistência Oncológica no Estado do Espírito Santo, a partir do Sistema Integrador dos Registros Hospitalares de Câncer; 2000 – 2014. *Informativo Vigilância do Câncer*. <https://saude.es.gov.br/Media/sesa/DANTS/INFORMATIVO%20VIGILANCIA%20DO%20CANCER%20-%20RHC%2020%2012%202017.pdf>

Araújo, M. S. M., Sardinha, A. H. L., Neto, J. A. F., Silva, E. L., & Lopes, M. L. H. (2019). Caracterização social e clínica dos homens com câncer de próstata atendidos em um hospital universitário. *Revista Mineira de Enfermagem*, 19(2), 196-203. <http://www.dx.doi.org/10.5935/1415-2762.20150035>

- Baldomir, R. A. (2017). *Aplicação do Algoritmo Apriori para Detectar Relacionamento entre Empresas nos Processos Licitatórios do Governo Federal* [Monografia, Universidade de Brasília]. <https://bdm.unb.br/handle/10483/19987>
- Campbell, A. (2021). *Data Visualization Guide: Clear Introduction to Data Mining, Analysis and Visualization*. Independently Published.
- Cassell, A., Yunusa, B., Jalloh, M., Mbodji, M. M., Diallo, A., Ndoye, M., Kouba, S. C., Labou, I., Niang, L. & Gueye, S. M. (2019). A Review of Localized Prostate Cancer: An African Perspective. *World Journal Of Oncology*, 10 (4-5). <https://doi.org/10.14740/wjon1221>
- Cavalcanti, G. & Kruger, F. P. G. (2018). Conhecimento e Atitudes sobre o Câncer de Próstata no Brasil. *Revisão Integrativa. Revista Brasileira de Cancerologia*, 64 (4), 561-567. <https://doi.org/10.32635/2176-9745.RBC.2018v64n4.206>
- Ferrão, V., Berrinelli, L. A. & Portella, M. R. (2017). Vivências de Homens com Câncer de Próstata. *Revista de Enfermagem UFPE On Line*, Recife; 11 (10), 4157 - 4164. <https://periodicos.ufpe.br/revistas/revistaenfermagem/article/download/231178/25153>
- Frank, E., Hall, M. A. & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. (4. ed.). Morgan Kaufmann.
- Hussain, L., Ali, A., Rathore, S., Saeed, S., Idris, A., Usman, M. U., Iftikhar, M. A. & Suh, D. O. (2019). Applying Bayesian Network Approach to Determine the Association Between Morphological Features Extracted from Prostate Cancer Images. *IEEE Access*, 7, 1586-1601. <https://ieeexplore.ieee.org/document/8579592>.
- IBGE. (2019). Instituto Brasileiro De Geografia E Estatísticas. Conheça o Brasil – População Cor ou Raça. *IBGE educa*. <https://educa.ibge.gov.br/jovens/conheca-obrasil/populacao/18319-cor-ou-raca.html>.
- IBGE. (2020a). Instituto Brasileiro de Geografia e Estatísticas. *Pesquisa nacional de saúde: 2019: percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal: Brasil e grandes regiões / IBGE, Coordenação de Trabalho e Rendimento*. IBGE. <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101764.pdf>
- IBGE. (2020b). Instituto Brasileiro De Geografia E Estatísticas. *Projeção da população do Brasil por sexo e idade para o período 2010-2060*. <https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-dapopulacao.html?edicao=21830&t=resultados>.
- IBGE. (2020c). Instituto Brasileiro De Geografia E Estatísticas. *Educação: 2019*. IBGE. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf
- INCA. (2010). Instituto Nacional de Câncer José de Alencar Gomes da Silva. *Registros Hospitalares de Câncer: Planejamento e Gestão*. (2. ed.). INCA.
- INCA. (2020). Instituto Nacional de Câncer José de Alencar Gomes da Silva. Perfil da Assistência Oncológica no Brasil entre 2012 e 2016. *Informativo Vigilância do Câncer*. (7). <https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/informativo-vigilancia-do-cancer-n7-2020.pdf>
- Jamsa, K. (2021). *Introduction to Data Mining and Analytics with Machine Learning in Rand Python*. World Headquarters Jones & Bartlett Learning.
