Análise dos Fatores Associados ao Diagnóstico da Hipertensão Arterial: Um Estudo Utilizando Algoritmos de Aprendizado de Máquina e Regressão logística

Lucas Dirk Gomes Ferreira

Instituto de computação - Universidade Federal Fluminense (UFF)

Niterói, Rio de Janeiro

1. Introdução

De acordo com a OMS, a hipertensão arterial é uma condição de saúde que afeta milhões de pessoas em todo o mundo, sendo considerada um importante fator de risco para doenças cardiovasculares, como o acidente vascular cerebral (AVC) e a insuficiência renal crônica.

Com isso, compreender os fatores que influenciam o diagnóstico da hipertensão arterial é fundamental para o desenvolvimento de estratégias eficazes de prevenção e diagnóstico precoce.

Ao investigar essas variáveis, espera-se obter informações sobre os fatores que influenciam na incidência do diagnóstico positivo da hipertensão arterial. Objetivando-se contribuir para o aprimoramento de estratégias de prevenção e controle da doença.

1.1 Objetivo

Neste contexto, o objetivo geral deste estudo é investigar os fatores associados ao diagnóstico da hipertensão arterial, em uma amostra representativa da população brasileira, através de algoritmos de aprendizado de máquina e regressão logística, buscando compreender os determinantes dessa condição de saúde. Para tanto, foram selecionadas diversas variáveis de estudo, oriundas da pesquisa nacional de saúde, incluindo informações demográficas, comportamentais, de saúde e estilo de vida. Através da aplicação desses algoritmos, foi possível realizar previsões da ocorrência da hipertensão arterial em indivíduos, contribuindo para uma melhor compreensão e identificação dos fatores de riscos associados a essa condição de saúde.

2. Fonte de dados

De acordo com o Governo Federal, a pesquisa nacional de saúde (PNS) é um inquérito de saúde realizado pelo Ministério da Saúde em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE).

No presente estudo, foram utilizadas informações referentes à pesquisa de 2019 da PNS. Essa pesquisa abrange uma variedade de tópicos relacionados à saúde da população, incluindo condições de saúde, acesso aos serviços de saúde, estilos de vida, etc.

Com essas informações, a PNS oferece uma visão abrangente sobre a situação de saúde da população brasileira, fornecendo dados importantes para formulação de políticas de saúde.

2.1 Variáveis de estudo

No presente estudo, foram selecionadas algumas variáveis com o objetivo de investigar sua influência no diagnóstico da hipertensão arterial. As variáveis selecionadas são as seguintes:

1) **V0001 - Unidade da Federação**
2) **V0026 - Tipo de situação censitária**
3) **C006 - Sexo**
4) **C008 - Idade**
5) **C009 - Cor ou Raça**
6) **D001 - Se sabe ler ou escrever**
7) **D00901 - Escolaridade**
8) **N004 - Quando o(a) Sr(a) sobe uma ladeira, um lance de escadas ou caminha rápido no plano, sente dor ou desconforto no peito ?**
9) **N005 - Quando o(a) Sr(a) caminha em lugar plano, em velocidade normal, sente dor ou desconforto no peito**

10) N010 - Nas duas últimas semanas, com que frequência o(a) Sr(a) teve problemas no sono, como dificuldade para adormecer, acordar frequentemente à noite ou dormir mais do que de costume?

11) N011 - Nas duas últimas semanas, com que frequência o(a) Sr(a) teve problemas por não se sentir descansado(a) e disposto(a) durante o dia, sentindo-se cansado(a), sem ter energia ?

12) P00104 - Peso

13) P00404 - Altura

14) P01001 - Em geral, o(a) Sr(a) costuma comer esse tipo de verdura ou legume:

15) P019 - Em geral, quantas vezes por dia o(a) Sr(a) come frutas?

16) P02002 - Em quantos dias da semana o(a) Sr(a) costuma tomar refrigerante?

17) P02102 - Que tipo de refrigerante o(a) Sr(a) costuma tomar?

18) P034 - Nos últimos três meses, o(a) Sr(a) praticou algum tipo de exercício físico ou esporte?

