CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

# Mineração de dados no diagnóstico de hipertensão baseado na Pesquisa Nacional em Saúde 2019

# Data mining in the diagnosis of hypertension based on National Health Survey 2019

# Minería de datos en el diagnóstico de hipertensión con base en la Encuesta Nacional de Salud 2019

Nicolau Machado de Carvalho[1], Marco Paulo Soares Gomes[2], Luis Enrique Zárate[2]

1 Bac., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil
2 Dr., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil

Autor correspondente: Prof. Dr. Luis Enrique Zárate
E-mail: zarate@pucminas.br

**Resumo**

A hipertensão é uma doença que atinge grande parte da população brasileira. Por ser uma doença muito comum, alguns de seus fatores de risco são conhecidos, mas conhecer a ordem de relevância pode trazer novos insights, principalmente quando o objetivo é o diagnóstico da doença. Recentemente foi disponibilizada a Pesquisa Nacional em Saúde 2019, que traz novas informações sobre a saúde da população brasileira. O Objetivo é auxiliar no diagnóstico dos indivíduos que sofrem de Hipertensão Arterial Sistêmica por meio de um método para descoberta de conhecimento e classificação por Floresta Aleatória. Resultados alcançaram um F1-score médio de 75%. As conclusões apontam que a ingestão de sal, manter-se fora do peso ideal, não praticar atividades físicas moderadas, e fumar, nessa ordem, são fatores muito importantes para diagnóstico da doença.

**Descritores:** Hipertensão; Mineração de dados; Descoberta de Conhecimento.

**Abstract**

Hypertension is a disease that affects a large part of the Brazilian population. As it is a very common disease, some of its risk factors are known, but knowing the order of relevance can bring new insights, especially when the objective is to diagnose the disease. The 2019 National Health Survey was recently made available, which provides new information about the health of the Brazilian population. The objective is to assist in the diagnosis of individuals suffering from Systemic Arterial Hypertension through a method for discovering knowledge and classification by Random Forest. Results achieved an average F1-score of 75%. The conclusions indicate that salt intake, staying outside the ideal weight, not practicing moderate physical activities, and smoking, in that order, are very important factors for diagnosing the disease.

**Keywords:** Hypertension; Data Mining; Knowledge Discovery.

**Resumen**

La hipertensión es una enfermedad que afecta a gran parte de la población brasileña. Al tratarse de una enfermedad muy común, se conocen algunos de sus factores de riesgo, pero conocer el orden de importancia puede indicar nuevos conocimientos, sobre todo cuando el objetivo es diagnosticar la enfermedad. Recientemente se puso a disposición la Encuesta Nacional de Salud de 2019, que proporciona nuevas informaciones sobre la salud de la población brasileña. El objetivo es ayudar en el diagnóstico de personas que sufren Hipertensión Arterial Sistémica a través de un método de descubrimiento de conocimiento y clasificación por Random Forest. Los resultados alcanzaron una puntuación media F1 del 75%. Las conclusiones indican que el consumo de sal, mantenerse fuera del peso ideal, no practicar actividades físicas moderadas y fumar, en ese orden, son factores muy importantes para el diagnóstico de la enfermedad.

**Descriptores:** Hipertensión; Mineria de datos; Descubrimiento de Conocimiento.

## Introdução

A análise da saúde de uma população desempenha um papel fundamental na identificação e compreensão de problemas de saúde pública, permitindo que os órgãos públicos de saúde adotem medidas preventivas e estratégias de intervenção. Dentro desse contexto, a Hipertensão Arterial Sistêmica (HAS) é uma doença cardiovascular crônica, multifatorial, caracterizada pelos níveis de pressão arterial persistentemente elevados. A prevalência estimada da doença é de 32% na população adulta brasileira, sendo que essa proporção aumenta com a idade. Os níveis elevados da pressão arterial conferem um aumento significativo do risco de eventos cardiovasculares, como infarto, acidente vascular encefálico (AVC) e insuficiência cardíaca, muitos destes preveníveis pelo tratamento precoce e adequado da doença.

O tratamento e controle da hipertensão arterial envolve mudanças no estilo de vida, como dieta saudável, exercícios físicos regulares, redução do consumo de álcool e tabagismo, e em alguns casos, é necessário o uso de medicamentos para controlar a pressão arterial.

No contexto da Ciência de Dados, especialmente na área da saúde, o processo de descoberta de conhecimento em banco de dados (Data Mining) é bastante requisitado, pois permite analisar dados acerca de um domínio de problema, numa perspectiva multifatorial, a partir de algoritmos computacionais de aprendizado.

