*South Health and Policy.* 2026; 5:364
doi: 10.56294/shp2026364

**REVISIÓN**

# Prediction of Cardiovascular Diseases Using Machine Learning Models

# Predicción de Enfermedades Cardiovasculares mediante Modelos de Aprendizaje Automático

Michael Rafael Rodríguez Rodríguez[1], Claudia Alejandra Delgado Calpa[1], Héctor Andrés Mora Paz[1]

[1]Universidad CESMAG, Facultad de Ingeniería, Ingeniería de Sistemas. Pasto, Colombia.

**Citar como**: Rodríguez Rodríguez MR, Delgado Calpa CA, Mora Paz HA. Prediction of Cardiovascular Diseases Using Machine Learning Models. South Health and Policy. 2026; 5:364. https://doi.org/10.56294/shp2026364

**Enviado:** 09-02-2025 **Revisado:** 08-06-2025 **Aceptado:** 24-12-2025 **Publicado:** 01-01-2026

**Editor:** Dr. Telmo Raúl Aveiro-Róbalo

**Autor para la correspondencia:** Michael Rafael Rodríguez Rodríguez

## ABSTRACT

The study addressed the global problem of cardiovascular diseases, which were one of the leading causes of mortality and morbidity according to the World Health Organisation. Multiple risk factors, both modifiable and non-modifiable, were identified, and the need to implement technologies that would enable early and accurate detection was emphasised. Given this scenario, the use of machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN), combined with traditional and alternative kernel functions, was proposed. A comparative approach was developed to validate the hypothesis that under-explored kernel functions could improve predictive performance in terms of accuracy and response time. To this end, models were trained with data extracted from recognised platforms such as Kaggle and UCI, and metrics such as accuracy, recall and F1-score were applied. The models were adjusted with hyperparameter optimisation techniques using random search. The results demonstrated that certain alternative kernel functions offered improvements in the error-time ratio, in some cases outperforming conventional kernels. The research not only contributed methodological advances in the development of predictive models, but also provided a support tool for clinical decision-making, particularly useful in contexts where timely diagnosis is crucial. Finally, the project contributed to strengthening artificial intelligence in public health, promoting well-being through the prevention and proactive management of cardiovascular diseases.

**Keywords:** Cardiovascular Diseases; Machine Learning; Kernel Functions; SVM; Preventive Diagnosis.

## RESUMEN

El estudio abordó la problemática global de las enfermedades cardiovasculares, las cuales representaron una de las principales causas de mortalidad y morbilidad según la Organización Mundial de la Salud. Se identificaron múltiples factores de riesgo, tanto modificables como no modificables, y se enfatizó en la necesidad de implementar tecnologías que permitieran una detección temprana y precisa. Frente a este panorama, se propuso el uso de algoritmos de aprendizaje automático como Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN), combinados con funciones kernel tradicionales y alternativas. Se desarrolló un enfoque comparativo para validar la hipótesis de que funciones kernel poco exploradas podrían mejorar el rendimiento predictivo en cuanto a exactitud y tiempo de respuesta. Para ello, se entrenaron modelos con datos extraídos de plataformas reconocidas como Kaggle y UCI, y se aplicaron métricas como precisión, recall y F1-score. Los modelos fueron ajustados con técnicas de optimización de hiperparámetros mediante búsqueda aleatoria. Los resultados demostraron que ciertas funciones kernel alternativas ofrecieron mejoras en la relación error-tiempo, superando en algunos casos a los kernel convencionales. La investigación no solo aportó avances metodológicos en el desarrollo de modelos predictivos, sino que también ofreció una

herramienta de apoyo para la toma de decisiones clínicas, especialmente útil en contextos donde el diagnóstico oportuno es crucial. Finalmente, el proyecto contribuyó al fortalecimiento de la inteligencia artificial en salud pública, promoviendo el bienestar mediante la prevención y el manejo proactivo de enfermedades cardiovasculares.

**Palabras clave:** Enfermedades Cardiovasculares; Aprendizaje Automático; Funciones Kernel; SVM; Diagnóstico Preventivo.

