**APLICAÇÃO DO CLASSIFICADOR K-NEAREST NEIGHBORS (KNN) NA ÁREA DA SAÚDE:
RELAÇÃO CINTURA-QUADRIL E PRESSÃO ARTERIAL****Application of k-Nearest Neighbors classifier (KNN) in healthcare:
Waist-to-hip Ratio (WHR) relationship and blood pressure**

ISSN: 2178-7514

Vol. 13 | Nº. 2 | Ano 2021

Ricardo Pablo Passos^{1,2,3}; Luís Felipe Silio^{1,2,4}; Bráulio Nascimento Lima^{1,2}; José Ricardo Lourenço de Oliveira^{1,2};
Adriano de Almeida Pereira^{1,2}; Gustavo Celestino Martins^{1,6}; Leandro Borelli de Camargo¹; Carlos Henrique Prevital Fileni¹;
Marcelo Francisco Rodrigues^{1,2,5}; Guanís de Barros Vilela Junior^{1,2}

RESUMO

Um dos pontos críticos nos algoritmos de Inteligência Artificial é a eficiência do classificador utilizado nos diferentes cenários de uma rede neural recursiva. O KNN (K-Nearest Neighbors) é um classificador simples capaz de rotular cada elemento de um conjunto de dados a partir da distância destes em relação aos seus vizinhos mais próximos em um conjunto de dados semelhante treinado previamente. Na área da saúde é amplamente conhecido que a Relação Cintura Quadril (RCQ) e a Pressão Arterial (PA) estão fortemente correlacionadas. Objetivo: averiguar o comportamento da acurácia, sensibilidade e precisão do KNN para diferentes valores de k (3, 5 e 7) em relação à identificação de hipertensos em um conjunto de dados de RCQ. Métodos: A amostra é composta por 499 mulheres, usuários do Programa de Saúde da Família (PSF) na cidade de Poços de Caldas, Minas Gerais. Todas as participantes assinaram o Termo de Consentimento Livre e Esclarecido (TCLE). Todas as participantes assinaram o Termo de Consentimento Livre e Esclarecido (TCLE). Para aferição da PA foi utilizado um esfigmomanômetro de coluna de mercúrio (0-300 mmHg) da marca Biotec[®], a medida das circunferências da cintura e do quadril foram realizadas com uma fita antropométrica, com escala em milímetros da marca Sanny[®]. O algoritmo desenvolvido na linguagem Python 3.0 e bibliotecas específicas do mesmo foram utilizadas. Foram utilizados 80% dos dados para o treinamento do algoritmo e 20% para a fase de testagem do mesmo. Resultados: os valores da acurácia, sensibilidade e precisão variaram para os diferentes valores de K testados na KNN. Para K=3 foram obtidos os melhores valores para acurácia (0,89) sensibilidade (0,89) e precisão (0,89) para este classificador de mulheres hipertensas a partir do RCQ das mesmas. Conclusão: o método KNN é um eficiente classificador em redes neurais recursivas aplicadas na área da saúde.

Palavras-chave: Inteligência Artificial, KNN, Saúde.

ABSTRACT

One of the critical points in Artificial Intelligence algorithms is the efficiency of the classifier used in different scenarios of a recursive neural network. The KNN (K-Nearest Neighbors) is a simple classifier capable of labeling each element of a dataset based on their distance from their nearest neighbors in a previously trained similar dataset. In healthcare it is widely known that the Waist-Hip Ratio (RCQ) and Blood Pressure (PA) are strongly correlated. Objective: to investigate the accuracy, sensitivity and precision behavior of KNN for different k values (3.5 and 7) in relation to the identification of hypertensive patients in a WHR dataset. Methods: The sample is composed of 499 subjects of both sexes, users of the Family Health Program (PSF) in the city of Poços de Caldas, Minas Gerais. All participants signed the Informed Consent Form (TCLE). To measure BP, a mercury column sphygmomanometer (0-300 mmHg) from the brand Biotec[®] was used. Waist and hip circumferences were measured with an anthropometric tape, with a scale in millimeters from the brand Sanny[®]. The algorithm developed in Python 3.0 language and its specific libraries were used. 80% of the data were used for training the algorithm and 20% for its testing phase. Results: the accuracy, sensitivity and precision values varied for the different K values tested in KNN. For K=3 the best values for accuracy (0.89) sensitivity (0.89) and precision (0.89) were obtained for this classifier of hypertensive women from their RCQ. Conclusion: the KNN method is an efficient classifier in recursive neural networks applied in healthcare.

