doi.org/10.51891/rease.v1i3.13312

A TECNOLOGIA EM BENEFÍCIO À SAÚDE: CLASSIFICAÇÃO DE SINAIS DE ELETROCARDIOGRAMA COM O USO DE INTELIGÊNCIA ARTIFICIAL

Emanuelle Passos Martins[1] Alisson Assis Cardoso[2]

RESUMO: Extrassístoles ventriculares são um tipo de arritmia caracterizada por impulsos ventriculares isolados e podem acometer tanto indivíduos saudáveis quanto cardiopatas. Daí a importânciado diagnóstico, que normalmente é feito por uma equipe médica a partir de um exame eletrocardiograma (ECG) do paciente. Diante disso, propõe-se o uso de dois métodos de inteligência artificial para a classificação dos sinais de ECG, em auxílio a tais profissionais, são eles: Regressão Linear Múltipla (RLM) e algoritmo *k-means*. Utilizou-se dados de ECG *MIT-BIH Arrhythmia Database*, com um total de 155 batimentos cardíacos referentes a um paciente, e obteve-se acurácia de 74,19% com o uso de ambos RLM e *k-means*. Assim, foi demonstrado que a aplicação de RLM e *k-means* na classificação de batimentos cardíacos a partir do exame ECG associado à avaliação da equipe médica pode gerar melhores análises no diagnóstico de extrassístoles ventriculares.

Palavras-chave: Complexos ventriculares prematuros. *K-means*. Regressão linear múltipla.

Área Temática: Medicina. Inteligência Artificial.

ABSTRACT: Ventricular extrasystoles are a type of arrhythmia characterized by isolated ventricularimpulses and can affect both healthy individuals and those with heart conditions. Hence, the importance of diagnosis, which is typically made by a medical team through an electrocardiogram (ECG) examination of the patient. Considering this, the use of two artificial intelligence methods for classifying ECG signals is proposed to assist these professionals: Multiple Linear Regression (MLR) and k-means algorithm. Besides that, MIT-BIH Arrhythmia Database was used, comprising a total of 155 heartbeats from a patient, and an accuracy of 74.19% was achieved using both MLR and k-means. Thus, it has been demonstrated that the application of MLR and k-means in the classification of heartbeats from the ECG examination, combined with the evaluation of the medical team, can lead to improved analyses in the diagnosis of ventricular extrasystoles.

Keywords: Ventricular premature complexes. K-means. Multiple linear regression.

[1]Universidade Federal de Goiás Goiânia, Goiás.
[2] Universidade Federal de Goiás, Goiânia, Goiás.

INTRODUÇÃO

Extrassístoles ventriculares (ESV), ou contrações ventriculares prematuras (CVP), também conhecidas como falha do batimento cardíaco, são doenças cardiovasculares caracterizadas por impulsos ventriculares isolados, podendo haver ou não a presença de sintomas (MITCHELL, 2023). A ESV é um tipo de arritmia que acomete tanto indivíduos saudáveis quanto cardiopatas.

Daí a importância do diagnóstico de ESV, que geralmente é dado por uma equipe médica a partir de análises do eletrocardiograma (ECG) do paciente. O ECG é um exame composto por sinais que representam o batimento cardíaco, que pode ser dividido nas regiões PQRST, sendo cada uma delas uma parte dos impulsos elétricos que dizem sobre a atividade elétrica cardíaca (MORSCH, 2018).

De forma sintética, em um ritmo sinusal, a onda P diz sobre a contração dos átrios, o complexo QRS sobre a contração dos ventrículos e a onda T, onda de recuperação, representa a repolarização ventricular, isto é, a alteração elétrica dos ventrículos em preparação para o próximo batimento cardíaco (MORSCH, 2018). Assim, o diagnóstico de ESV é marcado por alterações no formato dessa onda, normalmente caracterizado por largos complexos QRS, com a ausência de P e seguido por uma pausa compensatória (MITCHELL, 2023).

Na literatura, diversos métodos têm sido utilizados para se automatizar a classificação de sinais de ECG. Os resultados do programa computacional desenvolvido por Guimarães (2019) que identifica sinais de eletrocardiograma apontam a importância e eficácia de técnicas de aprendizado de máquina como forma de identificação de doenças cardíacas. Aspuru *et al.* (2019) propôs o uso do algoritmo de regressão linear para analisar sinais de ECG com baixo custo computacional e detectar informações sobre os picos de onda P, Q, R, S e T relevantes no diagnóstico de doenças cardíacas. Yakut, Bolat, Efe (2021) basearam o seu estudo na clusterização de classes de arritmia ao desenvolver um método de detecção com *k-means*.

