Caracterização da População Brasileira que sofre de Catarata-Hipertensão - Um Estudo Baseado na Base de Dados PNS 2019

Anna Luiza Silva[1], **Luis E. Zárate**[1]

[1]Curso de Ciência de Dados e Inteligência Artificial
Laboratório de Inteligência Computacional Aplicada - LICAP
Pontificia Universidade Católica de Minas Gerais (PUC-Minas)
CEP — 30140-100 — Belo Horizonte — MG — Brasil

annaluizabh10@gmail.com, zarate@pucminas.br

Abstract. *This study investigates the relationship between hypertension and cataracts, two conditions with high prevalence and significant public health impact. Cataracts account for 51% of blindness cases worldwide, while hypertension affects 45% of the adult Brazilian population. Using data from the 2019 National Health Survey, the study analyzed three groups: individuals with no diagnosis, individuals with hypertension, and those with both hypertension and cataracts. Machine learning techniques were applied to create predictive models aimed at early diagnosis and personalized treatment. Two models were evaluated: Random Forest, with an accuracy of 72%, and Decision Tree, with 68%, highlighting the variable "last medical consultation"as the most relevant.*

Resumo. *Este estudo investiga a relação entre hipertensão arterial e catarata, condições com alta prevalência e impacto na saúde pública. A catarata é responsável por 51% dos casos de cegueira no mundo, enquanto a hipertensão afeta 45% da população adulta brasileira. Utilizando dados da Pesquisa Nacional de Saúde de 2019, o estudo analisou três grupos: pessoas sem nenhum diagnóstico, hipertensas e pessoas com hipertensão e catarata. Técnicas de aprendizado de máquina foram aplicadas para criar modelos preditivos visando diagnóstico precoce e personalização do tratamento. Dois modelos foram avaliados: RandomForest, com acurácia de 72%, e Árvore de Decisão, com 68%, destacando a variável "última consulta médica"como a mais relevante.*

1. Introdução

A catarata e a hipertensão arterial são importantes desafios de saúde pública pela alta prevalência e impacto na qualidade de vida. Segundo a OMS, a catarata causa 51% dos casos de cegueira no mundo e afeta cerca de 20 milhões de pessoas com 550 mil novos casos anuais somente no Brasil (SBO). Já a hipertensão atinge 33% da população mundial e 45% dos adultos brasileiros, sendo fator de risco para doenças cardiovasculares e oculares, como a retinopatia hipertensiva e a catarata.

Estudos indicam que existe comorbidade entre hipertensão e doenças oculares, como o glaucoma e a catarata. O estudo JPHC-NEXT (Japão, 2013–2015) aponta a relação entre pressão arterial sistólica e pressão intraocular. Outros estudos destacam a

associação da hipertensão com fatores como colesterol LDL e diabetes tipo 2, e muitos deles reforçarem o papel do estilo de vida na prevenção dessas doenças.

Na computação, técnicas de aprendizado de máquina (AM) têm sido utilizadas para prever casos de catarata [Ishii et al. 2021], [Ranran et al. 2024] e hipertensão [Santhanam and Ahima 2019], [Hirohiko et al. 2024], com relevantes resultados aplicando regressão logística, SVM, floresta aleatória e aprendizado profundo. No entanto, a análise conjunta dessas condições ainda é pouco explorada [Nunez et al. 2022].

Este trabalho busca através do metodologia CAPTO identificar os principais fatores associados à comorbidade catarata-hipertensão na população brasileira, utilizando dados da PNS 2019 do IBGE [IBGE 2020], distinguindo três grupos de indivíduos: saudáveis, com hipertensão apenas, e com catarata e hipertensão simultaneamente.

2. Trabalhos Relacionados

A relação entre doenças sistêmicas e condições oculares, como o aumento da pressão intraocular (PIO) e a catarata, tem sido amplamente estudada. O estudo JPHC-NEXT mostrou que a pressão arterial sistólica (PAS) e diastólica (PAD) estão positivamente associadas à PIO, com maior impacto da PAS. Isso reforça a importância do controle da pressão arterial para prevenir complicações oculares, como o glaucoma.

