

Predicción de Fallos Cardiacos usando Machine Learning: Una revisión sistemática de la literatura

Prediction of Heart Failures using Machine Learning: A Systematic Literature Review

Jesús Sebastian Senmache Ly^{1,a}, Yamil Alexander Quiñones Nieto^{2,b}

¹ Universidad Nacional de Trujillo, Facultad de Ciencias Físicas y Matemáticas. Trujillo, Perú

² Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática. Lima, Perú

^a Autor de correspondencia: : t052701720@unitru.edu.pe, ORCID: <https://orcid.org/0009-0006-7858-504>

^b E-mail: yquinonesn@unmsm.edu.pe, ORCID: <https://orcid.org/0000-0003-4474-0556>

Resumen

El desarrollo de la Inteligencia Artificial (IA), hoy está en su auge, y eso despierta el interés de la comunidad científica para hacer estudios usando Machine Learning (ML) una de sus ramas de la IA, que mediante algoritmos o modelos entrenados se puede predecir fallos cardiacos. Según la búsqueda de literatura que se realizó, se encontró que en los estudios se utilizan variables para predecir fallos cardiacos las más utilizadas es la edad, el sexo, la glucosa en ayunas, la presión arterial sistola y el colesterol LDL. Así mismo se realiza un preprocesado de datos, el cual tiene fases y las más usadas son el Reescalamiento, Limpieza de datos, Agrupamiento, Codificación de datos, Detección de valores atípicos. También se ha visto que la mayoría de los estudios proponen sus metodologías y las fases más empleadas son el Preprocesamiento, Aplicación de modelos, Análisis de resultados, Train-test, Clasificación de datos, Selección de modelo o algoritmo.

Palabras clave: Inteligencia Artificial, Machine Learning, fallos cardiacos

Abstract

The development of Artificial Intelligence (AI) is currently at its peak, sparking the interest of the scientific community to conduct studies using Machine Learning (ML), a branch of AI. Through trained algorithms or models, ML can predict heart failures. According to the literature review conducted, it was found that the most commonly used variables in studies to predict heart failures are age, sex, fasting glucose, systolic blood pressure, and LDL cholesterol. Additionally, data preprocessing is carried out, which involves various phases, with the most commonly used being Rescaling, Data Cleaning, Clustering, Data Encoding, and Outlier Detection. It was also observed that most studies propose their own methodologies, with the most frequently employed phases being Preprocessing, Model Application, Results Analysis, Train-test, Data Classification, and Model or Algorithm Selection.

Keywords: Artificial Intelligence, Machine Learning, heart failures

Recibido: 03-10-2024 - Aceptado: 16-12-2024 - Publicado: 30-12-2024

Citar como:

Senmache Ly, J. & Quiñones, Y. (2024). Predicción de Fallos Cardiacos usando Machine Learning: Una revisión sistemática de la literatura. *Revista Peruana de Computación y Sistemas*, 6(2):111-118. <https://doi.org/10.15381/rpcs.v6i2.29141>

© Los autores. Este artículo es publicado por la Revista Peruana de Computación y Sistemas de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. Este es un artículo de acceso abierto, distribuido bajo los términos de la licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) [<https://creativecommons.org/licenses/by/4.0/deed.es>] que permite el uso, distribución y reproducción en cualquier medio, siempre que la obra original sea debidamente citada de su fuente original.

1. Introducción

Los fallos cardiovasculares representan un desafío considerable a nivel global y requieren la atención urgente de la comunidad médica internacional. Según la Organización Mundial de la Salud (OMS), los fallos cardíacos son responsables de aproximadamente 18 millones de muertes anuales en todo el mundo, con una proyección de aumento a 23.6 millones de muertes para el año 2030 [1][4]. Esta creciente prevalencia subraya la necesidad de mejorar las estrategias para la detección y prevención temprana de estos eventos.

En los últimos años, el machine learning ha captado el interés en diversos campos, incluida la medicina, gracias a su capacidad para resolver problemas complejos [2]. Su aplicación en el ámbito clínico puede optimizar la toma de decisiones, algo crucial para garantizar una atención médica de alta calidad. Sin embargo, los médicos se enfrentan a grandes volúmenes de datos, que a menudo son difíciles de manejar manualmente y poco prácticos para el proceso de toma de decisiones [3]. Además, el machine learning no solo facilita este proceso, sino que también permite una detección temprana de fallos cardíacos, lo que reduce tanto el impacto físico en los pacientes como los costos económicos asociados [4].