À vista disso, o presente trabalho propõe a classificação de batimentos cardíacos, em normais ou extrassistólicos, utilizando-se dois métodos de inteligência artificial, são eles: Regressão Linear Múltipla (RLM) e algoritmo *k-means*.

METODOLOGIA

O conjunto de dados utilizado foi o *MIT-BIH Arrhythmia Database*, que se trata de uma base de dados de ECG de vários pacientes (MOODY; MARK, 2001). Foram utilizados

divididos com a seguinte proporção: 80% dos dados para treinamento e 20% para teste, como mostra a Tabela 1.

Tabela 1 – Divisão dos dados

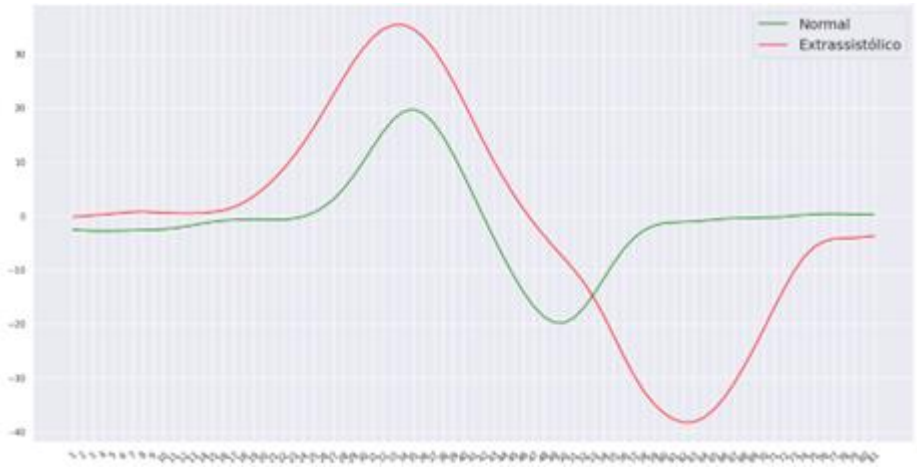
	Batimentos normais	Extrassístoles ventriculares
Treinamento	89	35
Teste	22	9

Fonte: autoria própria.

Ademais, cada batimento foi dividido em 81 partes. Na Figura 1 são mostrados dois batimentos do ECG em análise, podendo-se observar o padrão de comportamento de um batimento normal e de uma extrassístole ventricular.

É importante destacar que na base de dados, para representar os batimentos normais adotou-se o valor 0, e 1 para os batimentos extrassistólicos.

Figura 1 – Primeiro batimento normal (verde) e primeiro batimento extrassistólico (vermelho) do paciente



Fonte: autoria própria.

Para a classificação dos batimentos cardíacos, em normais ou extrassistólicos, fez-se uso de dois métodos: Regressão Linear Múltipla (RLM) e algoritmo *k-means*, que são supervisionados e não-supervisionados, respectivamente.

A RLM consiste em relacionar uma variável dependente (y) com n variáveis independentes (x), onde x são as 81 partes em que se dividiu cada batimento e y é a saída verdadeira, isto é, ter ou não extrassístole ventricular:

$$y = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n \qquad (1)$$

Assim, o cálculo dos coeficientes (b_1, b_2, ... , b_n) pode ser feito através da Pseudo-Inversa de Penrose-Moore, sobre a qual multiplica-se tanto x quanto y pela transposta de x (LAWSON; HANSON, 1995):

$$\hat{b} = (x^T x)^{-1}(x^T y) \qquad (2)$$

em que, $\hat{b}$ é o coeficiente de regressão.

A seguir, é possível estimar $\hat{y}$ isto é, uma predição que classifica o tipo do batimento cardíaco de um novo conjunto de teste:

$$\hat{y} = x_{teste}\hat{b} \qquad (3)$$

Resumindo, os coeficientes foram gerados com base no conjunto de treinamento, e a saída da regressão a partir desses coeficientes e do conjunto de dados de teste.