19) P035 - Quantos dias por semana o(a) Sr(a) costuma (costumava)praticar exercício físico ou esporte?

20) P038 - No seu trabalho, o(a) Sr(a) anda bastante a pé?

21) P039 - No seu trabalho, o(a) Sr(a) faz faxina pesada, carrega peso ou faz outra atividade pesada que requer esforço físico intenso?

22) P03905 - Em um dia normal, quanto tempo o(a) Sr(a) passa andando bastante a pé ou realizando essas atividades pesadas ou que requerem esforço físico no seu trabalho?Horas

23) P040 - Para ir ou voltar do trabalho, o(a) Sr(a) faz algum trajeto a pé ou de bicicleta?

24) P050 - Atualmente, o(a) Sr(a) fuma algum produto do tabaco?

25) P051 - E no passado, o(a) Sr(a) fumou algum produto do tabaco diariamente?

26) **P05402 - Quantos por dia de tabaco o senhor fumava?**

27) **Q00201 - Algum médico já lhe deu o diagnóstico de hipertensão arterial (pressão alta)?**

28) **Q03001 - Algum médico já lhe deu o diagnóstico de diabetes?**

29) **Q060 - Algum médico já lhe deu o diagnóstico de colesterol alto?**

30) **Q06306 - Algum médico já lhe deu o diagnóstico de uma doença do coração, tal como infarto, angina, insuficiência cardíaca ou outra?**

31) **Q068 - Algum médico já lhe deu o diagnóstico de AVC (Acidente Vascular Cerebral) ou derrame?**

32) **Q124 - Algum médico já lhe deu o diagnóstico de insuficiência renal crônica?**

33) **P02602 - Em quantos dias da semana o(a) Sr(a) costuma substituir a refeição do almoço por lanches rápidos como sanduíches, salgados, pizza, cachorro quente, etc?**

34) **P02601- Considerando a comida preparada na hora e os alimentos industrializados, o(a) Sr(a) acha que o seu consumo de sal é**

Além das variáveis mencionadas, foram criadas duas novas variáveis para este estudo: O índice de Massa Corporal (IMC) e a Região Geográfica.

A primeira variável, IMC, é uma medida que estabelece a relação entre peso e altura de um indivíduo. O cálculo do IMC é realizado dividindo-se o peso e a altura ao quadrado de um indivíduo.

A variável Região Geográfica se refere à localização geográfica em que o participante reside. Essa variável foi introduzida para investigar possíveis diferenças nas prevalências de hipertensão arterial entre as diferentes regiões do país.

3. Metodologia

3.0 Análise exploratória de dados

A análise exploratória de dados é uma etapa crucial para entender a estrutura e os padrões dos dados antes da modelagem. Nela esperamos encontrar e tratar valores ausentes e outliers,compreender as relação das variáveis e obter uma visão geral do conjunto de dados. Essas informações ajudam na tomada de decisões sobre o pré-processamento dos dados e a seleção de técnicas de modelagem adequadas.

3.1 Regressão logística

A regressão logística é uma técnica amplamente utilizada em estatística e ciência de dados para resolver problemas de classificação. Este método, é um modelo estatístico usado para modelar e analisar relações entre variáveis dependentes binárias.

No modelo de regressão logística, a curva logística é ajustada aos dados permitindo que se calcule a probabilidade de ocorrência de um evento de interesse (Silveira et al., 2021), temos que a sua função é dada por:

$F(x) = 1/(1 + e^{-Z})$, onde f(x) assume valores entre 0 e 1 e z valores entre -∞ a +∞.

3.2 Árvore de decisão

As árvores de decisão são um tipo de algoritmo de aprendizado de máquina que utiliza as características dos dados para sua tomada de decisões, utilizando uma estrutura de árvore. Para realização das divisões, durante sua construção, o algoritmo utiliza de medidas como entropia para terminar qual melhor caminho a ser tomado.

$Entropia = -p_1 * log_2(p_1) - ..... - -p_n log_2(p_n)$, onde p representa a proporção das classes dos dados.