Actualmente, se reconocen numerosos factores predictivos de fallos cardíacos, los cuales se pueden dividir en dos grandes categorías: disfunciones cardíacas y enfermedades crónicas del adulto [6]. Estos factores son ideales para ser aprovechados por machine learning, ya que permiten utilizar datos clínicos previamente recopilados para realizar análisis predictivos con fines diagnósticos [5]. Con el auge de la inteligencia artificial, se han resuelto muchos problemas basados en datos, abriendo nuevas oportunidades para mejorar el diagnóstico y tratamiento de enfermedades cardíacas [7].

La clasificación y predicción de fallos cardíacos ha sido objeto de múltiples investigaciones, que han explorado una variedad de métodos de machine learning. Desde dolencias comunes hasta enfermedades graves, estas tecnologías permiten a los médicos utilizar una amplia gama de herramientas para el diagnóstico [8]. Un enfoque popular es el uso de algoritmos de aprendizaje supervisado, donde los modelos se entrenan con conjuntos de datos etiquetados para identificar la relación entre las características de entrada y la presencia o ausencia de fallo cardíaco [9].

Si bien existen metodologías estándar para desarrollar proyectos de machine learning, en la investigación actual muchos autores proponen metodologías personalizadas que se adaptan mejor a sus necesidades específicas. Por ello, el presente estudio tiene como objetivo investigar las variables clínicas más relevantes para futuros trabajos, así como las fases

del preprocesamiento de datos y las metodologías más eficaces utilizadas en la predicción de fallos cardíacos.

El presente documento está organizado de la siguiente manera: sección 2, donde se describe la metodología en la investigación, sección 3, se analiza los resultados obtenidos y en la sección 4 se realizan las conclusiones.

2. Metodología

El desarrollo de la revisión sistemática se realizó siguiendo la metodología propuesta por [21], que consiste en 4 fases:

- Preguntas de revisión: Preguntas clave que guían la revisión sistemática y buscan identificar el objetivo principal de la investigación.
- Estrategia de búsqueda: Plan que describe el cómo se localizarán los artículos relevantes.
- Criterio de inclusión y exclusión: Requisitos que determinan qué artículos serán considerados o descartados.
- Proceso de selección: Pasos que se sigue para la revisión de los artículos recuperados y se decide cuales cumplen con los criterios

2.1. Preguntas de revisión

Se plantearon 3 preguntas de investigación con el objetivo de identificar las técnicas de análisis de datos empleadas para la predicción de fallos cardíacos.

P1: ¿Qué variables han sido empleadas para predecir la presencia de fallos cardíacos?

P2: ¿Qué fases se ha utilizado para realizar para el preprocesamiento de datos?

P3: ¿Qué fases se han utilizado para realizar los proyectos de ML para predecir fallos cardíacos?

Para responder a las preguntas anteriormente planteadas, se realizó la búsqueda sistemática de la información en las bases de datos de Scopus, PubMed, Scielo, Sciece Direct.

Se considero como términos de búsqueda (“heart failure prediction”, “data mining”, “machine learning”), entre los años 2022 y 2024. Una vez que la recopilación de artículos relacionados con el tema de estudio, se estableció los criterios de inclusión y exclusión, como se observa en la tabla 1.

2.2. Estrategia de búsqueda

En la Figura 1, se muestra la estrategia que se va a utilizar para la revisión de artículos científicos.

2.3. Criterios de inclusión y exclusión

En la tabla 1, se presenta los criterios de inclusión y exclusión para la selección de artículos.

Figura 1

Niveles de Calidad de uso de la aplicación "SearchM" percibidos por los estudiantes de la asignatura de Inteligencia Artificial en la escuela de ingeniería en computación e informática - UNPRG

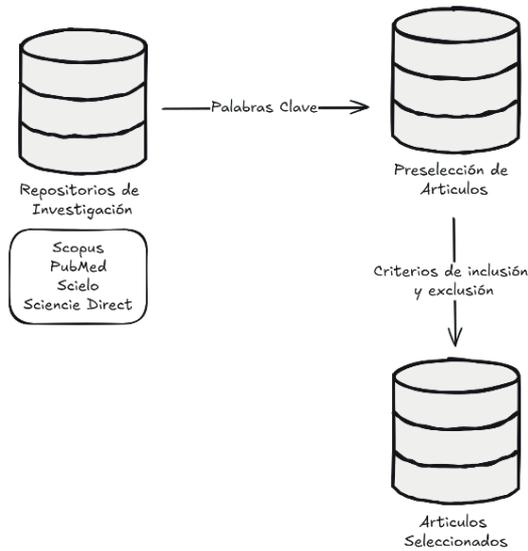


Tabla 1

Criterios de selección y exclusión

Inclusión	Exclusión
Los artículos de investigación deben estar dentro del rango de años desde el 2022 hasta el 2024	Los artículos que no se encuentren en el rango de fechas establecidas
Los artículos deben contener variables, técnicas y metodologías para predecir fallos cardíacos	Los artículos que no respondan a las preguntas de investigación planteadas
Los artículos deben estar publicados en revistas reconocidas por Scimago	Los artículos que no estén publicados en revistas indexadas en Scimago o que provengan de conferencias, repositorios sin revisión