Ademais, foi estabelecido um limiar, y = 0.5, em que todos os valores abaixo de 0.5 foram classificados como batimentos normais e todos os valores acima de 0.5 como extrassístoles ventriculares, devido ao fato de ter sido utilizado um modelo de regressão para classificação. Dessa forma, os erros foram condicionados ao limiar.

Por outro lado, a utilização do algoritmo *k-means* consiste em estabelecer a quantidade de classes (ou *clusters*) em que se deseja agrupar os dados, e posteriormente definir centroides aleatórios, que representam tais classes.

Assim, definiu-se o número de *clusters*, isto é, agrupamentos, igual a 2, sendo as duas classes: batimento normal e extrassístole ventricular. Logo, a quantidade de centroides geradosfoi igual a 2, por conta de que a quantidade de classes determina a de centroides.

Então, com o uso da métrica distância euclidiana (d), cuja equação pode ser vista a seguir, mediu-se a distância entre os centróides e os pontos.

$$d(A, c_1) = \sqrt{\sum_{i=0}^{n}(P_i - Q_i)^2} \qquad (4)$$

Em que A é um ponto; c_1 é o centroide 1; P = (p_1,p_2, ... , p_n), sendo p_n referente à coordenada y; e Q = (q_1,q_2, ... , q_n), sendo q_n referente à coordenada x.

Dessa forma, identificando-se a menor distância de cada um dos pontos até cada um doscentroides, determinou-se a qual *cluster* (representado pelos centroides) cada ponto

estava pertencendo.

Posteriormente, os centroides foram ajustados aos dados através do cálculo do ponto médio, sobre o qual soma-se as coordenadas dos pontos pertencentes a um certo centroide e divide-se pela quantidade de pontos que pertencem a ele, obtendo desta forma as novas coordenadas dos centroides.

$$c_{pm} = \left(\frac{\sum_{j=0}^{m} x_{i0}}{m}, \frac{\sum_{j=0}^{m} x_{i1}}{m}, \dots, \frac{\sum_{j=0}^{m} x_{in}}{m}\right), X_i \in C_i$$

Esses procedimentos, cálculo da distância euclidiana e do centroide no ponto médio são refeitos para reposicionamento dos centroides, até que a alteração seja mínima entre os reposicionamentos ou até que sejam concluídas todas as iterações, devendo tal valor ser previamente definido. Além disso, foi feita a normalização dos dados para que eles estivessemem uma mesma escala, então foram definidos os parâmetros do *k-means*, que foi treinado comos dados de treino.

Após isso, a identificação de um ponto que estava mais próximo de um determinado centroide do que de outro permitiu detectar a classe dos dados de teste. Por exemplo, ao verificar que um ponto A, que está mais próximo do centroide c_1 do que do c_2, possui valor real 1, conclui-se que tal centróide (c_1) representa os batimentos extrassistólicos. Dessa forma,estabeleceu-se uma *label*, isto é, um rótulo para classificação dos dados em análise, tornandosupervisionado tal método que era não-supervisionado.

Acerca da parametrização do *k-means*, é interessante citar que há vários parâmetros, que podem ser melhor detalhados na sua documentação (PEDREGOSA *et al.*, 2011). Diante dos que foram utilizados, *n_clusters* corresponde à quantidade de classes em que se deseja classificar os dados, enquanto *init* diz sobre a maneira como o algoritmo será inicializado, podendo ser de forma aleatória ou com *k-means++*. Dentre as várias possibilidades de inicialização, o *k-means++* calcula a distribuição de probabilidade empírica dos pontos que contribui na inércia geral, em outras palavras, isto ajuda a definir com mais assertividade a posição inicial dos *clusters*. Por outro lado, *n_init* se refere à quantidade de vezes em que o algoritmo é executado com diferentes centroides e *max_iter* define o número máximo de iterações do algoritmo a cada execução.

RESULTADOS E DISCUSSÃO

Após aplicar a RLM nos dados de treinamento, obteve-se os seguintes coeficientes angular e linear indicados na Figura 2.