No caso da catarata, há evidências de associação com doenças como hipertensão e diabetes. [Tomoyo et al. 2021] identificou que PAD elevada, colesterol LDL e controle glicêmico inadequados aumentam o risco de catarata em pacientes com diabetes tipo 2, especialmente com início precoce e progressão acelerada. A revisão de [Ang and Afshari 2021] também apontou a catarata como comum em distúrbios metabólicos, como obesidade e hipertensão, reforçando a importância de mudanças no estilo de vida.

Modelos de aprendizado de máquina (AM) vêm sendo aplicados na predição de catarata e hipertensão [Santhanam and Ahima 2019], [Hirohiko et al. 2024]. Em [Zafar et al. 2023], técnicas supervisionadas como regressão logística e redes neurais foram usadas para prever catarata relacionada à idade com base em variantes genéticas, com destaque para a regressão logística. Já [Lin et al. 2020] desenvolveu um modelo com floresta aleatória e boosting para diagnosticar catarata congênita em recém-nascidos, alcançando alta precisão.

Apesar dos avanços, ainda há lacunas na aplicação conjunta de dados sobre pressão arterial e catarata em modelos de AM. A integração desses dados pode aprimorar a predição da catarata, especialmente em populações vulneráveis, e contribuir para diagnósticos precoces e intervenções personalizadas.

3. Materiais e Métodos

3.1. Descrição da Base de Dados

Este estudo utiliza a Pesquisa Nacional de Saúde (PNS) 2019, conduzida pelo IBGE [IBGE 2020], que reúne informações demográficas, condições de saúde, uso de serviços médicos, estilo de vida e doenças crônicas. A base possui 293.726 registros e 1.088 atributos, dos quais 5.191 correspondem a indivíduos diagnosticados simultaneamente com catarata e hipertensão arterial.

O objetivo é investigar fatores associados à coexistência dessas condições. Para isso, foi realizada uma análise exploratória considerando região, estado e faixa etária. Não foram identificados padrões relevantes por localização geográfica, mas observou-se maior incidência da comorbidade a partir dos 60 anos, definindo o recorte populacional adotado.

Dessa forma, o estudo concentra-se em indivíduos com 60 anos ou mais, organizados em três grupos: (1) saudáveis (2.419 instâncias), (2) com ambas as condições (863) e (3) com hipertensão apenas (1.970).

3.2. Entendimento do Problema e seleção conceitual de atributos

Para o entendimento do domínio do problema e a seleção conceitual de atributos, utilizou-se o método CAPTO [Gonçalves et al. 2024], fundamentado no Modelo Espiral do Conhecimento [Kuriakose et al. 2010]. Esse método busca integrar o conhecimento tácito de especialistas com o conhecimento explícito proveniente de literatura, relatórios técnicos e dicionários de dados. A partir dessa integração, constrói-se um Modelo Conceitual (MC) unificado, que orienta a identificação dos atributos mais relevantes para projetos de ciência de dados.

O MC é estruturado em três níveis: dimensões (que representam diferentes perspectivas do domínio), aspectos (fatores específicos dentro de cada dimensão) e atributos (variáveis potenciais ligadas aos aspectos). Esses atributos são, então, vinculados às fontes de dados disponíveis, funcionando como um filtro para selecionar variáveis consistentes com os objetivos do estudo.

No presente trabalho, as dimensões identificadas foram: a) hábitos de alimentação; b) hábitos de saúde; c) genética; d) condições socioeconômicas; e) idade e condições físicas; f) exposição ambiental; e g) condições de trabalho. Cada uma dessas dimensões foi associada a aspectos relevantes extraídos da literatura, conforme indicado na Tabela 1.

Com base nesse processo, foram identificados 100 atributos relacionados à comorbidade catarata-hipertensão, posteriormente vinculados à base PNS 2019. Esses atributos compõem o conjunto de dados utilizado nas próximas etapas de preparação e análise. A descrição detalhada das variáveis da PNS encontra-se no dicionário de dados [IBGE 2020].

