ISSN 1870-4069

Detección de enfermedades cardiacas: Implementación de clasificador probabilístico en un dispositivo embebido

Marcos Julián Benítez-Rodríguez, Diego Pérez-Vega,
Jorge Luis Pérez-Ramos, Selene Ramírez-Rosales,
Luis Antonio Díaz-Jiménez, Ana Marcela Herrera-Navarro,
Hugo Jiménez-Hernández, Daniel Cantón-Enríquez

Universidad Autónoma de Querétaro,
Facultad de Informática,
México

{mbenitez07, dperez126}@alumnos.uaq.mx, daniel.canton@uaq.mx

Resumen. El diagnóstico temprano de enfermedades cardíacas desempeña un papel crucial en la toma de decisiones médicas para mejorar la salud de los pacientes con afecciones cardíacas. Una estrategia eficaz para este propósito implica la aplicación de técnicas de aprendizaje automático, las cuales facilitan la identificación de patrones y la comprensión de síntomas relacionados con estas enfermedades. En este trabajo, se ha implementado un clasificador probabilístico en un dispositivo embebido para la detección oportuna de enfermedades cardíacas. Este modelo utiliza la regla de Bayes junto con hipótesis simplificadoras que se basan en la suposición de independencia probabilística entre las variables clínicas. Durante el proceso de aprendizaje de los parámetros se asumen que cada variable clínica cuenta con una distribución normal. Posteriormente, se evalúa el rendimiento del modelo utilizando métricas de la matriz de confusión. Los resultados obtenidos muestran que el clasificador probabilístico propuesto mejora su rendimiento en comparación con otros trabajos consultados en la literatura. Además, se resaltan las ventajas de implementar el modelo propuesto en un dispositivo embebido. Así como, los trabajos a futuro a realizar en la investigación.

Palabras clave: Aplicaciones médicas, aprendizaje automático, detección de enfermedad cardiaca, razonamiento probabilístico, sistema embebido.

Heart Disease Detection: Implementation of Probabilistic Classifier in an Embedded Device

Abstract. Early diagnosis of heart disease plays a crucial role in medical decision making to improve the health of patients with cardiac conditions. An effective strategy for this purpose involves the application of machine learning techniques, which facilitate the identification of patterns and the understanding of symptoms related to these diseases. In this work, a probabilistic classifier has been implemented in an embedded device for timely detection of heart disease. This model uses Bayes' rule together with simplifying assumptions based on the

assumption of probabilistic independence between clinical variables. During the parameter learning process, each clinical variable is assumed to have a normal distribution. Subsequently, the performance of the model is evaluated using confusion matrix metrics. The results obtained show that the proposed probabilistic classifier improves its performance compared to other works consulted in the literature. Furthermore, the advantages of implementing the proposed model in an embedded device are highlighted. As well as the future works to be carried out in the research.

Keywords: Coronary artery disease, machine learning, heart disease detection, probabilistic reasoning, embedded system.

1. Introducción

Las enfermedades cardiovasculares son la principal causa de muerte en el mundo, cobrando la vida de aproximadamente 17.9 millones de personas cada año según la Organización Mundial de la Salud [1]. En México, los infartos de miocardio y los accidentes cerebrovasculares son responsables de alrededor de 150,000 muertes anuales, lo que destaca la importancia de prestar atención al infarto agudo de miocardio como una prioridad en la atención médica [2].

En esta situación [3], varios elementos dificultan la identificación temprana de enfermedades cardíacas: *i)* la falta de disponibilidad de cardiólogos, profesionales médicos especializados en estas afecciones; *ii)* una distribución desigual de los cardiólogos en el territorio, con una concentración en las principales zonas urbanas, y *iii)* factores relacionados con la alimentación, como el alto consumo de sal, la falta de actividad física y el tabaquismo.