Keywords: Artificial Intelligence, KNN, Healthcare.

1 Núcleo de Pesquisas em Biomecânica Ocupacional e Qualidade de Vida (NPBOQV, CNPq)

2 UNIMEP

3 Kroton Anhanguera Educacional

4 Centro Universitário São Lucas/Afya

5 Faculdade mogiana - Unimogi

6 FEUC- Faculdade Euclides da Cunha

Autor de correspondência

Ricardo Pablo Passos - Rppasso@gmail.com

INTRODUÇÃO

Um dos desafios atuais na área da saúde é desenvolver e validar algoritmos utilizados na inteligência artificial (IA), na perspectiva de aprimorar a acurácia e velocidade nos diagnósticos. Neste cenário, o método “K-Nearest Neighbors” (KNN) torna-se uma ferramenta bastante utilizada em procedimentos de classificação e/ou predição de patologias.

É importante a diferenciação entre conceitos que não raro são confundidos na área da saúde são eles: Acurácia, precisão, sensibilidade, F1-Score. Figliola, Beasley ⁽¹⁾.

Acurácia: indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;

Precisão: dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;

Sensibilidade: dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas;

F1-Score: média harmônica entre precisão e sensibilidade, ou seja, trata-se da razão entre o número de elementos e a soma do inverso dos mesmos.

Figura 01: VP = Verdadeiro Positivo, VN = Verdadeiro Negativo, FP = Falso Positivo, FN = Falso Negativo

ACURÁCIA	PRECISÃO	SENSIBILIDADE	F1-SCORE
$\frac{VP + VN}{VP + VN + FP + FN}$	$\frac{VP}{VP + FP}$	$\frac{VP}{VP + FN}$	$\frac{2 (Precisão * Sensibilidade)}{Precisão + Sensibilidade}$

Fonte: Figliola, R. S., Beasley, D. E. (2007). Teoria e projeto para medições mecânicas. Rio de Janeiro: LTC

O KNN pode ser um classificador eficiente para identificação de indivíduos hipertensos a partir da RCQ.

A hipertensão arterial essencial ou primária é caracterizada pelos níveis pressóricos de 140 mmHg e 90 mmHg da pressão arterial sistólica e da pressão arterial diastólica respectivamente, confirmadas por no mínimo duas mensurações em pelo menos dois dias separados, ou pela ingestão de medicação anti-hipertensiva para fins de controle da pressão arterial. A hipertensão

essencial representa 95% de todos os casos relatados e é um fator de risco para doenças cardiovasculares (DCV) e mortalidade prematura. Fatores de riscos não modificáveis (genéticos), e modificáveis (inatividade física e dietas com alto teor de sódio e gordura) contribuem para o desenvolvimento da hipertensão essencial. Os outros 5% dos casos relatados fazem referência a hipertensão secundária. Presença de doença renal crônica, estenose da artéria renal, feocromocitoma, hiperaldosteronismo primário

e apneia do sono representam as suas principais causas ^(2; 3; 4; 5). De acordo com Barroso et al. ⁽⁶⁾, em pesquisa nacional realizada no ano de 2013, 21,4% dos adultos brasileiros auto relataram hipertensão arterial, no entanto, levando em consideração as medidas de pressão arterial mensuradas e uso de medicação anti-hipertensiva, o percentual de adultos com hipertensão arterial (≥ 140 mmHg por 90 mmHg) atingiu 32,3 %. Segundo Melo et al. ⁽⁷⁾, os óbitos por DCV representaram 27,3% de 1.312.623 mortes, sendo responsável por 22,6% das mortes prematuras no Brasil. Em relação aos custos do Sistema Único de Saúde (SUS), com hospitalizações, procedimentos ambulatoriais e medicamentos, a estimativa é que houve um gasto de US \$523,7 milhões.

Existem índices antropométricos que associam de forma direta o excesso de peso corporal dos indivíduos a níveis elevados da pressão arterial. Tran et al. ⁽⁸⁾ reportaram que a Relação Cintura e Quadril (RCQ) maior que 0,95 para homens e maior que 0,85 para mulheres, que caracteriza acúmulo de gordura na região do tronco, demonstra indivíduos mais predispostos a doenças cardiovasculares como a hipertensão arterial sistêmica. Adicionalmente, outros estudos mostraram que a RCQ apresenta alto valor preditivo para complicações da obesidade, inclusive para hipertensão arterial sistêmica ^(9; 10; 11).