3.3. Pré-Processamento e preparação de dados

Após a seleção conceitual dos atributos, o conjunto de dados passou por um processo de preparação de dados. Foram mantidos apenas os registros com entrevistas completas (*V0015* = 1) referentes aos grupos populacionais definidos para o estudo. Atributos categóricos, como Gênero (*C006*), "Gravidez"(*P005*),"Menopausa"(*R028*), "Exposição ao sol"(*M011031*), "Uso de óculos"(*G033*), "Diabetes"(*Q03001*), e "Artrites"(*Q079*), foram recodificados em binários utilizando a técnica label encoding, no qual a coluna recebe 1 para quando o entrevistado se aplica à questão, e 0 caso contrário.

Para atributos que expressam frequência/intensidade, foram aplicados mapas de valores, simplificando a escala original. Por exemplo, a variável "Limitação física"(*G083*) foi transformada em uma escala de 0 (nenhuma limitação) a 3 (não consegue de modo algum), e o "Estado de saúde"(*J001*) foi reescalonado de 1 a 5 para 0 a 3. O mesmo

Domínio de problema: Comorbidade Catara - Hipertensão		
Dimensão: Hábitos de alimentação		
Aspectos	**Atributos**	**Atributos Mapeados**
Injestão de Alimentos	Qualidade, Frequência, Tipo, Quantidade	Módulo P – Estilos de Vida: P6a até P26a
Injestão de Bebidas	Qualidade, Frequência, Tipo, Quantidade	Módulo P – Estilos de Vida: P6b até P24a
Dimensão: Condição Física		
Aspectos	**Atributos**	**Atributos Mapeados**
Alteração Hormonal	Alteração Hormonal, Gravidez, Hormônio Feminino, Exames Clínicos, Tratamentos hormonais	Módulo Q – Doenças Crônicas: P5 e Q30b; Módulo R - Saúde da Mulher: R025 a R030
Alteração Metabólica	Exames Clínicos	Indisponível
Fragilidade Ossea/Muscular	Diagnóstico de Artrite, Osteoporose	Módulo Q – Doenças Crônicas: Q079 a Q083
Deficiência	Diagnóstico de deficiência intelectual ou física	Módulo G – Pessoas com Deficiências
Dimensão: Genética		
Aspectos	**Atributos**	**Atributos Mapeados**
Histórico Familiar	Histórico familiar de catarata ou hipertensão	Indisponível
Pré-Disposição Individual	Diagnóstico de outras doenças relacionadas a hipertensão e catarata, como diabetes, AVC, obesidade, etc	Módulo Q – Doenças Crônicas
Dimensão: Características do Indivíduo		
Aspectos	**Atributos**	**Atributos Mapeados**
Gênero	Sexo	Módulo C – Características gerais dos moradores: C6
Idade	Idade	Módulo C – Características gerais dos moradores: C7 e C8
Raça	Raça	Módulo C – Características gerais dos moradores: C009
Dimensão: Condições de Saúde		
Aspectos	**Atributos**	**Atributos Mapeados**
Uso de Serviços de Saúde	Motivo de saúde que requereu o uso dos serviços de saúde, Diagnóstico médico, Última consulta, Uso de medicamentos	Módulo J - Utilização dos serviços de saúde e Módulo Q – Doenças Crônicas: J4a, J7, J11a, J14, J15a, Q32a, Q33b, Q34c, Q38a3 e Q39a
Consumo de Drogas	Quantidade, Frequência, Tipo	Módulo P – Estilos de Vida: P27, P28a, P29, P50, P54, P56, P67 e P67a
Atividade Física	Tipo, frequência, Tempo de uso de dispositivos eletrônicos (Sedentarismo)	Módulo P – Estilos de Vida: P34, P35, P37 e P36; Módulo H - Atendimento médico
Descanso	Duração do sono, Cochilos durante o dia, Estresse, Problemas para dormir	Módulo N - Percepção do estado de saúde: N010 a N016; Módulo H - Atendimento médico: H024
Dimensão: Condições de Trabalho		
Aspectos	**Atributos**	**Atributos Mapeados**
Trabalho	Tipo de Trabalho, Horário de Trabalho, Ambiente de trabalho, Níveis de estresse do trabalho	Módulo E - Características de trabalho das pessoas de 14 anos ou mais de idade
Dimensão: Condições Socioeconômicas		
Aspectos	**Atributos**	**Atributos Mapeados**
Instabilidade Financeira	Tipo, Renda	Módulo E – Características de trabalho das pessoas 14 anos ou mais de idade e Módulo F – Rendimentos de outras fontes: E16, E18 e F1a até F14a
Apoio Social	Auxilio de unidades públicas de saúde	Módulo I- Cobertura de Plano de Saúde: I1a, I1b, I4
Acessibilidade	Acesso a água potável, Acesso a alimentos de qualidade, Acessibilidade a postos e hospitais, Acesso a atividade física	Módulo A-Informações do Domicílio : A5a a A9a
Dimensão: Exposição Ambiental		
Aspectos	**Atributos**	**Atributos Mapeados**
Exposição Ambiental	Exposição a luz solar excessiva, Exposição a poluíção, Exposição a toxinas	Módulo M - Características do trabalho e apoio social: M11