2.4. Proceso de selección

En la Tabla 2, se muestran que en un inicio se encontraron 72 artículos relacionados con el tema de estudio, teniendo en cuenta los criterios de inclusión y exclusión se preseleccionaron 41 y se seleccionaron 20.

2.5. Análisis

En esta sección, se responderá a las preguntas planteadas, por esta razón, se analizará minuciosamente todos los artículos de investigación seleccionados.

Tabla 2

Artículos encontrados

Fuente	Artículos encontrados	Artículos preseleccionados	Artículos seleccionados
PubMed	15	7	1
Scopus	35	20	12
Scielo	7	4	0
Science Direct	15	10	7

2.5.1. ¿Qué variables han sido empleadas para predecir la presencia de fallos cardíacos?

Las variables clínicas, son extraídas de data sets que ya fueron usados en investigaciones para machine learning, después de la extracción, limpieza y depuración de los datos obtenidos, se seleccionan las principales para el respectivo estudio.

En las investigaciones [1], [2] y [7], se utilizó el UCI Heart Disease dataset, el cual tiene 76 atributos, pero en esta investigación se utilizó solamente 14 variables. En el artículo [3], se utilizó el récord medico de pacientes hospitalizados en el hospital Sheba Medical Center de Israel, entre 2007 y 2017. En el trabajo de investigación [4] y [18], se utilizó el Kaggle Cardiovascular Disease Dataset, que consiste en el historial de 70000 pacientes y de 11 variables. En la investigación [5], se utilizó la combinación de las variables de 5 datasets (Combined, Z-AlizadehSani dataset, Framingham heart study, SA heart dataset, Cleveland UCI), las cuales tienes 13, 53, 15, 9, 13 variables respectivamente, de las cuales se seleccionaron para el estudio 9,34,9,7,9 respectivamente.

Los fallos cardíacos son causados por diferentes factores, que incluyen la genética, el estilo de vida, y la presencia de enfermedades tanto cardíacas como no cardíacas. A lo largo de los años, los datos recopilados fueron de mucha ayuda para muchos investigadores campo de la medicina como del análisis de datos, ya que se almacenan en base de datos en internet o de hospitales. En el trabajo [6] y [17], se utilizó el MIMIC-IV dataset, que contiene pacientes con el historial medico de los pacientes. En el trabajo [8], [9] y [20], se utilizó el Heart disease Cleveland UCI dataset, que contiene datos con 13 variables. En la investigación [10], se utiliza el MIMIC-III dataset, que contiene 14568 diagnósticos. En el trabajo [11], se utilizó la combinación de 5 datasets (Cleveland, Hungary, Switzerland, Long Beach VA, Stalog (liver)), que tienen en común sus 13 variables. En el trabajo de investigación [12], se empleó su propio dataset, que contiene 7 variables, que va desde la edad hasta depresión del ST.

En el trabajo [13], las variables que se usaron fueron tomadas de referencia de The National Institute for Cardiovascular Outcomes Research (NICOR), mientras que el dataset que utilizaron es privado. En el trabajo [14], se utilizó un dataset de 308205 personas, y se tomaron las características clínicas mas importantes, como la circunferencia abdominal, la hemoglobina, ácido úrico, etc. En la investigación [15], se utilizó un dataset de 540 personas de indonesia con sus 21 características clinicas. En la investigación [16], se recolecto información en Framingham, Massachusetts, se trabajó con un dataset que contiene 4238 pacientes y 15 variables. En el trabajo de investigación [19], se utilizó el dataset de HeartCarer, que contiene 2,992 pacientes y 13 variables clinicas.