Existem diversos trabalhos que têm procurado diagnosticar e caracterizar o perfil de pessoas que possuem hipertensão por meio de técnicas de Aprendizado de Máquina. O artigo de Elshawi e colaboradores[1] aplica o classificador Floresta Aleatória para identificar diversos fatores que caracterizam a hipertensão para uma específica população adulta na China. O artigo[2] busca identificar qual é o melhor algoritmo para se utilizar na predição do diagnóstico de hipertensão arterial. Os autores compararam os algoritmos: K-Vizinhos mais próximos (KNN), Support Vector Machine, Árvores de Decisão e Naive Bayes.

Diferentemente dos outros trabalhos citados, LaFreniere e colaboradores[3] relatam a utilização de um modelo baseado em redes neurais artificiais com dois objetivos: prever a presença de hipertensão na população Canadense, e entender os perfis de indivíduos

que podem se tornar hipertensos. Em relação aos trabalhos que buscam prever o diagnóstico de hipertensão, o artigo[4] busca prever estágios diferentes da doença, se o paciente está no estágio de pré-hipertensão, hipertensão nível 1 ou hipertensão nível 2.

Com a exploração da base de dados PNS 2019 (https://www.pns.icict.fiocruz.br/) como objetivo deste trabalho busca-se compreender os fatores que caracterizam o perfil do indivíduo hipertenso, e para isso, foi aplicada uma metodologia para descoberta de conhecimento em bases de dados. Como técnica de aprendizado é utilizado o classificador Floresta Aleatória e por meio da análise do modelo foi possível apontar alguns fatores importantes relacionados à doença presentes na população brasileira.

## Materiais e Métodos

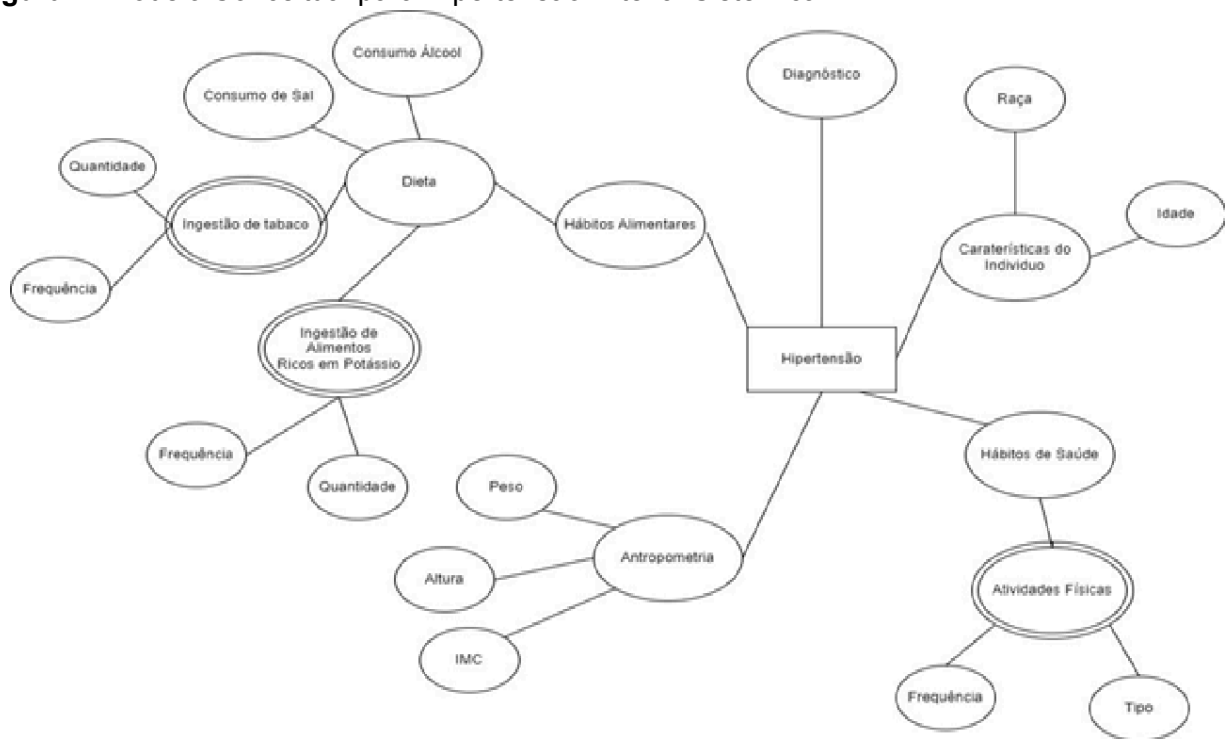
A Pesquisa Nacional de Saúde (PNS) é um estudo realizado no Brasil pelo Instituto Brasileiro de Geografia e Estatística (IBGE). O objetivo da PNS é fornecer informações sobre a saúde pública da população brasileira, incluindo aspectos relacionados às comorbidades, acesso aos serviços de saúde, estilo de vida, uso de medicamentos, prática de atividade física, entre outros aspectos que podem estar ligados à saúde do indivíduo. A base de dados considerada possui 293.726 registros e 1.088 atributos organizados em 26 módulos de questões. A seguir, serão apresentadas as principais etapas da metodologia adotada para a descoberta de conhecimento.