3.3 KNN

O KNN (K-Nearest Neighbors) é um algoritmo que se baseia na distância entre os dados para realizar as classificações.

Ele calcula distância euclidiana entre os pontos de dados e rotula um novo ponto de acordo com a maioria dos seus vizinhos mais próximos. No presente estudo foi utilizado k = 3, ou seja, classifica o novo rótulo de acordo com os três mais próximos.

A distância euclidiana é dada por :

$$D_p = ((x_2 - x_1)^2 + (y_2 - y_1)^2 + \ldots + (z_1 - z_2)^2)^{1/2}$$

3.4 Random Forest

O Random Forest consiste em criar um conjunto de árvores de decisão independentes, onde cada árvore é treinada com uma parte aleatória dos dados, usando somente alguns atributos. Durante a previsão, as árvores fornecem uma resposta e a classe é determinada pela média ou maioria das respostas individuais das árvores. É um excelente algoritmo para evitar o problema do overfitting.

3.5 SVM

O SVM (Support Vector Machine) é um algoritmo que busca encontrar um hiperplano de separação ótimo que maximize a margem entre as classes. Durante o processo de classificação, o SVM classifica um novo ponto com base na sua posição em relação ao hiperplano. o SVM possui diversos tipos de Kernel para modelagem das relações não lineares dos dados, são tipos de kernel: Linear, Polinomial, Radial e Sigmoid.

3.6 Curva ROC

Na regressão logística, é comum a avaliação do desempenho do modelo por meio de métricas como a curva ROC (Receiver Operating Characteristic). A curva ROC é uma representação gráfica onde podemos avaliar a capacidade do modelo na sua classificação.

A curva ROC é construída traçando-se a taxa de verdadeiros positivos no eixo y e a taxa de falsos positivos no eixo x, em diferentes pontos de corte do classificador.

Além disso, uma outra forma de olhar para a curva ROC é olhar a sua área, pois ela fornece uma medida numérica para o desempenho do modelo. A área sob a curva ROC

é um valor que varia de 0 a 1, onde 1 indica um modelo perfeito e um valor próximo a 0.5 indica que o modelo não é melhor que a própria aleatoriedade.

3.7 Stepwise Selection

O método stepwise é uma etapa fundamental na construção de um modelo de regressão logística, pois este método, visa buscar um subconjunto ótimo de variáveis explicativas para inclusão no modelo de regressão.

O método stepwise é realizado em etapas, onde são feitas adições e remoções sequenciais de variáveis com base no seu p-valor. Na adição, é testado a inclusão de cada variável no modelo e avalia-se se a sua adição melhora o ajuste do modelo. Na remoção, as variáveis já presentes são testadas e avalia-se se a sua remoção melhora o ajuste. O processo age de forma de iterativa até que nenhum critério de adição ou remoção seja atendido.

3.8 Undersampling

Undersampling é uma técnica de amostragem muito utilizada na regressão logística e algoritmos classificadores para abordar o desequilíbrio entre as classes.

Esta técnica consiste em reduzir o número de instâncias da classe majoritária, com o intuito de equilibrar a proporção entre as classes. Para isso, as instâncias da classe majoritária são removidas aleatoriamente até que a proporção entre as classes sejam equilibrada.

3.9 Especificidade e Sensibilidade

A especificidade é uma medida que indica a capacidade do modelo em identificar corretamente os casos negativos. Representa a proporção de casos negativos corretamente classificados em relação ao total de casos negativos reais.

A sua fórmula é dada por:

Especificidade = Verdadeiros Negativos / (Verdadeiros Negativos + Falsos Positivos)

Por fim, a sensibilidade é uma medida que nos indica a capacidade do modelo de identificar corretamente os casos positivos. Representa a proporção de casos positivos corretamente identificados pelo modelo em relação ao total de casos positivos reais.

A sua fórmula é dada por:

Sensibilidade = Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Negativos).

3.10 Normalização

A normalização de dados foi realizada visando garantir que as variáveis apresentassem uma escala comparável e adequada para a análise da regressão logística. A presente técnica transforma os dados de forma que se tenha média zero e desvio padrão igual a um.