Tabla 3

Variables clínicas empleadas para predecir la presencia de fallos cardiacos

VARIABLES CLÍNICAS	FUENTES
Edad (Age)	[1] [2] [3] [4] [5] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20]
Sexo (Sex/Gender)	[1] [2] [3] [4] [7] [8] [9] [10] [11] [13] [14] [16] [17] [18] [19] [20]
Glucosa en ayunas	[1] [2] [3] [4] [5] [7] [9] [10] [11] [14] [15] [16] [17] [18] [19] [20]
Presión Arterial sistola	[1] [2] [4] [5] [8] [9] [10] [11] [12] [14] [15] [16] [18] [19] [20]
Colesterol LDL	[1] [2] [4] [8] [9] [10] [11] [12] [13] [14] [17] [20]
Depresión del ST	[1] [2] [5] [7] [8] [9] [10] [11] [12] [20]
Frecuencia cardíaca máxima	[1] [2] [5] [8] [9] [10] [11] [12] [15] [20]
Angina inducida por ejercicio	[1] [2] [5] [7] [8] [9] [10] [11] [19] [20]
Electrocardiograma en reposo	[1] [2] [7] [8] [9] [10] [11] [19] [20]
Número de vasos coloreados	[1] [2] [5] [7] [8] [9] [19] [20]
Presión arterial diastólica (dbp)	[14] [15] [16] [18] [19] [20]
Hemoglobina	[3] [10] [13] [14] [15] [17]
Frecuencia cardíaca (Heart Rate)	[9] [10] [15] [16] [19] [20]
Creatinina	[3] [10] [13] [15] [17]
WBC (Células blancas)	[3] [10] [13] [15] [17]
IMC (Índice de masa corporal)	[14] [15] [16] [17]
Potasio/Kalium	[3] [15] [17]
Urea	[3] [15] [17]
Triglicéridos	[14] [17]
Plaquetas	[3] [17]
Colesterol HDL	[14] [17]
Bilirrubina Total	[3] [10]
AST (Aspartato aminotransferasa)	[3] [14]
Circunferencia abdominal	[14]
Ácido úrico	[14]
Volumen corpuscular medio	[3]
Hemoglobina corpuscular media	[3]
Fosfatasa alcalina	[3]
Eosinófilos	[3]
Basófilos	[3]
Procalcitonina	[3]
Fracción de eyección del ventrículo izquierdo	[15]
Distribución del tamaño de glóbulos rojos	[3]
Frecuencia respiratoria	[15]
Creatina quinasa	[17]
Concentración corpuscular media de hemoglobina	[3]
Volumen plaquetario medio	[3]

En la tabla 3, se puede observar 37 variables clínicas de distintos trabajos de investigación, donde el objetivo es predecir la presencia de fallos cardiacos de manera binaria. Las variables clínicas más usadas es la edad, el sexo, la glucosa en ayunas, la presión arterial sistola y el colesterol LDL.

2.5.2. ¿Qué fases se ha utilizado para realizar para el preprocesamiento de datos?

Para el trabajo [1], para la parte del preprocesamiento de los datos, se consideran la limpieza de datos, Codificación de datos categóricos y reescalamiento. Para la investigación [2], se realizó la detección de valores atípicos, Codificación de variables y escalado para los dataset. Para el artículo de investigación [3], se usó el Agrupamiento, Filtro de valores faltantes, Eliminación de Outlier, Winsorización y Imputación de datos. Para el trabajo [4], se utilizó la eliminación de Outliers, Verificación de distribución, Sección y Reducción. Para el trabajo de investigación [5], se empleó, Codificación de datos, Estandarización de datos y Regularización de datos. Para la investigación [6], se empleó la Generación de subgrupos, Asignación de clases y Selección de características. Para el trabajo [7], se consideró la Normalización, Categorización Eliminación de datos faltantes, División de datos. Para la investigación [8], se trabajó con Escalamiento MinMax y la Eliminación de datos faltantes. Para el trabajo [9], se consideró Limpieza de datos como eliminación de datos faltantes, Transformación de datos como Codificación One-hot, Normalización y Estandarización usando escalamiento Min-Max. Para el articulo [10], se utilizó Limpieza de datos, Estandarización, Clasificación y Clustering.

Para la investigación [11], se tuvo en cuenta solamente la normalización para los 12 atributos. Para el trabajo de investigación [12], se tomó la Limpieza de datos, Visualización de mapa de correlación, Eliminación de variable y Clustering. Para el trabajo [13], se consideró Limpieza de datos, Estandarización y Clustering. Para la investigación [14], se tomó en cuenta Limpieza de datos, Filtración y Agrupación. Para la investigación [15], se empleó Cohorte retrospectivo del dataset, Manejo de datos faltantes, y Agrupamiento. Para el trabajo [16], se empleó Limpieza de valores faltantes y duplicados, Discretización de atributos, Normalización de valores, Manejo de Desbalanceo de Clases.