**Etapa 1: Entendimento do domínio de problema:** Em relação a esta etapa, é ressaltada a necessidade de entendimento do domínio da aplicação e da importância do conhecimento a priori acerca desse domínio.

Devido à grande quantidade de informações (definidos como atributos no contexto da Ciência de Dados) disponíveis na base de dados PNS, é proposta uma seleção conceitual dos principais atributos que podem contribuir para traçar o perfil de pessoas hipertensas. Para isto, foi considerado o método CAPTO[5] o qual utiliza do conhecimento explícito (baseado na literatura médica) e conhecimento tácito (baseado na contribuição de especialistas de domínio) para construção de um Modelo conceitual. O conhecimento extraído é estruturado por meio de dimensões (perspectivas do domínio) e aspectos

relevantes dentro de cada dimensão. O modelo conceitual é apresentado na Figura 1 e as principais dimensões/perspectivas e aspectos associados à HAS foram: Hábitos alimentares (dieta, ingestão de tabaco, ingestão de alimentos ricos em potássio), Hábitos de saúde (atividade física) antropometria, e características do indivíduo (raça e idade), etc.

**Figura 1-** Modelo Conceitual para Hipertensão Arterial Sistêmica




**Etapa 2: Seleção conceitual de atributos:** A partir do modelo conceitual e da intervenção do especialista, foram selecionados os atributos relacionados com o modelo conceitual e da base original PNS. Na Tabela 1 são indicados os atributos extraídos (31) da base original PNS.

**Etapa 3: Fusão e transformação de atributos:** Antes de aplicar o processo de categorização (discretização) dos atributos, alguns atributos foram fundidos ou criados a partir de cálculos contendo combinação de atributos. A Tabela 2 mostra os atributos envolvidos no processo de função e categorização. Para justificar o procedimento de fusão de atributos, a literatura científica foi consultada, como indicado a seguir:

a) **CONSUMO_POTASSIO:** Baseado na literatura, foi considerado que frutas, suco de frutas, verduras, legumes e peixe são as fontes mais ricas de potássio. Dentro da PNS não existe uma variável que represente o consumo de potássio. Daí, optou-se por agrupar grupos de alimentos.

*CONSUMO_POTASSIO_SEMANA = P00901 + P015 + P01601 + P018*

**Tabela 1 -** Atributos da PNS selecionados a partir do modelo conceital

| Atributo-Questão | Descrição do atributo |
|---|---|
| **C008** | Idade |
| **C009** | Cor ou raça |
| **P00104** | Peso |
| **P00404** | Altura |
| **P00901** | Quantidade de dias da semana que costuma comer pelo menos um tipo de verd/legume |
| **P015** | Quantidade de dias da semana que costuma comer peixe |
| **P01601** | Quantidade de dias da semana que costuma tomar suco natural |
| **P018** | Quantidade de dias da semana que costuma comer frutas |
| **P02601** | Consumo de sal |
| **P02801** | Quantidade de dias da semana de consumo de bebida alcoólica |
| **P035** | Quantidade de dias da semana que costuma praticar exercícios ou esporte |
| **P03701** | Tempo de duração em horas dos exercícios físicos ou esporte |
| **P03702** | Tempo de duração em minutos dos exercícios físicos ou esporte |
| **P03904** | Número de dias da semana que anda bastante a pé ou faz atividade pesada |
| **P03905** | Tempo de duração em horas que anda bastante a pé ou faz atividade pesada |
| **P03906** | Tempo de duração em minutos que anda bastante a pé ou faz atividade pesada |
| **P04001** | Número de dias da semana que faz trajeto a pé ou de bicicleta, considerando ida e volta |
| **P04101** | Tempo de duração em horas que faz trajeto a pé ou de bicicleta, considerando ida e volta |
| **P04102** | Tempo de duração em minutos que faz trajeto a pé ou de bicicleta, considerando ida e volta |
| **P042** | Nas atividades habituais. Número de dias da semana de deslocamento a pé ou de bicicleta |
| **P04301** | Na atividade habitual. Tempo de duração em hora de deslocamento a pé ou de bicicleta |
| **P04302** | Na atividade habitual. Tempo de duração em minutos de deslocamento a pé ou de bicicleta |
| **P050** | Atualmente fuma algum produto de tabaco |
| **P051** | No passado fumou algum produto de tabaco |
| **P053** | Idade ao início do fumo |
| **P05403** | Consumo semanal de cigarros industrializados |
| **P05406** | Consumo semanal de cigarros de palha ou enrolados a mão |
| **P05802** | Na tentativa de parar de fuma. Média de consumo de cigarros industrializados |
| **P05901** | Há quanto tempo (em anos) parou de fumar |
| **Q00201** | Diagnóstico clínico de pressão alta |
| **Q00202** | HAS apenas durante período de gravidez |

b) **EXERCICIO_FISICO_MOD:** Foi considerado como exercício físico moderado, ao exercício físico realizado habitualmente (semanalmente) no trabalho, ida ao trabalho, ou outra atividade a pé ou de bicicleta. A discretização da variável

ocorreu baseada na recomendação de exercício físico não intenso de acordo com a referência.