Por otro lado, [4] menciona que el rápido desarrollo económico y el estilo de vida acelerado pueden predisponer a las personas a enfermedades crónicas, donde los síntomas suelen manifestarse en etapas avanzadas, lo que dificulta los tratamientos efectivos en etapas tempranas. Por lo tanto, subraya la importancia de fortalecer la atención médica comunitaria y buscar alternativas para la detección temprana de estas patologías.

El monitoreo continuo de los índices fisiológicos más representativos a través del uso de tecnologías digitales ha tenido una aceptación creciente en los últimos años [5]. Esto ha ocasionado que aspectos como históricos de algunas variables fisiológicas sean utilizadas para la predicción de ciertas enfermedades.

No obstante, las técnicas de aprendizaje automático se erigen como un pilar fundamental en el desarrollo de herramientas computacionales que ayuden en la detección oportuna de enfermedades cardíacas. En estudios relacionados a resolver esta problemática, se ha contado con métodos predictivos, tales como, redes neuronales multicapa [6], *random forest* [7] y máquinas de soporte vectorial [8].

Por otra parte, el procesamiento de información proveniente de sensores, en comparación con las mediciones médicas convencionales, ha demostrado ser especialmente prometedor en la detección temprana de enfermedades cardiacas [9]. En este sentido las investigaciones destacan el papel crucial de estas tecnologías en la personalización de la democratización de la tecnología mediante un acercamiento global a la población, de forma que se puedan promocionar estilos de vida saludables

y accesibles para un mayor número de individuos, así como en la identificación temprana de la necesidad de atención médica [10].

En el presente trabajo, se implementa un clasificador probabilístico en un sistema embebido para la detección oportuna de un tipo de enfermedad cardiaca, arterias coronarias. Para ello, se utiliza una placa Raspberry Pi 4 como dispositivo de aplicación remota para la evaluación de nuevos usuarios que no fueron entrenados previamente.

El resto del artículo está organizado de la siguiente manera: se presentan los materiales y métodos utilizados para la implementación del clasificador probabilístico en un sistema embebido; después, se muestran los resultados obtenidos del aprendizaje y la evaluación del clasificador; posteriormente, la discusión de los resultados; por último, las conclusiones y trabajos a futuro.

2. Marco teórico

En esta sección se explican los fundamentos teóricos en los que se basa la presente investigación. Primero, se habla acerca del dispositivo utilizado, así como detalles técnicos del mismo. Luego, se habla de forma general acerca de las diferentes técnicas de aprendizaje automático que hay en el estado de arte. Por último, se revisa el modelo matemático del clasificador probabilístico implementado.

2.1. Dispositivo embebido

La placa Raspberry Pi 4, está orientada hacia sistemas embebidos y multipropósito. Algunas características de la placa Raspberry Pi 4 se describen en la Tabla 1 [11]. Por otro lado, se ha elegido la placa Raspberry Pi por su versatilidad para aplicar un clasificador probabilístico que aprende mediante análisis clínicos de pacientes que puedan tener una enfermedad de arterias coronarias.

Además, como trabajo a futuro se pretende utilizar un sensor conectado a los pines de la placa Raspberry Pi, por ejemplo, el registro de la actividad eléctrica del corazón. Por último, una característica poderosa de la placa Raspberry Pi 4 es su fila de pines GPIO (general-purpose input/output) a lo largo del borde superior de la placa, véase Figura 1.

2.2. Aprendizaje automático

El aprendizaje automático (AA) es parcialmente un campo de la inteligencia artificial, que se enfoca en la toma de decisiones en condiciones de incertidumbre a partir del aprendizaje de datos. Por otro lado, AA es un proceso de dos fases que implica la selección de características relevantes y la adaptación del modelo en función de estas características [12].

No obstante, el aprendizaje automático se centra en tres conceptos principales: datos, modelo y aprendizaje [13]. Los datos son fundamentales para el funcionamiento del aprendizaje automático, y los modelos se diseñan para detectar patrones útiles en los datos. Además, los algoritmos de aprendizaje automático pueden ser supervisados, no supervisados, o por refuerzo, dependiendo del conocimiento a priori disponible y los objetivos del aprendizaje.