Para la investigación [17], utilizó Conversión de datos categóricos, Manejo de valores atípicos y Manejo de valores faltantes. Para el trabajo [18], se utilizó Extracción de características relevantes, Normalización de datos, Clasificar características. Para la investigación [19], se empleó Peso local de los factores de Riesgos y Peso global de los factores de Riesgos. Para el trabajo de investigación [20], Se utilizo eliminación de valores vacíos.

En la tabla 4, se observar que son 25 fases de preprocesamiento de datos, las utilizadas en los trabajos de investigación, las fases más empleadas son el Reescalamiento, Limpieza de datos, Agrupamiento, Codificación de datos, Detección de valores atípicos.

Tabla 4

Fases empleadas para el preprocesamiento de datos para predecir fallos cardiacos ordenados secuencialmente

Fases	Fuentes
Limpieza de datos (eliminación de datos faltantes y duplicados)	[1][4][9][10][12][13][14][16][20]
Manejo de datos faltantes	[15][17]
Imputación de datos	[3]
Detección de valores atípicos (Outliers)	[2][3][4][16][17]
Eliminación de Outliers	[3][4]
Winsorización	[3]
Verificación de distribución	[4]
Reescalamiento (normalización, estandarización, escalado MinMax)	[1][2][5][7][8][9][10][11][13][16][18]
Regularización de datos	[5]
Codificación de datos (variables categóricas)	[1][2][5][9][17]
Discretización de atributos	[16]
Selección de características	[6][18]
Extracción de características relevantes	[18]
Eliminación de variables	[12]
Reducción de datos	[4]
Generación de subgrupos	[6]
Asignación de clases	[6]
Manejo de desbalanceo de clases	[16]
División de datos	[7]
Cohorte retrospectivo del dataset	[15]
Visualización de mapa de correlación	[12]
Agrupamiento (Clustering)	[3][10][12][13][14][15]
Clasificación	[10][18]
Peso local y global de factores de riesgos	[19]
Filtración	[14]

2.5.3. ¿Qué fases se han utilizado para realizar los proyectos de ML para predecir fallos cardiacos?

Hoy en día, existen metodologías que se utilizan para proyectos de machine learning, pero para este tema en particular la mayoría investigadores optan por proponer sus propias metodologías que se adapten a sus necesidades, así que se enumeraran sus respectivas fases en orden.

Para el trabajo [1], se empleó el Preprocesamiento, Análisis del dataset, Train-Test, Modelo propuesto-CNN. Para la investigación [2], se propuso Preprocesamiento de datos, Modelos de clasificación, Clasificación y Resultados. En el trabajo de investigación [3], se utilizó las siguientes fases, Preparación de los datos, K-means Cluster analisis, Agrupamiento de períodos consecutivos similares, Generación patrones migratorios, Post-procesado. Para la investigación [4], se empleó el Ingreso de datos, Preprocesado, Aplicación de análisis K-mode, Clasificación de datos, Train-Model, Probar modelo. Para el artículo [5], se utilizó preprocesamiento, Selección, Clasificación de algoritmos, Evaluación. Para la investigación [6], se empleó el Preprocesado, Aplicación de métodos de análisis, Análisis de resultados. Para el trabajo de investigación [7], se utilizó el Preprocesado, Entrenamiento del Modelo, Test-Model, Análisis de resultados. En el artículo [8], se propone usar, Preprocesamiento, Selección, Clasificación, Análisis de

resultados. En el trabajo [9], se utilizó el Preprocesado, Clasificación, Train-Model, Evaluación de modelo, Análisis de resultados. En la investigación [10], se emplea el Preprocesamiento, Exploración, Train-Model, Análisis de resultados.

En el Trabajo [11], se utiliza el Preprocesamiento, División de dataset, Train-Dataset, Validación, Análisis. En el artículo [12], se empleó el Preprocesamiento, aplicación de Árbol de Decisión, Evaluación, Conclusiones. Para el trabajo [13], se empleó el Preprocesamiento, Análisis de agrupamiento, Aplicación de técnica de ML, Análisis de Resultados. En la investigación [14], se propuso utilizar la metodología LAMP para la minería de datos, Aplicación del método random-forest, Análisis de datos. En el trabajo [15], se utilizó el Preprocesamiento de datos, Selección de modelo, Train-Model, Evaluación de resultados. En la investigación [16], se propuso utilizar Bussiness Understanding, Data Understanding, Data Preparation, Modeling, Evalution, que pertenecen a la metodología CRISP-DM.