## INTRODUCCIÓN

Las enfermedades cardiovasculares, según la OMS, son uno de los mayores problemas de salud pública a nivel mundial, siendo la enfermedad cerebro vascular la primera causa de mortalidad y morbilidad, conforme estimación se cobra 17,9 millones de vidas cada año.[1,2,3,4,5,6]

Además, los problemas principales que se presentan en las personas son infartos, ataque cerebrovascular, síndrome metabólico, cardiopatías, e hipertensión arterial, son causados por sobrepeso, obesidad, dislipidemia, tabaquismo, inactividad física, alimentación no saludable, estos son factores de riesgo existentes que se pueden modificar, y existen aquellos que no se pueden modificar, pero que son un factor causal para el desarrollo de una enfermedad cardiovascular como son la edad, género, antecedentes personales de enfermedad cardiovascular y antecedentes de familiares con ECV prematuros solo cuando hayan ocurrido en primer grado.[7,8,9,10]

Muchos son los retos que afrontan las personas con este tipo de patologías, que van desde de la calidad de atención en salud hasta las tecnologías concernientes para su monitorización y cuidado. Gran cantidad de usuarios con enfermedades cardiovasculares necesitan estancias hospitalarias frecuentes, puesto que la restringida tecnología concerniente a la salud está concentrada en los centros médicos, principalmente porque adquirirlos tiene un alto costo, al igual que su mantenimiento, sin dejar a un lado el hecho de que se necesita de un conocimiento especializado para interpretar y analizar los datos capturados por estos equipos de alta complejidad.[4] Igualmente, la ausencia de tecnologías guiadas a la detección y prevención de eventos cardiovasculares peligrosos para la salud nutren la elevada tasa de muertes y hospitalizaciones.[11,12,13,14,15]

Un precedente que se debe tener en cuenta es el aprendizaje de máquina y su aplicación en el campo de la salud, donde este se ha ido incrementando aproximadamente en un 30 % a 40 % año tras año con ayuda del análisis estadístico y probabilístico de otras investigaciones. Aun así, el reto de recopilación y manipulación de datos es complejo, tal cual lo es la predicción de muerte al padecer alguna enfermedad, y el desafío es aún mayor cuando se trata del ser humano y órganos tan complejos como el corazón o el sistema que lo compone. Pero el ser diagnosticado no es un final decisivo para las personas al existir varios métodos para mantener un corazón saludable.[8]

Por todo lo anteriormente dicho, se busca abordar en primera instancia la problemática desde una etapa previa para contribuir a que no se origine una patología mayor, por lo tanto, se ve la necesidad de utilizar modelos de IA que resulten útiles para el apoyo diagnóstico oportuno de enfermedades cardiovasculares. Aun así, los modelos de aprendizaje automático conocidos hasta ahora que se usan para predecir el diagnóstico de insuficiencia cardíaca logran resultados con una precisión variable, entre el 80 % y 90 % en sus mejores casos.[8]

Así pues, existen diferentes técnicas de aprendizaje automático que son idóneas para identificar las características más importantes tras procesar grandes cantidades de datos, logrando así una predicción y mejorando algunos sistemas.[9]

Por tanto, el conjunto de datos a utilizar ha sido empleado para entrenar modelos con algunas técnicas de aprendizaje de máquina, pero hay técnicas que aún no se han utilizado y que podrían arrojar resultados importantes.[16,17,18]

Por otro lado, es conveniente que las personas generen conciencia de detectar oportunamente alguna enfermedad crónica, antes de que se manifiesten algunos síntomas y llegando al caso de no obtener un diagnóstico oportuno en cualquier momento de la vida.[10]

Por lo tanto, con este estudio se pretende validar la hipótesis de que es posible encontrar mejoras en los modelos de predicciones de enfermedades cardiovasculares mediante la inspección del desempeño de los algoritmos SVM y ANN, introduciendo funciones kernel que no se hayan desarrollado y utilizado para este caso, determinando que este resultado final sea útil para estudios relacionados con el objeto de estudio.[11]

En efecto, el aprendizaje de máquinas ha demostrado su valor de aplicación en contextos médicos, siendo una herramienta novedosa y alternativa que apoya tareas complejas como el diagnóstico de enfermedades. Estas tecnologías pueden garantizar la seguridad de mejor manera más efectiva, mediante una detección rápida y confiable de la enfermedad, e incluso puedan aliviar las preocupaciones de toda una población.