Figura 2 – Coeficientes angular e linear da RLM

```
1 angular_coefficient = mlr_model.coef_
2 angular_coefficient
```

```
array([  0.03055001,   0.27695189,  -1.63938405,   0.14872164,
         3.77382002,  -3.74156395,   0.30087473,   2.49630976,
        -1.4018904 ,  -0.72746655,   0.56914281,  -1.81616312,
         3.56862244,  -0.64121949,  -3.68710709,   3.81370195,
        -1.91304717,   0.56529853,  -0.10813547,  -2.32983285,
         5.85949191,  -4.30904103,  -0.55953485,   3.18030175,
        -0.76233281,  -3.58750055,   4.5559263 ,  -1.09321078,
        -2.0599704 ,   2.40995122,  -2.91708537,   0.65788811,
         5.45549364,  -8.48773949,   6.5278382 ,  -4.96806917,
         3.32050736,   1.40304022,  -5.68042247,   7.83089729,
        -8.11774589,   5.0028183 ,   1.86411703,  -4.75114419,
         0.25218644,   0.5566472 ,   1.98004514,   1.66569711,
        -6.20867781,   2.58001153,   0.36455236,   6.00628327,
       -10.26905341,   4.86547558,   2.59580169,  -5.59950239,
         5.67017201,  -2.72425129,  -3.143713  ,   6.4693959 ,
        -6.58345497,   6.76886502,  -5.9962461 ,   1.00126384,
         0.80074493,   3.92312543,  -3.74903731,   0.05259225,
        -0.74898037,   2.38543018,  -1.55380491,   2.74688052,
        -8.36357572,  10.31531458,  -3.96918125,  -3.4640457 ,
         5.78672429,  -5.33739731,   2.65564945,   0.76110332,
        -0.80198593])
```

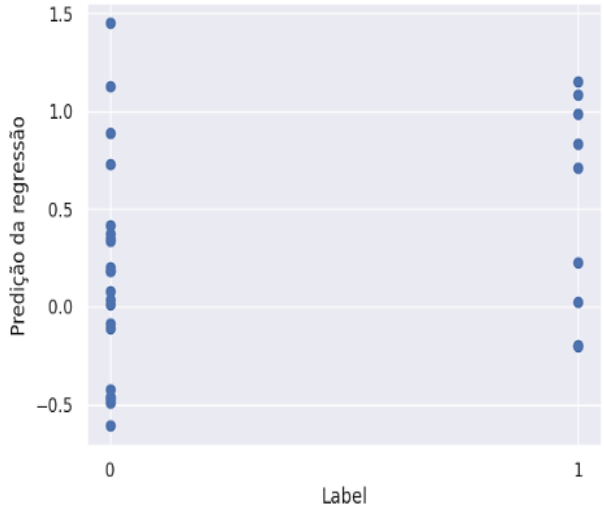
```
1 linear_coefficient = mlr_model.intercept_
2 linear_coefficient
```

```
0.4158934249927613
```

Fonte: autoria própria.

Dessa forma, atingiu-se os resultados apresentados na Figura 3.

Figura 3 – Predição da regressão diante da label de cada batimento



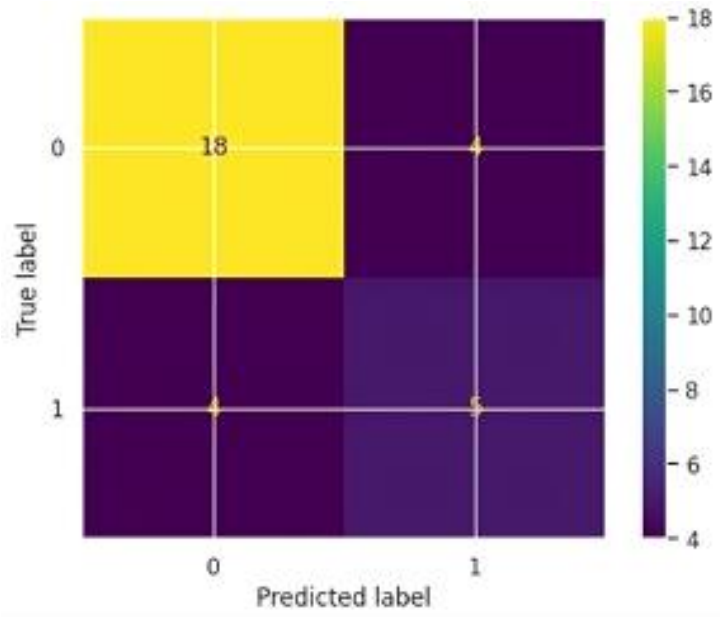
Fonte: autoria própria.