En el trabajo [17], Procesamiento de datos, Sección de Deep Learning, Sección de Machine Learning, en este trabajo se compara la eficiencia de los dos modelos, la última fase es la Comparación. En el trabajo [18], Preprocesamiento, Aplicación de algoritmo propuesto, Test-Model, Análisis de resultados. En la investigación [19], se empleó el Preprocesamiento de datos, Detección de anomalías, Training-Data, Testing-Data, Aplicación del modelo, Análisis de resultados. En el trabajo [20], se utilizó las siguientes fases: Preprocesamiento, Selección, Creación y Train-Model, Evaluación de los resultados.

En la Tabla 5, se visualiza las 18 fases utilizadas durante la realización de proyectos de machine learning para predecir fallos cardiacos, se muestra las fases ordenadas por número de incidencias.

Tabla 5

Fases empleadas para realizar proyectos de Machine Learning para predecir fallos cardiacos ordenados secuencialmente

Fases	Fuentes
Bussiness Understanding, Data Understanding (CRISP-DM)	[16]
Análisis del dataset / Exploración	[1][10]
Preprocesamiento de datos	[1][2][4][5][6][7][8][9][10][11][12][13][15][16][18][19][20]
División de dataset	[11]
Train-Test (incluyendo Train-Model, Testing-Model)	[1][4][7][9][10][11][15][19][20]
Selección de modelo / algoritmos	[5][8][15][16][20]
Modelo propuesto / Aplicación de modelos (CNN, K-means, K-mode, etc.)	[1][3][4][6][7][9][12][13][14][15][18][19]
Análisis de agrupamiento (K-means, K-mode, etc.)	[3][4][13]
Aplicación de técnicas de minería de datos (LAMP, random-forest)	[14]
Generación de patrones	[3]
Clasificación de datos	[2][4][5][8][9][13]
Detección de anomalías	[19]
Post-procesado	[3]
Evaluación / Análisis de resultados	[5][6][7][8][9][10][12][13][15][18][19][20]
Sección de Deep Learning / Machine Learning	[17]
Comparación de modelos	[17]
Conclusiones	[12]

Figura 4

Frecuencia de fases empleadas en el preprocesamiento de los datos en los trabajos presentados

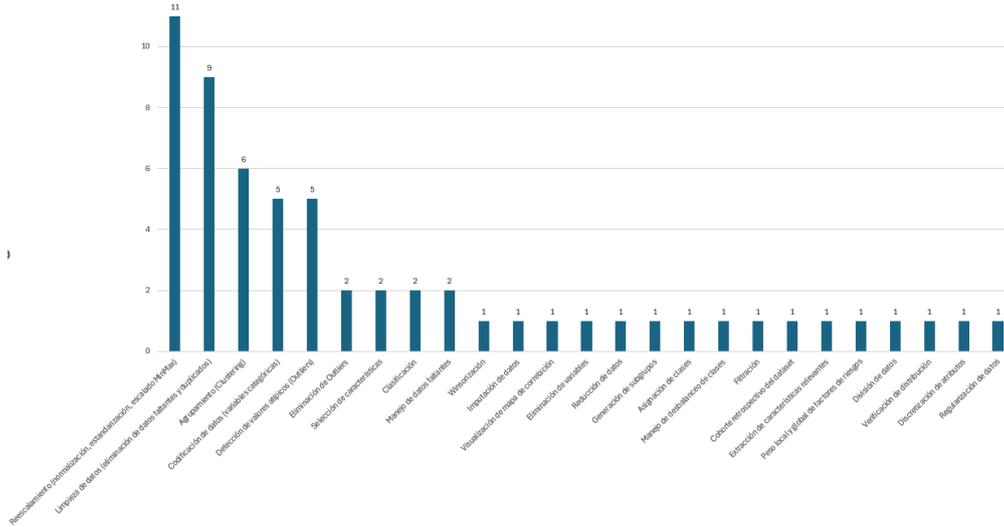
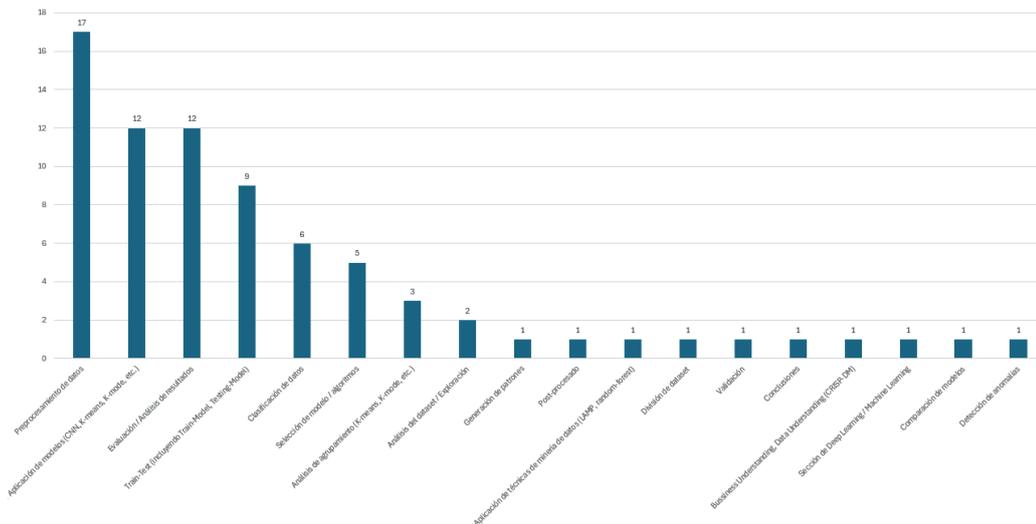