Por consiguiente, si no se implementa el aplicativo ni se realiza el estudio de comparación de las funciones kernel en el marco de apoyo al diagnóstico médico, se perdería la oportunidad de que las personas tomen

conciencia de la significancia de los efectos que sus hábitos y estilos de vida tienen hasta el momento.[19,20,21]

Es conveniente destacar que, probablemente, se pierda la sensibilización de la población en la prevención del riesgo cardiovascular, lo cual podría llevar a un aumento en la tasa de mortalidad y de mortalidad prematura por estas enfermedades.

En definitiva, mantener el contacto entre médicos y pacientes desde un enfoque de bienestar y estado de salud permite que haya una mayor y mejor cercanía, como se menciona en algunos artículos, el uso de estas tecnologías basadas en Inteligencia Artificial, complementan el conocimiento de los médicos y permite a estos pasar más tiempo con sus pacientes y mejorar el proceso de toma de decisiones compartido.[12] De no implementar el modelo predictivo, se perdería la posibilidad también de aplicar estas nuevas técnicas a tiempo, aumentaría los riesgos y se perdería la posibilidad de realizar un diagnóstico oportuno, por otro lado, no habría probabilidad de analizar grandes cantidades de datos rápidamente con coherencia y precisión, no se podría crear ese modelo de estudio tan importante para el conocimiento y además no se podría contribuir en ese estado de salud de cada persona, y que de carecer de salud cualquier individuo le sería muy difícil lograr algún éxito.[22,23,24]

Las enfermedades cardiovasculares son enfermedades de interés en salud pública y de las cuales mayormente produce muertes y deterioro progresivo de la salud, de las cuales no son diagnosticadas oportunamente, sino cuando estas se presentan en un estado avanzado, resulta de especial interés obtener un diagnóstico oportuno de estas enfermedades, y qué porcentaje de rangos de riesgo cardiovascular se presentan en mayor frecuencia, que van ligadas a factores que se pueden a futuro modificar, permitiendo adoptar medidas oportunas además que contribuye a promover la salud, y prevenir la enfermedad.[25,26,27,28]

El diagnóstico temprano de ECV es decisivo para disminuir las cifras de mortalidad y morbilidad, así como los riesgos que puedan alertar de manera más oportuna, por esta razón se han propuesto varios estudios para detectar enfermedades y riesgos cardiovasculares algunos a partir del procesamiento de la señal de electrocardiografía (ECG) y otros desde mediciones generales de salud pública, en ambos casos apoyados en técnicas de aprendizaje automático (Machine Learning).

Dentro de los distintos tipos de aprendizaje automático, se deriva el aprendizaje basado en kernel (Kernel Learning) o núcleos, el cual es un método para el análisis de patrones, cuyo algoritmo de aplicación más conocido es Máquinas de Soporte Vectorial (SVM). La tarea general del análisis de patrones es encontrar y estudiar tipos generales de relaciones en conjuntos de datos.[13]

Por ejemplo, por medio de diferentes modelos de aprendizaje profundo, como RNN (Red Neuronal Recurrente), LSTM (Red de memoria a corto plazo) e incluso CNN (Red neuronal convolucional) lograron resultados de hasta 96 % de precisión para detectar diferentes tipos de enfermedades cardiacas, entre las que se encuentran taquicardias ventriculares, fibrilación auricular y taquicardia sinusal,[4,14] otro ejemplo en el cual proponen un método novedoso a partir de un modelo de Random Forest hibrido con una componente lineal que logró una precisión de 88,7 % en la predicción de enfermedades cardiovasculares.[4,15]

Este estudio nace de la necesidad de aplicar un modelo predictivo basado en un análisis exhaustivo de las configuraciones en los algoritmos SVM y ANN, considerando que más características del kernel se aproximan al modelo óptimo de aprendizaje automático, haciendo uso de dataset previamente analizados, y recolección de información, para entrenamiento de algoritmos de predicción, para conocer el porcentaje en el rango de riesgo de enfermedades cardiovasculares, y que factores se puedan intervenir oportunamente por el médico.

La investigación será útil a nivel social porque brinda información a medida del estado de salud de la población y apoyo al profesional médico y cuando este puede de manera oportuna identificar si hasta el momento algún factor externo o interno influye en su calidad de vida, y evitar un deterioro que puede ser prevenible.