O estudo de Lima et al. ⁽¹²⁾ evidenciou a existência de correlação positiva, estatisticamente significativa entre RCQ com pressão arterial, demonstrando dessa forma que, quanto maiores

forem os valores dessa variável antropométrica, maiores serão as medidas da pressão arterial.

Em suma, neste trabalho, será gerado modelo de classificação a partir de características de mulheres (RCQ e pressão arterial) de diferentes faixas etárias. Esse modelo auxiliará na detecção de pressão arterial sistólica e pressão arterial diastólica elevadas sem diagnóstico prévio da doença, podendo gerar um alerta no Sistema Único de Saúde (SUS) para a atenção à paciente. No âmbito desse grupo de mulheres e da pressão arterial, isso melhoraria a relação entre os atendimentos e procedimentos realizados na atenção básica com os realizados em outro nível de complexidade.

Para isso propõe-se a regra dos KNN ou seja, K vizinhos mais próximos, na tentativa de estabelecer um método de classificação por meio do processamento de dados, distribuindo os conjuntos dos mesmos, alcançando as probabilidades entre as classes, que neste caso aplica-se a relação de cintura quadril e pressão arterial.

A classificação típica para determinação de decisões, inicia-se com o cálculo entre o processo de treinamento, a verificação de quais classes, os K que estão mais próximos, trazendo a associação entre o modelo de teste e a classe mais frequente entre os K.

Tal temática vem sendo debatida e bastante utilizada por diversos autores, tais como a classificação da pressão arterial não invasiva com base na fotopletismografia usando o KNN, um

dos métodos utilizados por Anggoro e Kurnia⁽¹⁴⁾

Tjahjadi e Raml⁽¹⁵⁾ utilizaram KNN, dado que o mesmo apresenta excelente desempenho em aplicações de classificação, como também é destacado por Tchuenté F, Baddour N, Lemaire⁽¹³⁾ que o método de classificação KNN, pode fornecer exatidão, sensibilidade, especificidade, precisão, pontuação, bastante relevantes.

Fati SM, Muneer A, Akbar NA, Taib SM. A⁽¹⁶⁾ realizaram a seleção automatizada com o método KNN pode ser realizada para selecionar os melhores recursos e estimar os indicadores de PA.

O objetivo do presente estudo foi avaliar a eficiência do classificador KNN em identificar sujeitos hipertensos através da RCQ em um algoritmo de inteligência artificial, considerando sua acurácia, precisão e sensibilidade.

MÉTODOS

As variáveis coletadas e o detalhamento metodológico relativo à KNN estão descritos a seguir.

Caracterização da amostra

A amostra é composta por 499 mulheres, usuários do Programa de Saúde da Família (PSF) na cidade de Poços de Caldas, Minas Gerais. Para aferição da PA foi utilizado um esfigmomanômetro de coluna de mercúrio (0-300 mmHg) da marca Biotec[®], a medida das circunferências da cintura e do quadril foram realizadas com uma fita antropométrica, com

escala em milímetros da marca Sanny[®].

Procedimentos metodológicos da KNN

O algoritmo desenvolvido na linguagem Python 3.0 e bibliotecas específicas do mesmo foram utilizadas. Foram alocados 80% dos dados para o treinamento do algoritmo e 20% para a fase de testagem do mesmo.

O algoritmo foi processado na interface Google Colab[®], com a importação das bibliotecas: pandas, matplotlib, KNeighborsClassifier, Numpy, scipy.

Como instrumento de classificação foi utilizado a KNN para identificar a classe de uma variável não contida em um conjunto de treinamento. Uma das características da KNN é sua sensibilidade em relação ao valor de K e da distância utilizada na mesma.

O algoritmo utilizado automatizou a melhor distância para o classificador KNN, sendo utilizado o postulado de Minkowski presente no mesmo.

Através da fórmula de Minkowski que relaciona os valores de X e Y em relação à potência $1/p$, onde $p=2$, reduzindo-a à distância euclidiana para a determinação dos K vizinhos mais próximos.