Tabla 1. Principales características de la placa Raspberry Pi 4. Elaboración propia.

Características	**Descripción**
Tamaño	88mm x 58mm x 19.5mm
Procesador	ARM Cortex A72
Velocidad	hasta 2.50GHz
RAM	de 2, 4 y 8 GB
Puertos GPIO	40 pines
Entradas	2 micro HDMI, 2 USB 2.0, 2 USB 3.0 Micro SD, conector de audio tipo jack y alimentación de USB-C
Sistema Operativo	Raspbian

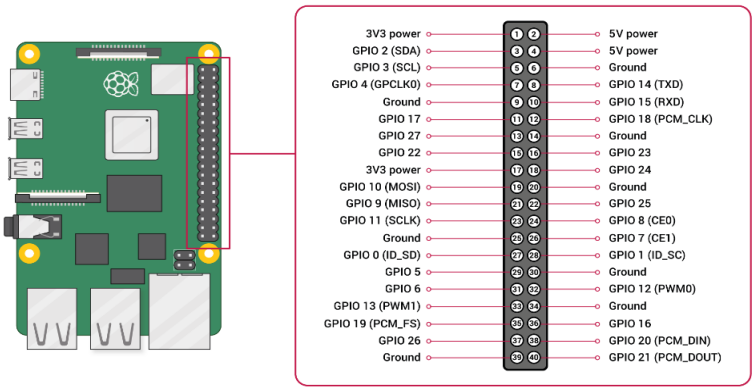

Fig. 1. Ubicación y posición de los pines integrados en la Raspberry Pi 4. Recuperado de [10].

El aprendizaje supervisado implica la deducción de una correspondencia entre entradas y salidas conocidas, mientras que el aprendizaje no supervisado busca modelar la estructura o distribución de los datos sin información previa sobre las salidas. El aprendizaje por refuerzo se basa en el proceso de ensayo y error para maximizar una función de recompensa a largo plazo [14].

Por otro lado, los algoritmos de aprendizaje automático se pueden dividir de acuerdo con el problema que se busca resolver. En la Figura 2, se muestran las principales tareas dentro del aprendizaje automático, las cuales varían de acuerdo con distintos autores [13, 14].

2.3. Clasificador probabilístico

El clasificador probabilístico es un modelo condicional que resuelve un problema de clasificación, representado por un vector $x = (x_1, \dots, x_n)$ donde n representa la n-ésima característica asignada como probabilidad, la misma se define como:

$$P(C_k|F_1, F_2, \dots, F_n), \tag{1}$$

Fig. 2. Tareas de aprendizaje automático de acuerdo con su aprendizaje.

donde cada k son los posibles resultados de una clase C_k. Dicha variable está condicionada al cumplimiento de ciertas variables independientes $x_1, x_2, \dots, x_n$, basadas en el teorema de Bayes, se reescribe la ecuación Ecuación 1 como Ecuación 2:

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{P(C_k) \cdot P(\boldsymbol{x}|C_k)}{P(\boldsymbol{x})}. \tag{2}$$

En la práctica, se presta especial atención en el numerador, ya que el denominador es independiente de C_k. Por lo tanto, se tienen claro conocimiento de los valores x_i de modo que el denominador es constante. Así mismo, en el numerador, se aplica la regla del producto para eventos dependientes, véase Ecuación 3:

$$P(x_1, x_2, \dots, x_n, C_k), \tag{3}$$

donde $P(x_1, x_2, \dots, x_n, C_k)$ es una probabilidad conjunta, es decir, $x_1, x_2, \dots, x_n, C_k \equiv x_1 \cap x_2 \cap \dots \cap x_n \cap C_k$, la cual se reescribe utilizando la regla de la cadena para eventos repetidos de la definición de probabilidad condicional, véase Ecuación 4:

$$\begin{aligned} P(x_1, x_2, \dots, x_n, C_k) = P(x_1|x_2, \dots, x_n, C_k) \cdot P(x_2|x_3, \dots, x_n, C_k) \cdot P(x_{n-1}|x_n, C_k) \\ \cdot P(x_n|C_k) \cdot P(C_k). \end{aligned} \tag{4}$$