$$MIN\_SEMANA\_EXERC\_FISICO\_MOD = P03904 * [(P03905 * 60) + P03906] + P04001 * [(P04101 * 60) + P04102] + P042 * [(P04301 * 60) + P04302]$$

**Tabela 2 -** Atributos Selecionados e envolvidos na fusão e categorização

| Dimensões/Atributos | Descrição | Regra de discretização |
|---|---|---|
| **RACA_PRETA_PARDA_INDIGENA** | Indicador binário se o entrevistado é de alguma dessas etnias (estudos mostram que são etnias que podem ter maior propensão a hipertensão) | Variável C009 é igual a 2, 4 ou 5 o indicador é 1, caso contrário 0 |
| **NIVEL_CONSUMO_POTASSIO** | Nível em que o entrevistado ingere potássio ao longo da semana | Dividido em 5 níveis baseado na soma das variáveis P00901, P015, P01601 e P018 |
| **EXERCICIO_FISICO_INTENSO** | Nível em que o entrevistado se exercita fisicamente de forma intensa na semana | Dividido em 5 níveis baseado nas variáveis P035, P03701 e P03702 |
| **EXERCICIO_FISICO_MOD** | Nível em que o entrevistado se exercita fisicamente de forma moderada na semana | Dividido em 5 níveis baseado nas variáveis P03904, P03905, P03906, P04001, P04101, P04102, P042, P04301, P04302 |
| **CONSUMO_SAL** | Nível em que o entrevistado consome sal durante suas refeições | Resposta da variável P02601 |
| **NIVEL_CONSUMO_ALCOOL** | Nível em que o entrevistado consome álcool durante a semana | Nível em que o entrevistado consome álcool durante a semana |
| **FAIXA_ETARIA** | Faixa etária do entrevistado | Dividido em 6 faixas etárias baseado na variável C008 |
| **NIVEL_IMC** | Nível de Índice de Massa Corporal do entrevistado | Cálculo de IMC através das variáveis P00104 e P00404, dividido em 5 níveis |
| **NIVEL_FUMO** | Nível de tabagismo do entrevistado | Dividido em 5 faixas etárias baseado nas variáveis P050, P051, P053, P05901 P05403, P05406, P05802 |
| **DIAGNOSTICO_OFICIAL** | Se a pessoa teve o diagnóstico positivo de hipertensão | Baseado nas variáveis Q00201 e Q00202 |

**c) EXERCICIO_FISICO:** A fusão dos atributo ocorreu baseado na recomendação de exercício físico intenso do [Ministério da Saúde 2004]:

$$MIN\_SEMANA\_EXERCICIO\_FISICO = P035 * [(P03701 * 60) + P03702]$$

d) **NIVEL_FUMO:** O atributo foi fundido da seguinte maneira:

$$TEMPO\_FUMO = C008 - P053 \; TEMPO\_FUMOU = C008 - P053 - P05901 \; QTD\_CIGARROS\_POR\_DIA =$$

*P05402 + P05405*

**e) IMC:** Foi criado o atributo IMC (índice de massa corpórea) a partir da altura e peso das pessoas:

*IMC = P00104 / (P00404*P00404)*

**Etapa 4: Categorização de atributos:** De forma a construir modelos com maior interpretabilidade alguns atributos foram submetidos a um processo de categorização, transformando alguns atributos em categóricos ordinais.

A discretização do atributo CONSUMO_POTASSIO_ SEMANA ocorreu de acordo com a seguine regra:

*Se.. CONSUMO_POTASSIO_SEMANA <= 2 → Nível 1*
*Se.. 2 < CONSUMO_POTASSIO_SEMANA <= 5 → Nível 2*
*Se.. 5 < CONSUMO_POTASSIO_SEMANA <= 7→Nível 3*
*Se.. 7 < CONSUMO_POTASSIO_SEMANA <= 9→ Nível 4*
*Se.. 9 < CONSUMO_POTASSIO_SEMANA → Nível 5*

A discretização do atributo MIN_SEMANA_EXERC_ FISICO_MOD ocorreu da seguinte forma:

*Se.. MIN_SEMANA_EXERC_FISICO_MOD <= 60 → Nível 1*
*Se.. 60 < MIN_SEMANA_EXERC_FISICO_MOD <= 120 → Nível 2*
*Se.. 120 < MIN_SEMANA_EXERC_FISICO_MOD <= 240 → Nível 3*
*Se.. 240 < MIN_SEMANA_EXERC_FISICO_MOD <= 300 → Nível 4*
*Se.. 300 < MIN_SEMANA_EXERC_FISICO_MOD → Nível 5*

A categorização do atributo NIVEL_FUMO seguiu as seguintes regras de discretização:

*Se.. (QTD_CIGARROS_POR_DIA >= 3 AND TEMPO_FUMO >= 20) OR (P05801 >= 3 AND TEMPO_FUMOU >= 20) → Nível 5*
*Se.. (QTD_CIGARROS_POR_DIA >= 3 AND TEMPO_FUMO >= 10 AND TEMPO_FUMO < 20) OR (P05801 >= 3 AND TEMPO_FUMOU >= 10 TEMPO_FUMOU < 20) → Nível 4*
*Se.. (QTD_CIGARROS_POR_DIA < 3 AND TEMPO_FUMO >= 20) OR (P05801 < 3 AND TEMPO_FUMOU>= 20) → Nível 3*
*Se.. (QTD_CIGARROS_POR_DIA < 3 AND TEMPO_FUMO >= 10 AND TEMPO_FUMO < 20) OR (P05801 <3 AND TEMPO_FUMOU >= 10 AND TEMPO_FUMOU < 20) → Nível 2*
*Se.. Não se enquadra em nenhum desses quesitos → Nível 1*

A discretização do atributo EXERCICIO_FISICO_INTENSO ocorreu baseada na recomendação de exercício físico intenso do Ministério da Saúde (https://bvsms.saude.gov.br/hipertensao-18):

*Se.. MIN_SEMANA_EXERCICIO_FISICO <= 30 → Nível 1*

*Se.. 30 < MIN_SEMANA_EXERCICIO_FISICO <= 60 → Nível 2*
*Se.. 60 < MIN_SEMANA_EXERCICIO_FISICO <= 120 → Nível 3*
*Se.. 120 < MIN_SEMANA_EXERCICIO_FISICO <= 200 → Nível 4*
*Se.. 200 < MIN_SEMANA_EXERCICIO_FISICO → Nível 5*

Para o atributo NIVEL_CONSUMO_ALCOOL, a discretização ocorreu baseado no recomendação da OMS:

*P02801 = 1 → Nível 1*
*P02801 = 2 → Nível 2*
*P02801 = 3 → Nível 3*
*P02801 IN (4,5) → Nível 4*
*P02801 IN (6,7) → Nível 5*

Para o atributo FAIXA_ETARIA, a discretização ocorreu dividindo de forma a distribuir equilibradamente o número de registros para cada faixa. A variável foi discretizada da seguinte forma:

*C008 < 25 → Nível 1*
*C008 >= 25 AND C008 <= 35 → Nível 2*
*C008 > 35 AND C008 <= 45 → Nível 3*
*C008 > 45 AND C008 <= 55 → Nível 4*
*C008 > 55 AND C008 <= 69 → Nível 5*
*C008 >= 70 → Nível 6*

O NIVEL_IMC foi discretizado baseado em referência da Fundação Oswaldo Cruz 2014 (https://centrodeobesidadeediabetes.org.br):

*IMC < 20 => 1 abaixo do peso*
*IMC >= 20 AND IMC <= 25 => 2 peso normal*
*IMC > 25 AND IMC <= 30 => 3 sobrepeso*
*IMC > 30 AND IMC <= 35 => 4 obesidade grau I*
*IMC > 35 AND IMC <= 40 => 5 obesidade grau II*
*IMC > 40 => 6 obesidade grau III*

Não houve necessidade do tratamento de dados ausentes, pois o número de dados ausentes foi baixo para o conjunto de atributos selecionados. Registros contendo dados ausentes foram retirados do conjunto de dados. Não houve necessidade do tratamento de outliers após a discretização realizada na preparação dos dados. A discretização já é uma forma de suavizar *outliers*, já que os valores muito fora do padrão acabam entrando em um nível da discretização, mantendo eles dentro do padrão, mesmo que se encaixem em um nível onde há menor quantidade de registros.

Após o processamento adotado, foram mantidos 88.737 registros da base original. Para o balanceamento de classes, foi aplicado o procedimento random undersampling da classe majoritária (registro de indivíduos que não são diagnosticados com hipertensão) para igualar as classes em 22.519 entrevistados para cada classe (com e sem HAS), totalizando 45.038 instâncias no conjunto de dados após o pré-processamento dos dados.