3.11 SOFTWARES E PACOTES UTILIZADO

Para a realização deste estudo, foram utilizados os softwares R e Python, juntamente com várias bibliotecas específicas, que desempenharam um papel fundamental na obtenção dos dados e na análise estatística.

3.11.1 R

No ambiente R, foram utilizadas as seguintes bibliotecas:

PNSIBGE - Permitiu o acesso aos dados do Programa Nacional De Amostra De Domicílios (PNAD), disponibilizados pelo Instituto Brasileiro De Geografia e Estatística.
Readr - Utilizada para ler e importar os dados em formato CSV.

3.11.2 Python

No ambiente Python, foram utilizadas as seguintes bibliotecas:

Pandas - Desempenhou papel na manipulação e análise dos dados.
Scikit-learn - Utilizado para construção da regressão e as suas métricas.
Matplotlib e Seaborn - Utilizada para visualização dos dados.

3.12 Validação Cruzada

A validação cruzada do tipo k-fold é uma técnica de validação em que o conjunto de dados é dividido, treino e teste, em seguida há o treinamento do modelo por diversos algoritmos. O processo é repetido k vezes, neste presente estudo k igual a 10, no final

as métricas de desempenho são calculados pela média dos resultados e desvio padrão dos resultados obtidos em cada iteração.

3.13 Teste Qui-Quadrado

O teste qui-quadrado é um teste estatístico que permite verificar a independência entre as variáveis.

A hipótese nula deste teste consiste em dizer que não há associação entre as variáveis testadas. Ou seja, são independentes.

O valor p é um indicador da significância estatística do teste. Se o valor-p for menor que um nível de significância pré determinado, podemos rejeitar a hipótese nula e concluir que existem evidências estatísticas para dizer que há uma associação significativa entre as variáveis.

4. Análise dos resultados

Nesta seção, será apresentado todo pré - processamento da base utilizada até os resultados dos algoritmos utilizados.

4.1 - Análise exploratória dos dados e Pré Processamento dos dados.

Inicialmente, verificou-se a quantidade de registros na base de dados selecionada. O arquivo possuía um total de 279 382 registros. A variável alvo, denominada ‘Q00201’, foi renomeada como ‘Hipertensão’. Dentre esses registros, 23 851 apresentavam hipertensão, 64 855 não e 190 646 tinham valores ausentes. Os valores ausentes foram removidos, resultando em uma base de dados com 88 736 registros.

Tabela 1: Distribuição de frequência da variável hipertensão.

Variáveis	n(%)
Hipertensão	
Sim	23851 (8.5)
Não	64885 (23.2)
N.A	190646 (68.2)

Fonte: Elaborado pelo autor com base nos dados da PNS 2019

Em seguida, foram criadas duas novas variáveis “Região Geográfica” e “IMC”. A variável “Região Geográfica” foi derivada da coluna “Estado”. A variável “IMC” foi calculada utilizando as informações de Altura e Peso de cada participante. Posteriormente, as colunas “Estado”, “Altura” e “Peso” foram removidas do conjunto de dados.

Logo depois, as colunas do conjunto de dados foram renomeadas para tornar os nomes mais intuitivos e compreensíveis. Os nomes das colunas foram substituídos por descrições mais claras e informativas, em vez de manter os códigos de identificação originais da pesquisa PNS.

Os dados faltantes foram verificados e observou-se que as variáveis referentes ao uso de cigarro (Se fumou no mês passado e quantos cigarros fumava por dia) possuem mais de 90% do seu valor absoluto em dados faltantes. Portanto, essas variáveis foram removidas do dataset.

Em relação às variáveis com valores ausentes remanescentes, foi adotada a estratégia de substituir todos os valores NA por -99. Essa abordagem permitiu identificar de forma clara quais observações originalmente possuíam valores ausentes, mantendo a integridade do conjunto de dados.

A tabela 2 apresenta a análise de associação entre a variável hipertensão e outras variáveis patológicas investigadas no estudo. Ao analisar os resultados, podemos observar que as variáveis diabetes, colesterol alto, doenças do coração e AVC demonstram dependência muito forte com a presença de hipertensão. Logo, podemos concluir que a presença de uma delas aumenta a probabilidade de ocorrência de hipertensão.