Luego, se toma en cuenta el concepto de independencia probabilística, donde, se asume que cada x_i variable es independiente de cualquier otra x_j para $i \neq j$ cuando se encuentran condicionadas a C_k, véase Ecuación 5:

$$P(x_i|x_{i+1}, \dots, x_n, C_k) = P(x_i|C_k). \tag{5}$$

En consecuencia, el modelo de probabilidad conjunta se expresa en la Ecuación 6:

$$\begin{aligned} P(C_k|x_1, x_2, \dots, x_n) \propto P(x_1, x_2, \dots, x_n, C_k) \propto P(C_k) \cdot P(x_1|C) \cdot P(x_2|C_k) \cdots \\ P(x_n|C_k) \propto P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k). \end{aligned} \tag{6}$$

Tabla 2. Variables del conjunto de datos utilizado. Elaboración propia.

Nombre	Descripción
Age	Edad en años
Sex	1 = hombre 0 = mujer
Cp	Tipo de dolor en el pecho: 1 = angina típica 2 = angina atípica 3 = pa sin angina 4 = asintomático
Trestbps	Presión arterial en reposo (en mm Hg al ingreso al hospital)
Chol	Colesterol sérico en mg / dl
Fbs	Azúcar en sangre en ayunas >120 mg/dl (1 – verdadero; 0 = falso)
Restecg	Resultados electrocardiográficos en reposo (0 = normal; 1 = teniendo ST-T; 2 = hipertrofia)
Thalach	Frecuencia cardiaca máxima alcanzada
Exang	Angina inducida por ejercicio (1 = sí; 2 = no)
Oldpeak	Depresión del ST inducida por el ejercicio en relación con el reposo
Slope	pendiente del segmento ST de ejercicio pico (1 = pendiente ascendente; 2 = plano; 3 = pendiente descendente)
ca	número de vasos principales (0-3) coloreados por la floración
Thal	1 = normal; 2 = defecto fijo; 3 = defecto reversible

De modo que, bajo los supuestos de independencia anteriores, el modelo probabilístico con enfoque Bayesiano se define en la Ecuación 7:

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{1}{Z} P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k), \tag{7}$$

donde Z es un factor que depende exclusivamente de la evidencia de las características. Por último, se combina el modelo Bayesiano con una regla de decisión, en este caso se escoge la hipótesis que sea más probable para minimizar la probabilidad de clasificación errónea. El clasificador resultante se define en la Ecuación 8:

$$\hat{y} = \underbrace{argmax}_{k \in \{1,\dots,k\}} [P(C_k) \prod_{i=1}^{n} P(x_i|C_k)], \tag{8}$$

3. Metodología

La metodología seguida en el proyecto se aborda en la Figura 3. Primero, se describe el conjunto de datos utilizado. Luego, se menciona la etapa de aprendizaje para el entrenamiento del clasificador probabilístico. Por último, se revisa la etapa de evaluación de la implementación.