Sendo os pontos azuis a, os valores abaixo do limiar (0.5) foram classificados como batimentos normais e aqueles acima ou igual ao limiar como extrassistólicos.

É possível verificar na matriz de confusão (Figura 4) que a maior parte dos dados de teste (23), aproximadamente 74,19%, foi classificada de forma correta, sendo a maioria deles

(18) batimentos normais. Também pode-se observar que 4 batimentos normais foram classificados de forma errada como sendo extrassistólicos e que 4 extrassístoles foram classificadas incorretamente como batimentos normais, sendo a porcentagem de classificaçõesincorretas igual a aproximadamente 25,81%.

Figura 4 – Matriz de confusão da classificação dos dados com RLM



Fonte: autoria própria.

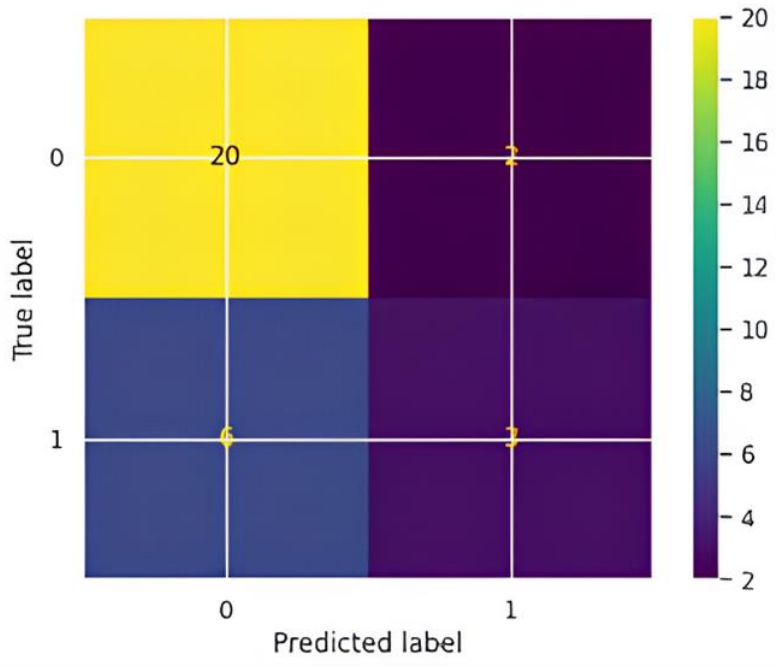
Sob outra perspectiva, a melhor performance do *k-means* foi obtida adotando-se os seguintes parâmetros: *n_clusters* = 2, *init* = 'k-means++', *n_init* = 10, *max_iter* = 500. Tal clusterização é apresentada na Figura 5, sendo cada uma das cores a representação de cada agrupamento, totalizando dois agrupamentos. Os resultados obtidos com o uso do *k-means* estão representados na matriz de confusão na Figura 6.

Com tal algoritmo também 23 batimentos cardíacos (74,19%) foram classificados corretamente, sendo 20 deles batimentos normais, valor superior ao obtido na classificação com o uso da RLM, enquanto os outros 8 foram classificados de forma errada. Dentre esses batimentos, 6 se tratavam de extrassístoles classificadas erroneamente como batimentos normais e 2 eram batimentos normais tidos como extrassistólicos.

Figura 5 – Clusterização dos dados com k-means

Fonte: autoria própria.

Figura 6 – Matriz de confusão da classificação dos dados com *k-means*



Fonte: autoria própria.

Como afirma Aspuru (2019), a principal vantagem do algoritmo de regressão linear se trata da sua simplicidade teórica, o que se aplica também ao *k-means*. Essa simplicidade, por sua vez, impulsiona um número de estudos na área de classificação de sinais de ECG maior com redes neurais do que com eles, devido à complexidade dessas muitas vezes ser associada

a melhores resultados. Ainda assim, se obteve acurácia acima de 74% para ambos, reconhecendo que métodos mais simples também têm o potencial de gerar contribuições significativas.

Apesar dos vários testes feitos em ambos os modelos para se atingir as melhores configurações principalmente quanto aos parâmetros, reconhece-se a importância da quantidade dos dados. Tendo sido utilizados os dados referentes somente a um paciente, associa-se à pequena quantidade de dados a performance em torno de 74%, se comparada às obtidas por Guimarães (2019) com os métodos *Support Vector Machine*, *Decision Tree*, *K Neighbors* e *Random Forest*, e por Yakut, Bolat, Efe (2021) com *k-means*, todos acima de 90%.