Variáveis	P-Valor	Teste Qui-Quadrado
Teste Qui-Quadrado hipertensão X:		
Diabetes	0	6645.4
Colesterol Alto	0	3470.1
Doença do coracao	0	1866.04
AVC	0	448.4

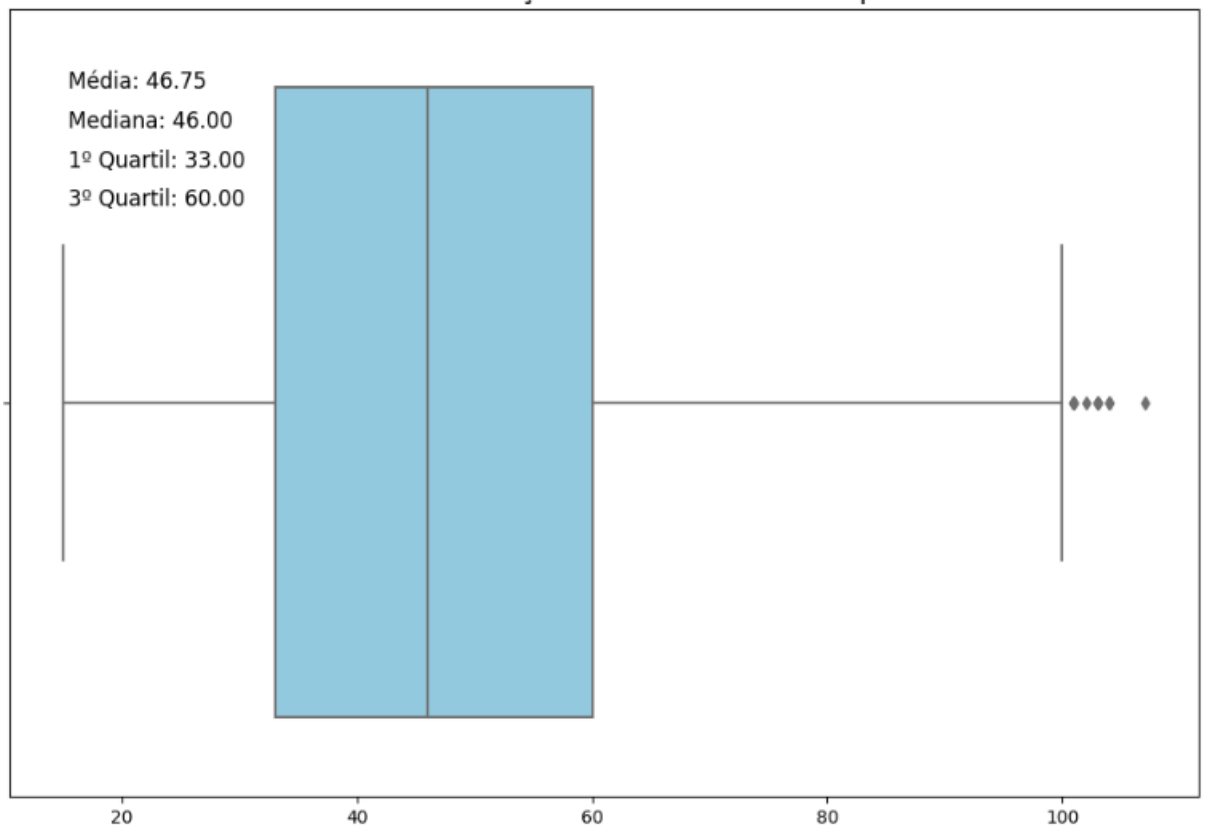
Fonte: Elaborado pelo autor com base nos dados da PNS 2019

Uma análise crucial em nosso estudo é a avaliação da distribuição das idades em nosso conjunto de dados, uma vez que é amplamente conhecido na literatura que a idade e a hipertensão estão correlacionadas. É essencial garantir que tenhamos uma distribuição equilibrada de idades em nossa amostra, a fim de obter resultados confiáveis e não enviesados.