**Etapa 5: Modelo computacional de aprendizado:** Foi utilizado o classificador Floresta Aleatória (FA) com os melhores hiperparâmetros selecionados pela técnica random search. Os hiperparâmetros selecionados foram: Quantidade de árvores = 40, Número mínimo de instâncias necessárias para dividir um nó interno da árvore = 5, Número mínimo de instâncias exigido para ser considerado um nó folha da árvore = 30, Máximo de atributos utilizados = 5, Profundidade máxima da árvore = 7.

**Etapa 6: Validação Cruzada:** O conjunto de dados completo foi dividido em 80% para treinamento e 20% para teste. Para treinamento foi aplicado o procedimento de validação cruzada com k-fold = 10.

## Resultados experimentais

Resultados do treinamento indicaram que para um intervalo de confiança de 95%, a medida F1-score encontra-se no intervalo [0.74,0.76]. A Tabela 3, destaca o resultado para o conjunto de testes (20% do conjunto de dados original).

**Tabela 3 – Resultados do procedimento de teste**

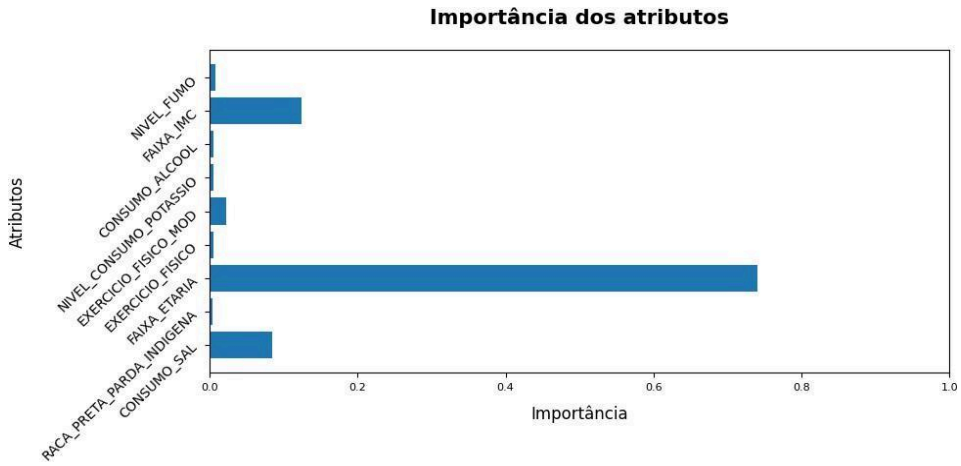
| Classe | Precisão | Sensibilidade | F1-score |
|---|---|---|---|
| **Sem HAS: 0** | 0,78 | 0,70 | 0,74 |
| **Com HAS: 1** | 0,73 | 0,80 | 0,76 |

O modelo testado apresentou resultados efetivos como um F1-score médio de 75% para o conjunto de teste e uma sensibilidade média de 75%, sendo que a sensibilidade do valor da classe que representa os hipertensos foi igual a 80%. No Gráfico 1, podemos observar a importância de cada atributos para o modelo utilizando o classificador Random Forest.

Pode-se observar que os atributos mais importantes para que o modelo consiga classificar as classes (1: hipertenso e 0: não hipertenso) são a faixa etária, o IMC, o consumo do sal, o exercício físico moderado, o tabagismo. Por outro lado, com menor importância estão o consumo de álcool, consumo de potássio, exercícios físicos intensos, e etnia. A importância da faixa etária pode ser evidenciada pelos próprios dados da

Sociedade Brasileira de Cardiologia, que dizem que mais de 70% da população brasileira de 70 anos ou mais tem diagnóstico de hipertensão. Outro ponto observado pelo Gráfico 1 é que pessoas com alto consumo de sal e pessoas com índice de massa corporal elevada têm maior tendência a serem hipertensas, explicando a importância desses outros atributos. O exercício físico moderado é praticado por muito mais pessoas que o exercício físico intenso, por isso apresenta uma maior importância, sendo que pessoas menos sedentárias têm menos chance de serem hipertensas. Mais um fato interessante é que o atributo Raça apresenta a menor importância para a classificação feita pelo modelo entre os atributos da base. O atributo foi incluído baseado em alguns estudos que estabelecem que pessoas que não se declaram brancas têm maior dificuldade de controlar e acompanhar a pressão arterial, porém este estudo não evidenciou este possível causador.