A fórmula de Minkowski é:

$$(\sum_i = (|X_{eu}-Y_{eu}|^p))^{(1/p)}$$

Foram utilizados diferentes valores de K, sendo eles, 3, 5 e 7 para identificação do melhor parâmetro.

O conjunto de variáveis na fase de treinamento do algoritmo foram RCQ, PAS/PAD e o diagnóstico (0 para hipertensos e 1 para normotensos).

Para verificação da acurácia do algoritmo obtido com a KNN, gerou-se uma matriz de confundimento, a fim de ratificar as métricas e identificar qual o melhor valor de K. Obteve-se também a precisão, a sensibilidade e o F1-score.

A seguir serão apresentados os resultados, encontrados na aplicação da KNN nos três diferentes cenários.

RESULTADOS E DISCUSSÃO

Após a importação dos dados foi estabelecido um treinamento de aprendizagem com as variáveis disponíveis que neste caso são

RCQ e PA e sequencialmente um teste de validação.

Para K=3, obteve precisão de 0,89, sensibilidade 0,89 e F1-score de 0,88, sendo a acurácia de 89%. Para K=5, obteve precisão de 0,88, sensibilidade 0,86 e F1-score de 0,85, sendo a acurácia de 86%. Para K=7, obteve precisão de 0,85, sensibilidade 0,81 e F1-score de 0,79, sendo a acurácia de 81%. Tais resultados estão sintetizados na tabela 1 onde observa-se que a maior acurácia foi para k=3.

Quanto mais próxima de 1 por a razão entre acurácia e precisão melhor será o valor de K correspondente. Observa-se que para K=7 esta razão é de 0,95; para K=5 a mesma é de 0,97 e para K=3 a razão é igual a 1, o que corrobora que os três vizinhos mais próximos são mais eficientes no presente algoritmo.

K	Precisão	Sensibilidade	Acurácia	F1-Score
3	0,89	0,89	0,89	0,88
5	0,88	0,86	0,86	0,85
7	0,85	0,81	0,81	0,79

Sendo assim, a precisão, sensibilidade, acurácia e F1-Score, tornam-se valiosas medidas de desempenho e interpretação de dados.

De acordo com Shana J e Venkatachalam T⁽¹⁷⁾, abordagens de classificação de mineração de dados são muito eficientes e a utilização do classificador KNN, é largamente aplicado na atualidade na área da saúde.

Da mesma forma, Fatchan, M e Akbar, Muhamad T and Hadikristanto, W e Firmansyah, frisa que os resultados obtidos em sua pesquisa são muito bons, aplicando a classificação KNN, com a observação precisão, sensibilidade, acurácia e F1-Score.

Por isso, implementar a ciência de dados, tais como modelos baseados em IA em

processos na área da saúde, ratificam e favorecem a compreensão do fenômeno, trazendo uma quantificação necessária, baseado em estatística e inovação.

CONCLUSÃO

Nos três contextos simulados para diferentes valores de K foi encontrada diferença entre a melhor (K=3) e a pior (K=7) acurácia de 8.9%, fato que corrobora a importância da testagem do algoritmo com diferentes valores de K.

Para o presente conjunto de dados de RCQ e PA o valor de K que apresentou melhor confiabilidade classificatória foi o K=3, divergindo portanto do valor padrão utilizado nas bibliotecas do Python. Isso confirma a importância de simulações a priori que antecedam a implementação final do classificador.

A precisão, sensibilidade e acurácia para K=3 foi de 0,89, valor este que sob o ponto de vista da aplicação precisa ser otimizado. Isto mostra a necessidade da utilização de outras variáveis além do RCQ para que este classificador apresente pelo menos 95% de acurácia.

O KNN é um classificador que apresenta significativa potencialidade de eficiência na área da saúde, especialmente, quando recorre a mais de uma variável para uma otimização do algoritmo.