Com base na análise do gráfico de boxplot abaixo, podemos observar que a distribuição de idade no conjunto não apresenta um viés significativo. As informações fornecidas

pelos quartis destacam que o primeiro quartil (25% dos dados) está em torno dos 33 anos, a mediana está aproximadamente 46 anos e o terceiro quartil (75%) está em torno dos 60 anos. Esses valores indicam uma distribuição equilibrada das idades, abrangendo um intervalo considerável.

Grafico 1: Distribuição da Idade dos Participantes



Fonte: Elaborado pelo autor com base nos dados da PNS 2019

Todas as variáveis categóricas da base de dados foram convertidas em variáveis quantitativas discretas para serem utilizadas no modelo de classificação. Essa transformação é necessária, pois a maioria dos algoritmos de aprendizado de máquina trabalham somente com variáveis quantitativas.

Agora verificando a associação entre as variáveis Hipertensão e Sexo, através da tabela de contigência abaixo, observamos que existem mais homens com hipertensão do que mulheres na nossa amostra, olhando proporcionalmente, observamos que 30.53% dos homens são hipertensos enquanto apenas 22% das mulher são hipertensas, mostrando uma possível associação entre Sexo e Hipertensão;

Tabela de contigência.

Sexo	Homem	Mulher
Hipertensão:		
Não	32904	31981
Sim	14467	9384
Total	47371	41365

Através da tabela 1, conseguimos observar que os dados apresentam um desequilíbrio significativo. Essa falta de balanceamento é um fator importante a ser considerado na construção de um modelo de classificação, pois a predominância de casos negativos podem afetar a capacidade do modelo de identificar corretamente os casos positivos. Para mitigar esses problemas, foi aplicado a técnica de undersampling, com isso ocorreu a redução da classe majoritária para equilibrar a distribuição entre as classes. Na tabela 3, podemos observar como ficou a nova distribuição dos dados após aplicação dessa técnica.

Distribuição de frequência da variável hipertensão após Undersampling.

Variáveis	n(%)
Hipertensão	
Sim	23851 (50)
Não	23851 (50)
Total	47702 (100)

Fonte: Elaborado pelo autor com base nos dados da PNS 2019

Após a aplicação da técnica de undersampling e a obtenção do conjunto de dados balanceado, foi realizada a etapa de normalização dos dados. A normalização é um processo importante na preparação dos dados, pois visa garantir escalas compatíveis entre as variáveis para não ocorrer influências negativas no resultado dos modelos.

Após a normalização dos dados, foi aplicado o método de seleção stepwise por meio da regressão logística. Este método é uma abordagem que visa selecionar as melhores variáveis para o modelo preditivo, levando em consideração a importância estatística e o impacto das variáveis na capacidade preditiva do modelo.

4.2 - Avaliação dos Modelos

Foi utilizado a seleção stepwise para realização do experimento**.** A técnica de validação cruzada foi aplicada, sendo executada 10 vezes com diferentes sementes aleatórias variando de 1 a 10.

As variáveis selecionadas pelo Stepwise para o experimento foram as: "Se anda no trabalho", "Escolaridade", "Idade", "Consumo de alimentos industriais", "Se sente dor no peito ao caminhar em ladeiras/escadas", "IMC", "colesterol", "Sexo", "Fuma quantos cigarros por dia?", "Se tem doença do coração", "Se já teve AVC", "Se fuma tabaco", "Cor", "Se sente muito cansaço", "Se sente dor no peito ao caminhar em lugares planos", "Situação censitária", "Região", "Se come legumes" e "Diabetes".

Essas variáveis foram utilizadas em conjunto com a técnica de validação cruzada para avaliar o desempenho do modelo. A abordagem com todas as variáveis foi aplicada para fins de comparação.

Foram testados diversos modelos de classificação para realizar o experimento, os modelos utilizados foram: Regressão logística, KNN (K=3), SVM com kernel linear, Random Forest e Decision Tree.