Figura 5

Frecuencia de uso de las fases empleadas en la metodología en los trabajos investigados



En la Figura 4, se observa en el gráfico las 25 fases utilizadas en el preprocesamiento de los datos en las diferentes investigaciones para predecir fallos cardíacos. Como se detalló, las fases más empleadas son: reescalamiento, limpieza de datos, agrupamiento, codificación de datos, detección de valores atípicos.

En la Figura 5, se muestra el gráfico donde se visualiza las 18 fases de las metodologías utilizadas en los trabajos de investigación. Las más utilizadas son: Preprocesamiento, aplicación de modelos, análisis de resultados, train-test, clasificación de datos, selección de modelo o algoritmo.

4. Conclusiones

En el presente trabajo se presenta una revisión sistemática de literatura para la predicción de fallos cardíacos usando Machine Learning. Por tal motivo, se buscó y recopiló 72 artículos relacionados con el tema de estudio, que después de aplicarle los criterios

de inclusión y exclusión, quedaron seleccionados 20 artículos.

En estos artículos se encontraron las variables usadas para predecir fallos cardíacos, las fases del preprocesado de los datos y las fases de las metodologías propuestas. Las variables clínicas más utilizadas es la edad, el sexo, la glucosa en ayunas, la presión arterial sistola y el colesterol LDL. Las fases más empleadas en el preprocesamiento de datos son el Reescalamiento, Limpieza de datos, Agrupamiento, Codificación de datos, Detección de valores atípicos. Las fases más utilizadas en las metodologías propuestas son el Preprocesamiento, Aplicación de modelos, Análisis de resultados, Train-test, Clasificación de datos, Selección de modelo o algoritmo.

En futuros trabajos se recomiendan usar las variables, fases para el preprocesamiento y fases de las metodologías encontradas. Adicionalmente se recomienda que se elija las fases dependiendo de los datos y método de trabajo.