**Gráfico 1 -** Importância dos atributos



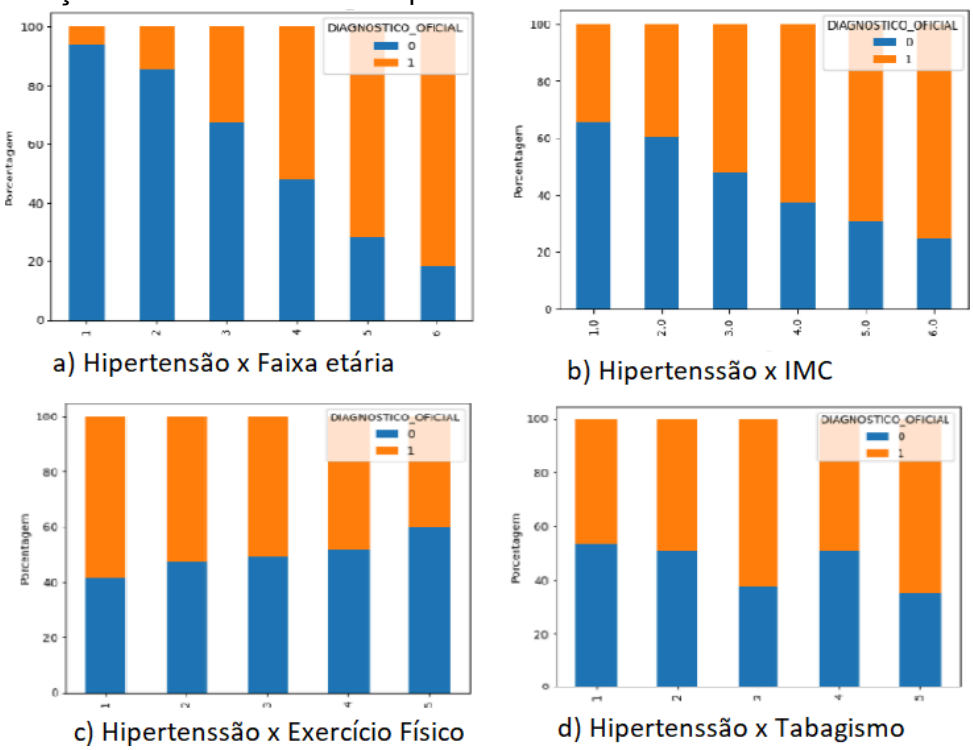

No Gráfico 2a podemos observar que há uma relação de aumento dos casos de diagnóstico positivo (classe = 1) conforme aumenta a faixa etária. As faixas etárias representadas por 5 e 6, onde há maior concentração de hipertensos, indicam pessoas de 55 ou mais anos de idade, idades em que a maioria da população brasileira apresenta quadro de hipertensão.

Como é evidenciado no Gráfico 2b, o IMC é um indicador de massa corporal e pessoas com sobrepeso têm uma maior chance de serem hipertensas. A partir da faixa

representada pelo número 3, considera-se as pessoas com algum tipo de sobrepeso e pode-se observar que é a partir dessa faixa que o número de hipertensos ( classe =1) aumenta em relação ao número de não hipertensos (classe= 0).

O Gráfico 2c mostra que quanto mais praticante do exercício físico moderado, menor a propensão a ser hipertenso, já que essas pessoas provavelmente não são sedentárias e a prática de atividade física é importante para o controle da pressão arterial.

**Grafico 2 -** Relações encontradas com a hipertensão




O tabagismo pode influenciar na pressão arterial. Um fato interessante que o Gráfico 2d mostra, é que o tempo de fumo do indivíduo se relaciona com maior assertividade sobre o diagnóstico de hipertensão do que a quantidade de cigarros fumados, já que os níveis 3 e 5 apresentam indivíduos que fumam há mais de 20 anos, embora o nível 3 fume menos cigarros por dia que o nível 4.

De forma a descrever o perfil de hipertensos da população brasileira a partir do PNS 2019, aplicamos a técnica de árvore de decisão (algoritmo J48, Weka, com poda). As regras obtidas permitem descrever o perfil e apontar hipóteses que podem ser

investigadas posteriormente. A Tabela 5, mostra as regras de decisão obtidas. A partir das regras, é possível observar, que além dos perfis esperados, como a maioria de indivíduos com idade menor a 45 anos não apresentarem HAS e maiores de 70 anos apresentam a doença; indivíduos entre 45 e 55 anos, podem não apresentar HAS, se o peso for normal ou abaixo do peso. Indivíduos entre 55 e 69 anos, com peso normal e consumo alto ou muito alto de sal podem não apresentar HAS. Esta é uma hipótese que deveria ser analisada.