As métricas utilizadas na avaliação foram: Média das acurácias, Médias da Sensibilidade, Média da Especificidade e média da área sob a curva ROC, além disso, também foi calculado os seus desvios padrões. Essas métricas são importantes para avaliação da capacidade de cada modelo, pois fornecem uma visão geral do desempenho do modelo.

Algoritmo	Acurácia	Sensibilidade	Especificidade	AUC-ROC
Decision Tree	0.66(Média)	0.65 (Média)	0.664	0.66
KNN	0.69(Média)	0.64 (Média)	0.716	0.681
Random Forest	0.749(Média)	0.76 (Média)	0.736	0.751
Regressão Logística	0.747(Média)	0.74 (Média)	0.746	0.746
SVM	0.750(Média)	0.75 (Média)	0.741	0.748

Fonte: Elaborado pelo autor com base nos dados da PNS 2019

Ao analisar os resultados dos modelos, verificamos que o Random Forest, a Regressão Logística e o SVM apresentaram as melhores métricas de desempenho. No entanto, considerando a natureza do problema de hipertensão, em que é crucial

identificar corretamente os casos positivos, a sensibilidade média torna-se um critério relevante na escolha do modelo.

Dessa forma, levando em conta a sensibilidade média como medida de performance, o modelo Random Forest se destaca como a melhor opção. Sua capacidade de identificar corretamente os indivíduos com hipertensão é fundamental para um diagnóstico preciso e eficaz. Portanto, optamos pelo uso do modelo Random Forest como a abordagem principal para a detecção de hipertensão neste estudo.

Utilizando o método de Feature importance, conseguimos olhar quais são as variáveis que mais influenciam no modelo do Random Forest. A partir da tabela abaixo conseguimos observar as 10 variáveis que mais influenciam no diagnóstico da hipertensão, é como esperado, a idade é um fator de suma importância para esse diagnóstico, assim como o IMC.

Variável	**Importância**
Idade	0.27
IMC	0.20
Escolaridade	0.08
Região	0.05
Consumo de alimentos indústrias	0.04
Se anda bastante no trabalho	0.04
Cor	0.04
Se come legumes	0.03
Se sente cansaço com muita frequência	0.03

5 Conclusão

Com base nos objetivos do estudo e nos resultados obtidos, foi possível investigar os fatores associados ao diagnóstico da hipertensão arterial em uma amostra representativa da população brasileira. Utilizando algoritmos de aprendizado de máquina e regressão logística, foram analisadas diversas variáveis relacionadas a características demográficas, comportamentais, de saúde e estilo de vida.

Os resultados indicaram o Random Forest como principal modelo para o diagnóstico de hipertensão neste estudo. Além disso, o modelo indicou variáveis de suma importância no diagnóstico da hipertensão, como a idade, o imc, escolaridade, etc. Podendo assim fazer um agrupamento de grupos de riscos.

Esses resultados mostram o potencial do aprendizado de máquina no diagnóstico da hipertensão arterial e compreensão dos fatores de risco. Como pesquisa futura pode ser considerada a inclusão de mais variáveis explicativas, tendo em vista que a pesquisa PNS possui centenas de variáveis que não foram investigadas no presente estudo.

6. Referências

Silveira, M. B. G. da ., Barbosa, N. F. M. ., Peixoto, A. P. B. ., Xavier, Érika F. M. . and Xavier Júnior, S. F. A. (2021) “Application of logistic regression in the analysis of risk factor associated with arterial hypertension”, *Research, Society and Development*, 10(16), p. e20101622964. doi: 10.33448/rsd-v10i16.22964.

Araujo, L. V., Miranda, M. H. da S., Fontenele, M. H. de S., Damasceno Neto, O. F., Batista, J. G., de Lima, A. F., & de Souza, D. A. (2022). Detecção do risco de Diabetes em estágio inicial utilizando redes ELM e seleção de features baseada em algoritmo genético: Early stage Diabetes risk prediction using ELM and ga-based feature selection. *Brazilian Journal of Development*, *8*(7), 54179–54190. https://doi.org/10.34117/bjdv8n7-339.

CARBONI, S. O uso de Árvores de Decisão na descoberta de desconhecimento para saúde. Tese (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul. Porto Alegre, 2003