Referencias

- [1] Honi, D. G., & Szathmary, L. (2024). A one-dimensional convolutional neural network-based deep learning approach for predicting cardiovascular diseases. *Informatics in Medicine Unlocked*, 49, 101535. <https://doi.org/10.1016/j.imu.2024.101535>
- [2] Mohapatra, S., Maneesha, S., Mohanty, S., Patra, P. K., Bhoi, S. K., Sahoo, K. S., & Gandomi, A. H. (2023). A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease. *Healthcare Analytics*, 3, 100133. <https://doi.org/10.1016/j.health.2022.100133>
- [3] Ramon-Gonen, R., Heart, T., Ben-Assuli, O., Shlomo, N., & Klempfner, R. (2022). Disease evolution and risk-based disease trajectories in congestive heart failure patients. *Journal of Biomedical Informatics*, 125, 103949. <https://doi.org/10.1016/j.jbi.2021.103949>
- [4] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2), 88. <https://doi.org/10.3390/a16020088>
- [5] Atimbire, S. A., Appati, J. K., & Owusu, E. (2024). Empirical exploration of whale optimisation algorithm for heart disease prediction. *Scientific Reports*, 14(1), 4530. <https://doi.org/10.1038/s41598-024-54990-1>
- [6] Jung, J., Kim, D., & Hwang, I. (2024). Exploring Predictive Factors for Heart Failure Progression in Hypertensive Patients Based on Medical Diagnosis Data from the MIMIC-IV Database. *Bioengineering*, 11(6), 531. <https://doi.org/10.3390/bioengineering11060531>
- [7] Alfadli, K. M., & Almagrabi, A. O. (2023). Feature-Limited Prediction on the UCI Heart Disease Dataset. *Computers, Materials & Continua*, 74(3). <https://doi.org/10.32604/cmc.2023.033603>
- [8] Shrivastava, P. K., Sharma, M., & Kumar, A. (2023). HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithms. *Measurement: Sensors*, 25, 100657. <https://doi.org/10.1016/j.measen.2022.100657>
- [9] Gayathri, R., Sangeetha, S. K. B., Mathivanan, S. K., Rajadurai, H., Malar, B. A., Mallik, S., & Qin, H. (2024). Enhancing heart disease prediction with reinforcement learning and data augmentation. *Systems and Soft Computing*, 6, 200129. <https://doi.org/10.1016/j.sasc.2024.200129>
- [10] Liang, Y., & Guo, C. (2023). Heart failure disease prediction and stratification with temporal electronic health records data using patient representation. *Biocybernetics and Biomedical Engineering*, 43(1), 124-141. <https://doi.org/10.1016/j.bbe.2022.12.008>
- [11] Mesran, M., Kmurawak, R. M., & Windarto, A. P. (2024). Integration of PSO-based advanced supervised learning techniques for classification data mining to predict heart failure. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 22(1), 76-85. <http://doi.org/10.12928/telkomnika.v22i1.25357>
- [12] Arifuddin, A., Buana, G. S., Vinarti, R. A., & Djunaidy, A. (2024). Performance Comparison of Decision Tree and Support Vector Machine Algorithms for Heart Failure Prediction. *Procedia Computer Science*, 234, 628-636. <https://doi.org/10.1016/j.procs.2024.03.048>
- [13] Soltani, F., Jenkins, D.A., Kaura, A. et al. Phenogrouping heart failure with preserved or mildly reduced ejection fraction using electronic health record data. *BMC Cardiovasc Disord* 24, 343 (2024). <https://doi.org/10.1186/s12872-024-03987-9>
- [14] Miyashita, Y., Hitsumoto, T., Fukuda, H., Kim, J., Washio, T., & Kitakaze, M. (2023). Predicting heart failure onset in the general population using a novel data-mining artificial intelligence method. *Scientific Reports*, 13(1), 4352. <https://doi.org/10.1038/s41598-023-31600-0>
- [15] Indriany, F. E., Siregar, K. N., Purwowyoto, B. S., Siswanto, B. B., Sutedja, I., & Wijaya, H. R. (2024). Predicting the Risk of Severity and Readmission in Patients with Heart Failure in Indonesia: A Machine Learning Approach. *Healthcare Informatics Research*, 30(3), 253-265. <https://doi.org/10.4258/hir.2024.30.3.253>
- [16] Neto, C., Ferreira, D., Ramos, J., Cruz, S., Oliveira, J., Abelha, A., & Machado, J. (2022). Prediction models for coronary heart disease. In *Distributed Computing and Artificial Intelligence*, Volume 1: 18th International Conference 18 (pp. 119-128). Springer International Publishing. https://doi.org/10.1007/978-3-030-86261-9_12
- [17] Huang, J., Cai, Y., Wu, X., Huang, X., Liu, J., & Hu, D. (2024). Prediction of Mortality Events of Patients with Acute Heart Failure in Intensive Care Unit Based on Deep Neural Network. *Computer Methods and Programs in Biomedicine*, 108403. <https://doi.org/10.1016/j.cmpb.2024.108403>
- [18] Theerthagiri, P. (2022). Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique. *Intelligent Systems with Applications*, 16, 200121. <https://doi.org/10.1016/j.iswa.2022.200121>
- [19] Zhou, C., Hou, A., Dai, P., Li, A., Zhang, Z., Mu, Y., & Liu, L. (2023). Risk factor refinement and ensemble deep learning methods on prediction of heart failure using real healthcare records. *Information Sciences*, 637, 118932. <https://doi.org/10.1016/j.ins.2023.04.011>
- [20] Zhou, W., Liu, H., Zhou, R., Li, J., & Ahmadi, S. (2024). An optimal method for diagnosing heart disease using combination of grasshopper evolutionary algorithm and support vector machines. *Heliyon*, 10(9). <https://doi.org/10.1016/j.heliyon.2024.e30363>
- [21] Duo, X. U., & Zeshui, X. U. (2024). Machine learning applications in preventive healthcare: A systematic literature review on predictive analytics of disease comorbidity from multiple perspectives. *Artificial Intelligence in Medicine*, 102950. <https://doi.org/10.1016/j.artmed.2024.102950>