- Junior, M. M. L., Reis, L. O., Ferreira, U., Cardoso, U. O., Barbieri, R. B., Mendonça, G. B. & Ward, L. S. (2015). Unraveling Brazilian Indian Population Prostate Good Health: Clinical, Anthropometric and Genetic Features. *International braz j urol*, 41 (2). <https://doi.org/10.1590/S1677-5538.IBJU.2015.02.23>
- Menezes, R., Menzes, M., Teston, E. F., Matumoto, S., Faller, J. W. (2019). Knowledge, Behaviour and Health Practices of Men Concerning the Prostate Cancer. *Revista de Pesquisa Cuidado é Fundamental Online*, 11(5), 1173-1179. <http://dx.doi.org/10.9789/2175-5361.2019.v11i5.1173-1179>.
- Ministério da Saúde. (1998). “Portaria nº 3.535”. https://bvsm.sau.gov.br/bvs/sau/legis/gm/1998/prt3535_02_09_1998_revog.html#:~:text=Estabelece%20crit%C3%A9rios%20para%20cadastramento%20de%20centros%20de%20atendimento%20em%20oncologia.
- Ministério da Saúde. (2005). “Portaria nº 741”. https://bvsm.sau.gov.br/bvs/sau/legis/sas/2005/prt0741_19_12_2005.html.
- Mota, T. R., Barros, D. P. O. (2018). Perfil dos pacientes com câncer de próstata em hospital de referência no estado de Pernambuco. *Revista brasileira de análises clínicas*, 50 (4), 334-338. <https://doi.org/10.21877/2448-3877.201900766>
- Oliveira, C. O., Marques, P. M. A., Filho, W. C. C. (2007). Grades computacionais na otimização da recuperação de imagens médicas baseada em conteúdo. *Radiologia brasileira*, 40 (4), 255-261. <https://doi.org/10.1590/S0100-39842007000400011>
- Panis, C., Kawasaki, A. C. B., Pascotto, C. R., Justina, E. Y. D., Vicentini, G. E., Lucio, L. C. & Prates, R. T. C. (2018). *Revisão crítica da mortalidade por câncer usando registros hospitalares e anos potenciais de vida perdidos*. 16 (1). <https://doi.org/10.1590/S1679-45082018AO4018>
- Preissler, A. (2016). *Data Mining para Definição dos Perfis de Pacientes com Câncer de Estômago*. [Monografia, Universidade Regional do Noroeste do Estado do Rio Grande do Sul]. <https://bibliodigital.unijui.edu.br:8443/xmlui/bitstream/handle/123456789/4216/Amanda%20Preissler.pdf?sequence=1&isAllowed=y>
- Prodanov, C. C., Freitas, E. C. (2013). *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico*. (2. ed.). Feevale.
- Rego, R. F. N. B., Barros, R. A., Pimenta, L. O. S., Rodrigues, J. V. C., ANJOS, E. B. (2020). Perfil Clínico Epidemiológico da População Atendida Num Programa de Rastreamento de Câncer de Próstata. *Revista de Atenção à Saúde*, 18 (65), 38-47. <https://doi.org/10.13037/ras.vol18n65.6647>
- Sacramento, R. S., Simião, L. J., Viana, K. C. G., Andrade, M. A. C., Amorin, M. H. C. & Zandonade, E. (2019). Associação de variáveis sociodemográficas e clínicas com os tempos para início do tratamento do câncer de próstata. *Ciência & Saúde Coletiva*, 24 (9), 3265-3274. <https://doi.org/10.1590/1413-81232018249.31142017>
- Santos, M. O. (2018). Estimativa 2018: Incidência de Câncer no Brasil. *Revista Brasileira De Cancerologia*, 64(1), 119–120. <https://doi.org/10.32635/2176-9745.RBC.2018v64n1.115>

Sharma, S. & Bansal, M. (2020). Real-Time Sentiment Analysis Towards Machine Learning. *International Journal of Scientific & Technology Research*, 9 (2). <https://10.13140/RG.2.2.27062.24643>

Silva, L. A., Peres, S. M. & Boscaroli, C. *Introdução à Mineração de Dados: com aplicações em R*. Elsevier.

Silva, J. S. & Nascimento, L. P. (2017). *Fatores Culturais Associados a não Adesão aos Exames Preventivos de Câncer de Próstata em Parintins* [Trabalho de Conclusão, Universidade do Estado do Amazonas]. <http://repositorioinstitucional.uea.edu.br/handle/riuea/759>

Stone, C. R., Courneya, K. S., McGregor, S. E., Li, H. & Friedenreich, C. M. (2019). Determinants of changes in physical activity from prediagnosis to post-diagnosis in a cohort of prostate cancer survivors. *Support Care Cancer*, 27, 2819-2828. <https://doi.org/10.1007/s00520-018-4578-2>

Wazlawick, R. S. (2009). *Metodologia de pesquisa para ciência da computação*. Elsevier.