El resultado será una contribución a la inteligencia artificial en el campo del aprendizaje automático, especialmente en el área de estudios comparativos de algoritmos. Esto incluirá una ruta de comparación, la implementación de las funciones kernel, modelos entrenados para hacer predicciones de enfermedades cardiovasculares y la visualización de dichos datos.

Finalmente, este proyecto se sintoniza con la dinámica de promoción de la salud y prevención de la enfermedad que viene trabajando Colombia en sus diferentes marcos contextuales, además que la enfermedad cardiovascular no cataloga y la puede padecer cualquier persona en su mayoría de edad, por eso es importante el estudio, análisis de los posibles factores variantes según sea, el índice de masa corporal, perímetro abdominal, el nivel socioeconómico en que se encuentra, hábitos, antecedentes de enfermedades personales y familiares, entre otros, además se obtiene un mejoramiento en funciones kernel, para propiciar la integridad en los modelos predictivos.

**Delimitación**

Este proyecto se desarrolló mediante experimentos sobre base de datos obtenidos de Herramientas de repositorio como datasets Kaggle, y datasets UCI, los modelos serán entrenados usando Redes Neuronales

(ANN) y Máquinas de Soporte Vectorial (SVM), Para evaluar los modelos se utilizarán métricas de clasificación asociadas a la exactitud y al tiempo.

Los Kernel a evaluar serán seleccionados dentro de un conjunto de funciones trascendentes, que cumplan las condiciones Karush-Kuhn-Tucker (KKT).

La configuración de los hiperparámetros se realizará utilizando las técnicas de búsqueda aleatoria. El proyecto se desarrollará en un tiempo estimado de 18 meses, comenzando en el periodo A del año 2023 y finalizando en el periodo B de 2024.

## DESARROLLO

### *Tópicos del marco teórico*

#### Antecedentes

Se llevaron a cabo la consulta de fuentes bibliográficas de los últimos 5 años, relacionadas con predicción del riesgo cardiovascular o en relación con enfermedades cardiacas implementando la inteligencia artificial, donde se encontraron artículos a nivel internacional y nacional, estos a su vez tienen correlación con aspectos conceptuales y técnicas asociadas para predecir el riesgo y el tipo de datos. Otra revisión realizada fue consultar fuentes que cuente con los algoritmos de aprendizaje automático utilizando funciones kernel y máquinas de soporte vectorial, en las que se obtuvieron:

*Antecedentes Internacionales*

Según Yang et al.[11], en el proyecto titulado "Estudio del modelo de predicción de enfermedades cardiovasculares basado en un bosque aleatorio en el este de China", publicado en el año 2020, (seleccionaron, 29 930 sujetos con alto riesgo de enfermedades cardiovasculares (ECV)) de 10 1056 personas en 2014, se realizó un seguimiento regular utilizando un sistema de registro de salud electrónico. El análisis de regresión logística mostró que casi 30 indicadores estaban relacionados con la ECV, incluidos el género, la vejez, los ingresos familiares, el tabaquismo, el consumo de alcohol, la obesidad, la circunferencia de cintura excesiva, el colesterol anormal, la lipoproteína de baja densidad anormal, la glucemia en ayunas anormal y otros. Se utilizaron varios métodos para construir el modelo de predicción, incluido el modelo de regresión multivariante, el árbol de clasificación y regresión (CART), Naïve Bayes, Bagged trees, Ada Boost y Random Forest. Utilizaron el modelo de regresión multivariante como punto de referencia para la evaluación del desempeño (Área bajo la curva, AUC = 0,7143). Los resultados mostraron que Random Forest fue superior a otros métodos con un AUC de 0,787 y logró una mejora significativa con respecto al punto de referencia. Proporcionaron un modelo de predicción de ECV para la evaluación del riesgo de ECV a 3 años. Se basó en una gran población con alto riesgo de ECV en el este de China utilizando el algoritmo Random Forest, que proporcionaría una referencia para el trabajo de predicción y tratamiento de ECV en China. De acuerdo con esto, se necesitan más estudios poblacionales del modelo de predicción de ECV propuesto en esta investigación, con más población, mayor tiempo de seguimiento, que cubran más lugares en China con validación externa.[4,16] Este aporte y desarrollo del proyecto permite comparar las técnicas clásicas y como estas actúan en diversos parámetros para obtener resultados similares o diferentes a los ya estudiados, además de determinar las variables que más aportan o afectan a la variable respuesta, que para esta ocasión se obtienen los factores de riesgo significativos en predicción de enfermedades cardiovasculares, así mismo se logra la comparación de métodos usados para la evaluación de desempeño teniendo así ese resultado como referencia y determinar cuán significativa es la escogencia de las variables a trabajar para un determinado resultado.