**Tabela 5 – Regras obtidas da árvore de decisão**

| | |
|---|---|
| FAIXA_ETARIA Faixa_Etária<25: N (3892.0/196.0) | FAIXA_ETARIA 55<Faixa_Etária<=69 |
| FAIXA_ETARIA 25<=Faixa_Etária<=35: N (7683.0/963.0) | \| FAIXA_IMC = Sobrepeso: S (5001.18/1545.79) |
| FAIXA_ETARIA 35<Faixa_Etária<=45: N (8681.0/2429.0) | \| FAIXA_IMC = Obesidade grau II: S (669.16/106.11) |
| FAIXA_ETARIA 70<Faixa_Etária: S (7920.0/1835.0) | \| FAIXA_IMC = Obesidade grau I: S (2414.57/507.38) |
| FAIXA_ETARIA 45<Faixa_Etária<=55 | \| FAIXA_IMC = Obesidade grau III: S (185.04/12.03) |
| \| FAIXA_IMC = Sobrepeso | \| FAIXA_IMC = Peso normal |
| \| \| CONSUMO_SAL = Adequado: N (2040.8/879.0) | \| \| CONSUMO_SAL = Adequado |
| \| \| CONSUMO_SAL = Baixo: S (1089.4/467.0) | \| \| \| RACA_PRETA_PARDA_INDIGENA = Não: S (783.0/361.0) |
| \| \| CONSUMO_SAL = Alto: S (309.8/145.0) | \| \| \| RACA_PRETA_PARDA_INDIGENA = Sim: N (1225.62/563.31) |
| \| \| CONSUMO_SAL = Muito Baixo: S (189.0/78.0) | \| \| CONSUMO_SAL = Baixo: S (1408.31/531.31) |
| \| \| CONSUMO_SAL = Muito alto: N (45.0/21.0) | \| \| CONSUMO_SAL = Alto: N (221.0/103.0) |
| \| FAIXA_IMC = Obesidade grau II: S (543.3/151.12) | \| \| CONSUMO_SAL = Muito baixo: S (235.0/71.0) |
| \| FAIXA_IMC = Obesidade grau I: S (1706.93/640.37) | \| \| CONSUMO_SAL = Muito alto: N (35.0/16.0) |
| \| FAIXA_IMC = Obesidde grau III: S (174.09/37.04) | \| FAIXA_IMC = Abaixo do peso |
| \| FAIXA_IMC = Peso normal: N (2766.51/922.9) | \| \| CONSUMO_SAL = Adequado: N (271.08/102.04) |
| \| FAIXA_IMC = Abaixo do peso: N (319.17/88.1) | \| \| CONSUMO_SAL = Baixo: S (179.04/70.04) |
| | \| \| CONSUMO_SAL = Alto: N (27.0/13.0) |
| | \| \| CONSUMO_SAL = Muito baixo: S (41.0/16.0) |
| | \| \| CONSUMO_SAL = Muito alto: S (4.0/1.0) |

## Conclusão

Observando os resultados do modelo, conclui-se que ele classificou corretamente a grande maioria dos registros. Porém, poucos dos atributos selecionados para classificar a presença de hipertensão foram importantes para a classificação do indivíduo como hipertenso. Atributos como raça, consumo de potássio, consumo de álcool foram quase irrelevantes para o modelo. Pode-se concluir que controlar o consumo de sal, praticar

atividades físicas e manter-se próximo ao seu peso ideal são importantíssimos fatores para não apresentar um quadro de hipertensão durante a vida. Ainda assim, é muito possível que com idade mais elevada, existam problemas com a pressão arterial. Apesar da efetividade do modelo ser aceitável para descrever o perfil de pessoas hipertensas, a base de dados da PNS tem seu questionário voltado principalmente para a oportunidade de realizar políticas públicas de saúde por parte do governo. Portanto, não é uma base de dados que apresenta informações tão precisas, e em alguns casos, como no atributo de consumo do potássio, que foi mencionado anteriormente neste artigo, apresenta limitações para determinar certos indicadores. Embora a base de dados PNS não seja perfeita para a modelagem de dados, e as informações não sejam tão precisas, já que para analisar doenças é necessário avaliar dados durante a vida, e a maioria dos dados da base são apenas do momento atual, ainda é possível construir modelos de dados a partir dela. Para outras análises de indicadores da saúde a PNS poderia ser mais efetiva.

## Referências

1. Malta, D. C., Bernal, R. T. I., Ribeiro, E. G., Moreira, A. D., Felisbino-Mendes, M. S., & Velásquez-Meléndez, J. G. (2022). Arterial hypertension and associated factors: National Health Survey, 2019. Revista De Saúde Pública, 56, 122.

2. Elshawi, R., Al-Mallah, M. H., Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. BMC medical informatics and decision making 19 (1): 1–32, 2019.

3. Kublanov, V. S., Dolganov, A. Y., Belo, D., and Gamboa, H. Comparison of machine learning methods for the arterial hypertension diagnostics. Applied bionics and biomechanics vol. 2017.

4. LaFreniere, D., Zulkernine, F., Barber, D., and Martin, K. Using machine learning to predict hypertension from a clinical dataset. In 2016 IEEE symposium series on computational intelligence (SSCI). IEEE, pp. 1–7, 2016.

5. Zarate, L., Petrocchi, B., Maia, C., Felix, C., and Gomes, M. P. Capto- a method for understanding problem domains for data science projects. Concilium 23:922–941, 2023.