Tabela 1. Dimensões e aspectos para Comorbidade Catarata - Hipertensão"

procedimento foi aplicado a outras variáveis, como "Tempo de uso de dispositivos eletrônicos"(*P04502*), "Última consulta médica"(*J01101*), "Problemas de sono"(*N010*), "Consumo de álcool"(*P027*), "Frequência de exercícios físicos"(*P035*), "Consumo de tabaco"(*P050*) e "Última medição de pressão"(*Q00101*), todas ajustadas para escalas mais compactas e adequadas à modelagem.

Valores ausentes foram tratados respeitando o significado de cada atributo. Para atributos como "Sexo", "Peso"e "Diabetes", os registros contendo valores faltantes foram removidos. No caso do atributo "Gravidez", aplicou-se a seguinte regra de imputação:

Se sexo == "Masculino": A coluna "Gravidez"= 0

Se sexo == "Feminino"E "Idade»= 55: "Gravidez"= 0 (considerando que essa é a idade média em que as mulheres entram na menopausa)

Outros casos foram tratados de maneira similar, como a variável Menopausa e a "Frequência de exercício físico". Esta última utilizando como referência o atributo "Atividade física nos últimos 3 meses"(*P034*). Inconsistências e redundâncias foram eliminadas, garantindo a qualidade dos dados. Além disso, o método IQR (Intervalo Interquartil) foi aplicado para identificar e remover outliers dos atributos numéricos como "Peso"e "Idade".

Após o processo de peparação de dados, observou-se que alguns atributos eram irrelevantes, como a "Gravidez", que estava preenchida apenas com 0 (indicando que nenhuma entrevistada do conjunto de dados selecionado estava grávida), e o atributo "Idade", que, por se tratar de um recorte da população de indivíduos com 60 anos ou mais, não traria informações relevantes. Ambas foram, portanto, eliminadas. A Tabela 2, mostra a quantidade de instâncias antes e após pré-processamento do conjunto de dados.

Classe	Quantidade de instâncias inicial	Quantidade de instâncias após processamento
Pessoas sem diagnóstico	2.419	2.351
Pessoas com hipertensão	1.970	1.879
Pessoas com hipertensão e catarata	863	809

Tabela 2. Quantidade de instâncias após pré-processamento dos dados

De forma a avaliar a capacidade discriminatória de cada atributo a_n, foi calculado o valor da entropia com respeito ao atributo classe= *{Sem hipertensão, Com hipertenão, Hipertensão-e-catarata}*.

$$\begin{array}{lcl} D_n & \Rightarrow & \text{Conjunto de valores possíveis para o atributo } n \\ & & D_n = \{d_{n,1}, d_{n,2}, ..., d_{n,|D_n|}\} \\ C & \Rightarrow & \text{Conjunto de classes} \\ P(x) & \Rightarrow & \text{Probabilidade de ocorrer x} \end{array} \quad (1)$$