Guimarães (2019) evidencia que as razões para as altas taxas de acerto obtidas em seu estudo se devem em especial ao grande número de dados de treinamento. Contudo, também demonstra que nem sempre uma mudança nessa quantidade é proporcional aos resultados, pois em alguns cenários a eficácia do modelo teve uma leve queda apesar do aumento de dados, concluindo que são necessários testes experimentais. Assim, para trabalhos futuros seria interessante o uso dos dados referentes a outros pacientes em prol de aumentar a base de dados.

Diferentemente do estudo de Yakut, Bolat, Efe (2021) que analisou somente o valor de acurácia, dedicou-se a conhecer mais profundamente os resultados obtidos. Ao se analisaros falsos negativos e falsos positivos pode-se na área da saúde retardar o tratamento de um paciente ou submetê-lo a condições desnecessárias, respectivamente. Sobre isso, os modelos obtiveram comportamentos diferentes, sendo a precisão da RLM melhor para com os batimentos normais e o *k-means* para com as extrassístoles.

CONCLUSÃO

Os resultados mostraram acurácia de 74,19% ao classificar batimentos cardíacos, em normais ou extrassistólicos, com o uso de ambos RLM e *k-means*, enquanto a taxa de erro foi de 25,81%. Apesar de tal valor ter sido o mesmo, a quantidade de batimentos normais classificados corretamente com o *k-means* foi maior em 2 unidades em relação à RLM, sendo 20 e 18, respectivamente. Por outro lado, quanto à classificação correta das extrassístoles o *k-means* obteve melhor performance, com 5 batimentos contra 3 com a RLM.

Na área da saúde, a análise de tais valores é importante pois ao se classificar erroneamente um batimento normal em extrassístole, pode-se submeter o paciente a um tratamento desnecessário, enquanto o contrário pode implicar complicações no estado de saúde dele, devido à falta de tratamento.

Assim, foi demonstrado que é possível a aplicação de métodos simples como a RLM e o *k-means* ao diagnóstico de extrassístoles ventriculares diante de um exame ECG em auxílio ao profissional da saúde, permitindo uma melhor análise da condição do paciente e escolha do tratamento. Vale destacar a importância da qualidade dos dados bem como da quantidade ao se utilizar métodos de inteligência artificial. Assim, para trabalhos futuros seria interessante aumentar a base de dados para que o modelo possa aprender com mais informações e desse modo, conseguir mais acertos nas predições, melhorando a sua performance.

REFERÊNCIAS BIBLIOGRÁFICAS

ASPURU, J. *et al*. Segmentation of the ECG signal by means of a linear regressionalgorithm. **Sensors**, v. 19, n. 4, p. 775, 2019.

GUIMARÃES, T. J. R. Identificação de doenças cardíacas a partir de Eletrocardiogramas utilizando Machine Learning. 2019. 39f. Monografia (Bacharelado em Engenharia Elétrica) - Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2019.

LAWSON, C. L.; HANSON, R. J. Solving least squares problems. **Society for Industrial and Applied Mathematics,** 1995.

MITCHELL, L. B. Extrassístole ventricular (ESV). **Manual MSD**, 2023. Disponível em: https://www.msdmanuals.com/pt-br/profissional/doen%C3%A7as-cardiovasculares/arritmias-card%C3%ADacas-espec%C3%ADficas/extrass%C3%ADstole-ventricular-esv. Acesso em: 04 jan. 2024.

MOODY, G.; MARK, R. The impact of the MIT-BIH Arrythmia Database. **IEEE Engineering in Medicine and Biology Magazine**, v. 20, n. 3, p. 45-50, 2001.

MORSCH, J. A. O que são as ondas do eletrocardiograma e como interpretar? **Telemedicina Morsch**, 2018. Disponível em: https://telemedicinamorsch.com.br/blog/ondas-do-eletrocardiograma. Acesso em: 04 jan.2024.

PEDREGOSA, F. *et al*. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

YAKUT, Ö.; BOLAT, E. D.; EFE, H. K-Means Clustering Algorithm Based Arrhythmic Heart Beat Detection in ECG Signal. **Balkan Journal of Electrical and Computer Engineering**, v. 9, n. 1, p. 53-58, 2021.