3.1. Descripción de los datos

El conjunto de datos utilizados fue "*Heart Disease*" del repositorio de datos para aprendizaje automático de la Universidad de California Irving [15]. Los datos fueron medidos y recolectados por la Fundación Clínica de Cleveland (FCC), con 1025 datos

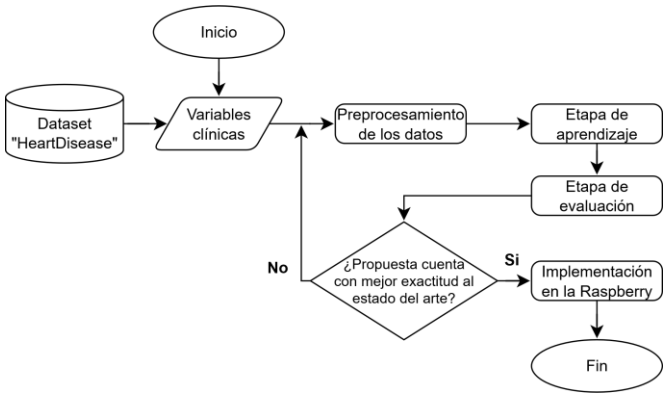

Fig. 3. Esquema de la implementación del clasificador probabilístico en un dispositivo embebido.

de pacientes como muestra. En la Tabla 2, se muestran las trece características del conjunto datos. Además, la distribución de los datos es de 499 casos con enfermedad de arterias coronarias (49%) y 526 casos con ausencia de enfermedad de arterias coronarias (51%). Para la experimentación, el conjunto de datos "*Heart Disease*" fue dividido en dos subconjuntos: el entrenamiento con 75% de los datos, y el de pruebas con 25% de la información restante, seleccionado de mediante un muestreo aleatorio simple.

3.2. Etapa de aprendizaje

Cada característica tiene una distribución empírica especifica, la cual depende de la naturaleza de los datos. En este trabajo, se asumen que las características utilizadas tienen una distribución Gaussiana, véase la Ecuación 9:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{9}$$

donde el parámetro μ es la media o valor esperado de la distribución, mientras que el parámetro σ es la desviación típica, véase en las Ecuaciones 10 y 11:

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{10}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ . \tag{11}$$

3.3. Etapa de evaluación

La evaluación el clasificador probabilístico se llevó a cabo mediante la matriz de confusión. Los elementos de la matriz de confusión son: *i)* Verdaderos Positivos (VP), pacientes que tenían enfermedades cardiacas y estaban correctamente diagnosticados; *ii)* Verdaderos Negativos (VN), pacientes que no tenían enfermedades cardiacas y

estaban correctamente diagnosticados; *iii)* Falsos Negativos (FN), pacientes que tenían enfermedades cardiacas y fueron mal diagnosticados; y *iv)* Falsos Positivos (FP), pacientes que no tenían enfermedades cardiacas y fueron mal diagnósticos. En la industria médica, los FN son las predicciones más peligrosas. Las diferentes métricas de rendimiento se calcularon utilizando una matriz de confusión. La fórmula de exactitud está dada por la Ecuación 12:

$$Exactitud = \frac{VP + VN}{VP + VN + FN + FP}. \tag{12}$$

La precisión es el valor positivo predicho definido por la Ecuación 13:

$$Precisión = \frac{VP}{VP + FP}. \tag{13}$$

La exhaustividad es la proporción de pacientes con enfermedades cardiacas, véase Ecuación 14:

$$Exhaustividad = \frac{VP}{VP+FN}, \tag{14}$$

El valor F1, también conocido como Score-F1, es considerado un promedio armónico entre la precisión y la exhaustividad, véase Ecuación 15:

$$F1_{Score} = 2\left(\frac{Precisión \cdot Exhautividad}{Precisión + Exhautividad}\right). \tag{15}$$

3.4. Implementación en la Raspberry

Los datos de los análisis clínicos se dividen en dos subconjuntos. Primero, el conjunto de entrenamiento es enviado a la placa Raspberry para el aprendizaje de los parámetros. Después, los datos del subconjunto de pruebas son enviados a la placa de Raspberry para predecir si una persona tiene o no una enfermedad de arterias coronarias. Posteriormente, el rendimiento del clasificador se visualiza en una pantalla o monitor que va conectado a la placa Raspberry Pi, véase Figura 4.