A Entropia do atributo a_n com respeito ao atributo classe, representado por C, é dado por:

$$
\begin{aligned}
&H_{(a_{n;C})} = \sum_{i=1}^{|D_n|} P_{(x=d_{n,i})} H(d_{n,i}; C) \\
&Onde: \\
&H_{(d_{n,i};C)} = \sum_{j=1}^{|C|} P_{(x=d_{n,i};c_j)} I_{(x=d_{n,i};c_j)} \\
&I_{(x=d_{n,i};c_j)} = \log_{10} \frac{1}{P_{(x=d_{n,i};c_j)}}
\end{aligned}
\tag{2}
$$

O conjunto ordenado de valores das entropias $\mathbb{H}$ calculado é definido como:

$$
\mathbb{H} = \left\{ \; H_{(a_i;C)} \in \mathbb{R} | H_{(a_i;C)} < H_{(a_{i+1};C)} \; \right\} \tag{3}
$$

Para o conjunto de dados, um atributo a_k, correspondente ao valor $minH_{(a_k;C)}$ pode representar um atributo fortemente relevante por estar mais relacionado diretamente com o atributo classe. O atributo correspondente a $maxH_{(a_k;C)}$ pode representar um atributo fracamente relevante para a classificação. Os atributos com entropia entre os valores extremos podem corresponder a atributos relevantes. Por meio de inspeção humana, e com auxílio de especialista de domínio, foi possível avaliar a relevância utilizando o conjunto $\mathbb{H}$. Na Tabela 3 estão relacionados os atributos finais após o pré-processamento e o valor da entropia em relação ao atributo classe calculado.

Atributo	Tipo do dado	Possíveis valores	Entropia
ultima_consulta	Categórico ordinal	'Nunca consultou', 'Até 2 anos', 'Mais de 2 anos'	1,372
ultima_pressao_medida	Categórico ordinal	'Até 1 ano', 'Até 2 anos', '2 anos ou mais'	1,410
trabalho_semana	Dicotômica	'Trabalha', 'Não trabalha'	1,500
exposicao_sol	Dicotômica	'Se expõe', 'Não se expõe'	1,533
estado_saude	Dicotômica	'Bom', 'Ruim'	1,546
tempo_disp_Eletronico	Dicotômica	'Menos de 2h por dia', 'Mais de 2h por dia'	1,547
sexo	Dicotômica	'Masculino', 'Feminino'	1,553
consumo_tabaco	Dicotômica	'Consome', 'Não consome'	1,558
usa_oculos	Dicotômica	'Usa', 'Não usa'	1,566
consumo_alcool	Dicotômica	'Consome mais de uma vez por mês', 'Consome pouco ou nada'	1,567
peso	Dicotômica	'Abaixo de 70kg', '70kg ou mais'	1,572
problemas_sono	Dicotômica	'Tem problema para dormir', 'Não tem problema para dormir'	1,578
exercicio	Dicotômica	'Até 3 vezes na semana', 'Mais de 3 vezes na semana'	1,583
raca	Dicotômica	'Branca', 'Não branca'	1,584

Tabela 3. Análise de Entropia

A partir da análise da entropia foi possível identificar os atributos com maior potencial discriminatório sendo "Última consulta", "Última pressão medida", "Trabalho na semana", e "Exposição ao sol"os de menor valor, sugerindo maior relevância para a distinção entre as classes. Foi confirmado por meio do valor da entropia que o atributo "Idade"está diretamente relacionado com a classe, o que poderia enviesar o modelo, então foi decidido retirá-lo do conjunto de dados.

Atributos menos relevantes: "Consumo de alcool", "Peso", "Problemas de sono", "Exercícios", e "Raça"possuem maior entropia e, portanto, são menos informativos mas podem complementar outros atributos. Os atributos apresentados não apresentam uma relação direta com a classe por terem entropias altas e muito parecidas entre si, alguns mais relevantes do que outros, mas nada que possa trazer um alerta de que pode enviesar

o modelo. Então todos os atributos presentes na Tabela 3 foram mantidos para construção do modelo de aprendizado supervisionado. A base de dados está disponível em: https://github.com/licapLaboratory/DataBase-PNS-Catarata-Hipertensao

4. Experimentos e Análise de Resultados

Foram construídos dois modelos. Um modelo caixa-branca, interpretável, baseado em Árvore de decisão, e um modelo Ensemble, baseado em Floresta aleatória.