REFERÊNCIAS

1. ROCHA, A. C; GUEDES JÚNIOR, D. P. Avaliação física para treinamento personalizado, academias e esportes. Uma abordagem didática, prática e atual. 1ªed. Phorte: São Paulo – SP, 2013.
2. CASSIANO, M. H; SILVA, H. T. D; LUZ, A. B. S; BEZERRA, M. S; BARBOSA, S. S; ARAÚJO, D. F. S. Correlação entre os índices antropométricos e pressão arterial de adolescentes e adultos jovens em municípios do nordeste brasileiro. *Ciência Plural*, v. 5, n. 2, p. 49-67, 2019.
3. MENEZES, M. H; REIS, V. H; DANTAS, D. B. G; SANTANA, A. P; DIAS, F. C. F; JUNIOR, C. A. R; CUELLAR, P. M. G. Hipertensão arterial sistêmica e eventos cardiovasculares no estado do Tocantins, Brasil. *Revista de Patologia do Tocantins*, v. 4, n. 2, p. 50-53, 2017.
4. ARAÚJO, G. S. B; MOTA, I. A. V; SILVA, J. D; AOYAMA, E. A; SOUZA, R. A. G. Hipertensão arterial: problema de saúde pública nos dias atuais. *Revista Brasileira Interdisciplinar de Saúde*, v. 1, n. 1, p. 39-43, 2019.
5. Diretrizes do ACSM para os testes de esforço e sua prescrição. 10ªed. Guanabara-Koogan – Rio de Janeiro, 2018.
6. BARROSO, W. K. S; RODRIGUES, C. I. S; BORTOLOTO, L. A; MOTA-GOMES, M. A; BRANDÃO, A. A; FEITOSA, A. D. M, et al. Diretrizes Brasileiras de Hipertensão Arterial – 2021. *Arq Bras Cardiol*, v. 116, n. 3, p. 516-658, 2020.
7. MELO, J. I. V; MATOS, A. C. L; PINTO, R. S; FREITAS, G. R. M. O impacto econômico dos serviços farmacêuticos na assistência à saúde de pacientes portadores de hipertensão: uma revisão sistemática. *J Bras Econ Saúde*, v. 13, n. 1, p. 66-77, 2021.
8. ANDRADE, M. S; LIRA, C. A. B. Fisiologia do Exercício. 1ªed. Manole: Barueri- SP, 2016.
9. LIMA, J. S; BEZERRA, L. A; GOMES, R. M. O. P; FONTES, A. A. B; NORONHA, V. F. C. M; JESUS, J. B; LIMA, S. O. Correlação entre indicadores antropométricos de obesidade e níveis pressóricos. *Brazilian Journal of Health Review*, v. 4, n. 4, p. 14725-14735, 2021.
10. NOGUEIRA, D. S et al. Relação cintura quadril em idosos hipertensos e normotensos treinados. In: Congresso Internacional de Atividade Física, Nutrição e Saúde. Anais. Aracaju: UNIT, 2016.
11. TRAN, N. T. T; BLIZZARD, C. L; LUONG, K. N; TRUONG, N. L. V; TRAN, B. Q; OTAHAL, P; NELSON, M; MAGNUSSEN, C; GALL, S; BUI, T. V; SRIKANTH, V; AU, T. B; HA, S. T; PHUNG, H. N; TRAN, M. H; CALLISAYA, M. The importance of waist circumference and body mass index in cross-sectional relationship with risk of cardiovascular disease in Vietnam. *PLOS ONE*, v. 13, n. 5, p. 1-13, 2018.
12. Tchuente F, Baddour N, Lemaire ED. Classification of Aggressive Movements Using Smartwatches. *Sensors (Basel)*. 2020;20(21).
13. Anggoro DA. Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. *International Journal of Emerging Trends in Engineering Research*. 2020;8(5):1689-94.

14. Tjahjadi, Ramli. Noninvasive Blood Pressure Classification Based on Photoplethysmography Using K-Nearest Neighbors Algorithm: A Feasibility Study. *Information*. 2020;11(2).
15. Fati SM, Muneer A, Akbar NA, Taib SM. A Continuous Cuffless Blood Pressure Estimation Using Tree-Based Pipeline Optimization Tool. *Symmetry*. 2021;13(4):686.
16. Shana J, Venkatachalam T. Efficient Data-Mining Classification Approach For Ecg Data In Health Care Application, 03 March 2021, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-277577/v1>]
17. Fatchan, Muhamad and Akbar, Muhamad Taufik and Hadikristanto, Wahyu and Firmansyah, Andri. (2021) Determination of Nutritional Status Using Classification Method Datamining Using K- Nearest Neighbord (KNN) Algorithm. In: InCEEES 2020, 17-18 July 2020, Bekasi, Indonesia.

OBSERVAÇÃO: Os autores declaram não existir conflitos de interesse de qualquer natureza. Alguns dos autores são bolsistas Capes.