4. Resultados

A continuación, se discute el rendimiento del clasificador probabilístico desarrollado en GNU Octave 8.2.0. En la Tabla 3, se contrasta los rendimientos del clasificador probabilístico con respecto a otros trabajos consultados en la revisión de la literatura.

Basado en los resultados, se demuestra que el clasificador probabilístico asumiendo que las características presentan una distribución Gaussiana conlleva a una mejora sustancial en casi todas las métricas, con excepción de la medición F1.

Los resultados anteriores, son importantes dado la importancia de detectar de forma temprana, si una persona tiene o no una enfermedad cardiaca. No obstante, el clasificador propuesto y los demás trabajos consultados pueden llegar a clasificar erróneamente a una persona como enferma cuando no lo esté. Así mismo, se puede clasificar a una persona como no enferma cuando si tenga una enfermedad cardiaca.

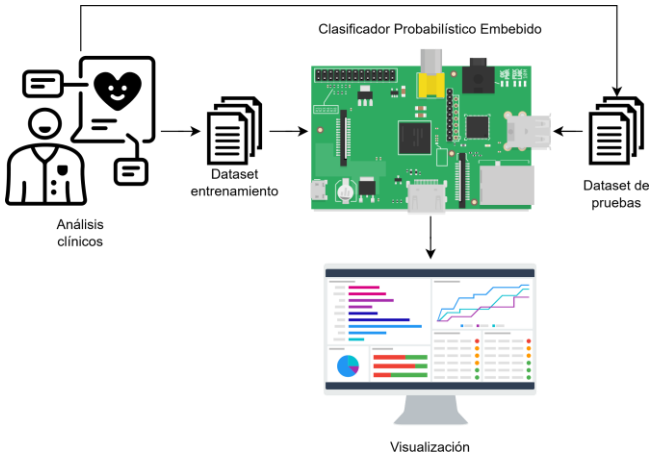

Fig. 4. Arquitectura general.

Tabla 3. Comparación del rendimiento del clasificador probabilístico.

Modelo	Exactitud	Precisión	Exhaustividad	F1
RNA Multicapa [5]	80%	–	–	–
Random Forest [6]	85%	86%	–	**92%**
SVM [7]	82%	90%	–	–
Propuesta (Gaussiana)	**87%**	**94%**	**82%**	88%

Por lo que, se recomienda que el diagnóstico dado por este clasificador sea validado por un médico cardiólogo.

Por otro lado, el uso de la Raspberry Pi 4 puede desempeñar un papel crucial en la detección temprana de enfermedades cardíacas al servir como plataforma de bajo costo. Además, de ser eficiente en la implementación algoritmos de aprendizaje automático, permitiendo así un diagnóstico más rápido y preciso en entornos médicos de difícil acceso.

5. Conclusiones

En este trabajo, se ha elaborado un clasificador probabilístico para realizar diagnósticos tempranos de la enfermedad de arterias coronarias. Este clasificador fue implementado en un sistema embebido de bajos requerimientos usando GNU-Octave. El fundamento de este modelo se basa en la regla de Bayes y en la suposición de independencia entre variables. Sin embargo, aunque el enfoque bayesiano dado al clasificador puede ser efectivo en diversos contextos, en ocasiones su desempeño disminuye debido a la falta de independencia condicional entre las características. Además, algunas características son de tipo discretas, por lo que la suposición de una distribución Gaussiana no siempre es la más apropiada.

Por otro lado, se destaca la importancia de adoptar nuevas tecnologías en el diagnóstico temprano de enfermedades cardíacas, enfatizando su eficacia y seguridad en entornos hospitalarios. La presente implementación se puede emplear estas en áreas remotas o rurales donde hay carencias o no hay disponibilidad de médicos cardiólogos.

Así como, facilitar a los pacientes un acceso más rápido y confiable en sus tratamientos críticos. No obstante, se recomienda que la detección estimada por el clasificador probabilístico sea validad por un médico cardiólogo.