Devido ao desbalanceamento de classes, ver Tabela 2, foi aplicado um processo de balanceamento undersampling, ajustando as instâncias de cada classe para 809, correspondente ao número de instâncias da classe minoritária de comorbidade (hipertensão+catarata). Para o process de treinamento e teste foi aplicado a técnica *hold-out*, sendo 70% para treino e 30% para teste.

Para construção dos modelos foi utilizado o software KNIME (ver Figura 1). O *dataflow* mostra os componentes utilizados. Para o modelo baseado em árvore, o fluxo começa com a leitura do conjunto de dados (*Excel Reader*), seguindo pela etapa de divisão de treino correspondente a 70% da base e teste (*Partitioning*) com 30% dos dados, o treinamento (*Decision Tree Learner*), o processo de teste (*Decision Tree Predictor*), e resultados de desempenho do modelo (*Score*).

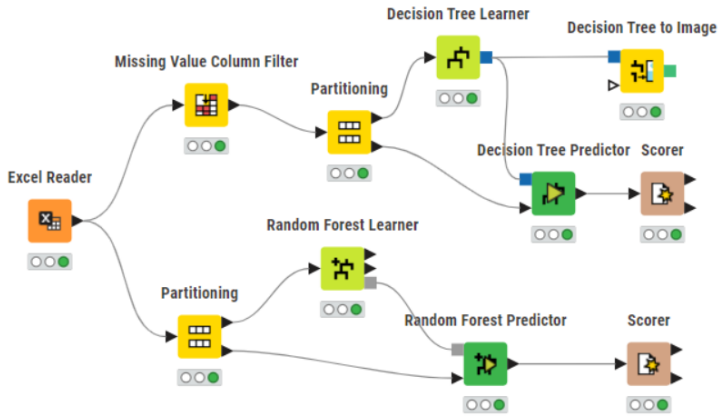

Figura 1. Fluxograma criado no KNIME

Após treinamento e teste dos modelos é possível obter a matriz de confusão e as medidas de desempenho. As Tabelas 4 e 6, correspondem aos resultados do modelo para a árvore de decisão. As Tabelas 5 e 7, mostram os resultados para o modelo baseado em floresta aleatória.

	Hipertensão + Catarata	**Pessoas saudáveis**	**Hipertensão**
Hipertensão + Catarata	172	15	56
Pessoas saudáveis	15	176	52
Hipertensão	33	59	151

Tabela 4. Matriz de confusão árvore de decisão

Os modelos utilizados apresentaram um desempenho geral aceitável, com a Floresta Aleatória se destacando em métricas de desempenho. Uma análise detalhada das

	Hip. + Catarata	Pessoas saudáveis	Hipertensão
Hipertensão + Catarata	182	11	50
Pessoas saudáveis	10	203	30
Hipertensão	36	63	144

Tabela 5. Matriz de confusão floresta aleatória

	Precisão	Sensibilidade	Especificidade	F-medida
Hipertensão + Catarata	0.782	0.708	0.901	0.743
Pessoas saudáveis	0.704	0.724	0.848	0.714
Hipertensão	0.583	0.621	0.778	0.602

Tabela 6. Resultados e acurácia árvore de decisão

	Precisão	Sensibilidade	Especificidade	F-medida
Hipertensão + Catarata	0.798	0.749	0.905	0.773
Pessoas saudáveis	0.733	0.835	0.848	0.781
Hipertensão	0.643	0.593	0.835	0.617

Tabela 7. Resultados e acurácia floresta aleatória

matrizes de confusão e das métricas sugere que ambos os modelos conseguem classificar corretamente a maioria dos casos, mas ainda enfrentam dificuldades para a classe "Hipertensão", indicando que provavelmente não foram identificados atributos inerentes às pessoas que sofrem unicamente de hipertensão.

A Árvore de Decisão apresentou uma medida F1 média de 68,6%. Teve um melhor desempenho na classe comorbidade "Hipertensão+catarata"(74,3%), mas com pior desempenho para a classe "Hipertensão"(60,2%). A Floresta Aleatória teve uma medida F1 média de 72,4%, com destaque para "Pessoas saudáveis"(Sensibilidade de 83,5% e medida F1 de 78,1%). O modelo demonstrou maior precisão para a classe "Hipertensão+catarata", sendo mais eficaz para este grupo em comparação com a Árvore de Decisão. Os atributos de maior relevância para esse modelo foram: “Última_consulta”, seguida por “Última_pressao_medida”, “Peso” e “Sexo”.

4.1. Regras Geradas

Modelos caixa preta baseados em Floresta Aleatória são eficazes, mas não são interpretáveis. Em contraste, árvores de decisão permitem extrair regras claras de classificação. Neste estudo, foram geradas 175 regras, das quais 58 apresentaram 100% de acurácia. A seguir, destacam-se algumas regras que caracterizam indivíduos com hipertensão e com a comorbidade hipertensão + catarata.

A *Regra 1* Indivíduos não expostos ao sol, sem problemas de sono, com uso moderado de dispositivos eletrônicos, não fumantes, aposentados, e com acompanhamento médico regular foram classificados como hipertensão + catarata (97,5% de cobertura; 357 registros).

A *Regra 2* Pessoas ainda ativas no trabalho, sem exposição ao sol, que avaliam sua saúde como ruim e realizam consultas médicas frequentes também foram classificadas como hipertensão + catarata (64,7% de cobertura; 22 registros).

A *Regra 3* A presença de hipertensão isolada foi associada a indivíduos que não trabalham, consomem tabaco e usam óculos (73% de cobertura; 66 registros).

Regra 4 Indivíduos com hipertensão, ainda em atividade laboral, sem exposição

ao sol e com boa autoavaliação de saúde apresentaram 80% de acurácia (108 registros).

Também foram identificadas regras menores (7 a 20 registros) com 100% de cobertura. No total, geraram-se 36 regras para hipertensão + catarata, 73 para hipertensão e 66 para pessoas saudáveis, com médias de cobertura de 68,8%, 82,8% e 83%, respectivamente. O foco do modelo não é preditivo, mas descritivo — visando caracterizar indivíduos com a comorbidade e distingui-los dos demais grupos.

As regras indicam que pessoas com hipertensão + catarata têm hábitos relativamente saudáveis e acompanhamento médico, enquanto indivíduos apenas com hipertensão, embora também monitorados, podem manter hábitos de risco, como o tabagismo. Essa observação sugere que o diagnóstico de doenças crônicas estimula práticas de autocuidado. Para confirmar essa hipótese, estudos longitudinais seriam necessários, permitindo avaliar a evolução dos hábitos de vida e seu impacto na saúde ao longo do tempo.

A variável "última consulta"revelou-se um dos atributos mais relevantes nos modelos preditivos, indicando que o tempo desde o último atendimento médico está fortemente associado à presença de catarata e hipertensão. Essa relação sugere que o acompanhamento regular da saúde pode contribuir para o diagnóstico precoce e a prevenção de complicações. Além disso, a análise dessa variável permite refletir sobre implicações clínicas, como a importância de programas de monitoramento contínuo e intervenções direcionadas para indivíduos com maior tempo sem consulta.

5. Conclusão

Este estudo explorou a relação entre hipertensão arterial sistêmica e catarata, condições relevantes para a saúde pública em populações idosas. Utilizando dados da PNS 2019 e técnicas de aprendizado de máquina, identificaram-se padrões e correlações entre essas doenças, com destaque para a maior incidência a partir dos 60 anos.

Foram testados dois modelos preditivos: Árvore de Decisão e Floresta Aleatória. Ambos apresentaram bom desempenho, com a Floresta Aleatória atingindo 72,6% de acurácia, superior à Árvore de Decisão (68,4%). Os atributos mais relevantes foram "última consulta", "última pressão medida", "peso"e "sexo", indicando a importância do acompanhamento médico no diagnóstico precoce. Apesar dos bons resultados, houve dificuldades na classificação da classe "hipertensão", sugerindo a necessidade de refinar ou adicionar atributos mais específicos da doença ao modelo.