Como futuras investigaciones, se sugiere considerar las dependencias entre las variables, es decir, construir una red Bayesiana que tome en cuenta las relaciones causales entre ellas. Luego, se propone identificar las variables más influyentes en el estudio mediante un análisis multivariable. Además, se destaca que este clasificador probabilístico puede ser aplicado en cualquier otro problema de muestreo. Por último, como trabajo a futuro se pretende utilizar diferentes sensores conectado a la placa Raspberry Pi, por ejemplo, para el análisis de la actividad eléctrica del corazón, entre otras aplicaciones a desarrollar.

Agradecimientos. Los autores agradecen al Centro de Investigación e Innovación en Ciencias de la Computación y Tecnología Educativa (CIICCTE) adscrito a la Facultad de Informática de la UAQ por el espacio brindado para la realización de este trabajo.

Referencias

1. OMS: Enfermedades cardiovasculares. Organización Mundial de la Salud (2021). https://www.who.int/es/health-topics/cardiovascular-diseases (2021).
2. UNAM. Enfermedades del corazón, pandemia permanente. Boletín UNAM de la Dirección de Comunicación Social. https://www.dgc (2020)
3. Schultz, W.M., Kelli, H.M., Lisko, J.C., Varghese, T., Shen, J., Sandesara, P., Sperling, L.S. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. Circulation, vol. 137, no. 20, pp. 2166–2178, (2018). DOI: 10.1161/CIRCULATIONAHA. 117.029652.
4. Huang, W.: Research on user Satisfaction of Older Community Care based on Structure Equation. In: 2012 Fourth International Symposium on Information Science and Engineering, pp. 489–492 (2012). DOI: 10.1109/ISISE.2012.118.
5. Marschollek, M., Gietzelt, M., Schulze, M., Kohlmann, M., Song, B., Wolf, K.H.: Wearable Sensors in Healthcare and Sensor-Enhanced Health Information Systems: All our Tomorrows?. Healthcare Informatics Research, vol. 18, no. 2, pp. 97–104 (2012). DOI: 10.4258/hir.2012. 18.2.97.
6. Durairaj, M., Revathi, V.: Prediction of Heart Disease using Back Propagation MLP Algorithm. International Journal of Scientific & Technology Research, vol. 4, no. 8, pp. 235–239 (2015).
7. Mohan, S., Thirumalai, C., Srivastava, G.: Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access. 7, pp. 81542–81554 (2019). DOI: 10.1109/ ACCESS.2019.2923707.
8. Dwivedi, A.K.: Performance Evaluation of Different Machine Learning Techniques for Prediction of Heart Disease. Neural Computing and Applications, vol. 29, pp. 685–693 (2018). DOI: 10.1007/s00521-016-2604-1.
9. Bonato, P.: Keynote: Digital Health Technologies and their Role in the Development of Precision Rehabilitation Interventions. In: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 200–200 (2021). DOI: 10.1109/PerComWorkshops51409.2021.9431126.
10. Aparicio-Montelongo, I., Celaya-Padilla, J.M., Luna-García, H., Galván-Tejada, C.E., Galván-Tejada, J.I., Rosales, H.G.: Predicción de enfermedades cardíacas derivadas de diabetes, mediante algoritmos genéticos: Caso de estudio. Research in Computing Science, vol. 151, no. 6, pp. 159–172 (2022).

11. Documentación de Raspberry Pi: https://www.raspberrypi.com/documentation/ (2024)
12. Plaza, E.: Tendencias en Inteligencia Artificial: Hacia la cuarta década. Nuevas tendencias en Inteligencia Artificial, pp. 379–425 (1992)
13. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press (2021)
14. Deisenroth, M.P., Faisal, A.A., Ong, C.S.: Mathematics for machine learning. Cambridge University Press (2020)
15. Universidad de California Irvine (UCI): Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/heart+disease (2024)