Os resultados deste estudo podem ser utilizados para apoiar políticas públicas voltadas à prevenção e controle da catarata e hipertensão, orientando campanhas de saúde e programas de triagem direcionados a grupos de risco, especialmente indivíduos com 60 anos ou mais. Além disso, os modelos preditivos desenvolvidos podem ser incorporados a sistemas de apoio à decisão em saúde, auxiliando profissionais na identificação precoce de pacientes com maior probabilidade de apresentar a comorbidade, otimização de recursos e planejamento de intervenções preventivas mais efetivas. Como próximos passos, recomenda-se incluir variáveis sobre hábitos de vida, histórico familiar e exames, além de testar outros algoritmos de aprendizado de máquina.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e da Fundação de Ampara à

Pesquisa do Estado de Minas Gerais (FAPEMIG), Processo PCE-00349-25.

Referências

Ang, M. J. and Afshari, N. A. (2021). Cataract and systemic disease: A review. *Clinical & Experimental Ophthalmology*, 49(2):118–127.

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

Hirohiko, K., Koshimizu, H., Nakamura, K., and Okuno, Y. (2024). Recent developments in machine learning modeling methods for hypertension treatment. *Hypertension Research*, 47(3):700–707.

IBGE (2020). Pesquisa nacional de saúde 2019 - instituto brasileiro de geografia e estatística. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=25921&t=resultados`. Acesso em: 2024-07-15.

Ishii, K., Ryo, A., Takashi, O., Shingo, M., Yuri, F., Hiroshi, M., Keiichi, O., Atsushi, N., Shuhei, Y., Akira, O., and Masaki, T. (2021). Predicting intraocular pressure using systemic variables or fundus photography with deep learning in a health examination cohort. *Sci Rep*, 11.

Kuriakose, K. K., Raj, B., Murty, S., and Swaminathan, P. (2010). Knowledge management maturity models – a morphological analysis. *Journal of Knowledge Management Practice*, 11(3):1–10.

Lin, D., Chen, J., Lin, Z., Li, X., Zhang, K., Wu, X., Liu, Z., Huangc, J., Li, J., Zhu, Y., Chen, C., Zhao, L., Xiang, Y., Guo, C., Wang, L., Liu, Y., Chen, W., and Lin, H. (2020). A practical model for the identification of congenital cataracts using machine learning. *eBioMedicine*.

Nunez, R., Harris, A., Szopos, M., Rai, R., Keller, J., Wikle, C., Robinson, E. L., Lin, M., Zou, D., Verticchio, A., Siesky, B. A., and Guidoboni, G. (2022). Clarifying the roles of high and low blood pressure in glaucoma via physiology-informed machine learning. *Invest. Ophthalmol. Vis. Sci.*, 63(7).

Ranran, C., Jinming, L., Yujie, L., Yiping, J., Xue, W., Hong, L., Yanlong, B., and Haohao, Z. (2024). Machine learning models for predicting 24-hour intraocular pressure changes: A comparative study. *Med Sci Monit.*, 3(30).

Santhanam, P. and Ahima, R. (2019). Machine learning and blood pressure. *J Clin Hypertens*.

Tomoyo, Y., Akiko, H., Kazumasa, Y., Kenya, Y., Miki, U., Yoko, O., Mariko, S., Kazuo, T., Norie, S., Kazuno, N., Shoichiro, T., and Hiroyasu, I. (2021). Hypertension and hypercholesterolemia are associated with cataract development in patients with type 2 diabetes. *High Blood Press Cardiovasc Prev*, 28:475–481.

Zafar, S., Khurram, H., Kamran, M., Fatima, M., Parvaiz, A., and Shaikh, R. S. (2023). Potential of gja8 gene variants in predicting age-related cataract: A comparison of supervised machine learning methods. *PLOS One*.