Por otro lado, según Chávez Olivera et al.[17], en su estudio titulado "Aplicación Móvil para Predecir la Probabilidad de Pertenecer al Grupo de Riesgo Cardiovascular Utilizando Machine Learning" publicado en el año 2022, Lima, Perú, el presente estudio se centra en la creación de una aplicación móvil que tiene por principal funcionalidad la predicción de pertenecer al grupo de riesgo cardiovascular en personas mayores de 50 años. Para lograr esto se ha investigado sobre distintas variables y algoritmos de machine learning que permiten lograr esta tarea. Es así como se decidió que el motor de inferencia sería un modelo ensamblado, donde el metaclasificador final es un modelo de Naive Bayes y los modelos base son Random Forest y Logistic Regresion. El proceso de validación de la aplicación lo realizó un especialista en cardiología, el cual comprobó el nivel de precisión del modelo. Se observó que los modelos ensamblados de Support Vector Machine, Random Forest, Naive Bayes y Logistic Regresion obtienen precisiones de 87,00 %, 88,00 %, 87,00 % y 86,00 % con una estabilidad de 8,00 %, 6,00 %, 2,00 % y 8,00 % respectivamente, pese a que el modelo ensamblado Random Forest tiene una mejor precisión, este es más inestable, por lo que se escogió el modelo ensamblado Naive Bayes, ya que tiene una precisión parecida y es más estable que los demás. El aporte que se provee de esta investigación es que muestra los resultados de los entrenamientos de distintos modelos de machine learning, y la combinación de estos, el ensamblaje permite la construcción de algoritmos más precisos y estables, además que al usar modelos Random Forest y Naive Bayes se puede determinar la mejor precisión, y tener un punto de comparación con la investigación.

Según Scavino et al.[18], en su informe titulado "Informe final publicable de proyecto Creación de algoritmos utilizando técnicas de clasificación supervisada y no supervisada para el diagnóstico de enfermedades cardiovasculares en una población de adultos mayores de bajos recursos en Uruguay" publicado en el año 2022, Uruguay. La presente investigación tiene como propósito la generación de algoritmos de aprendizaje automático para la identificación de la patología cardíaca, fibrilación auricular, a partir de datos de la señal electrocardiográfica de una sola derivación con un dispositivo móvil de tecnología electrónica. Los algoritmos de aprendizaje profundo con las arquitecturas consideradas no mostraron un buen desempeño. Una mayor cantidad de datos podría resultar en una mejora de la capacidad de clasificación de estos algoritmos. Por otra parte, las técnicas de aprendizaje estadístico aplicadas a un conjunto de características extraídas de la señal ECG sin procesar mostraron un mejor desempeño. Es conveniente destacar que la construcción de estos algoritmos permite entender el funcionamiento de estos y cómo se llega al diagnóstico final, la capacidad de poder interpretar los mecanismos internos de los métodos para ofrecer un resultado genera una fuente de conocimiento amplia con posibilidad de desarrollo a futuro, además de encontrar posibles causas por las que el uso de diferentes algoritmos llega a la variación en los diagnósticos.

En un cuarto estudio, según Polero et al.[19], en su proyecto titulado "Predicción de riesgo de sufrir un síndrome coronario agudo mediante un algoritmo de Machine Learning (ANGINA)" publicado en el año 2020, Buenos Aires, Argentina. La presente investigación pretende demostrar la capacidad de los clasificadores de machine learning para diagnosticar y predecir un SCA en pacientes que consultan de forma espontánea al SEM con dolor torácico de etiología no identificada, durante un período de seguimiento de 30 días. Se analizaron 161 pacientes que consultaron al SEM con dolor torácico. Se registró mediante un clasificador de machine learning las variables objetivas y subjetivas de caracterización del dolor. De esta manera se obtuvo que la edad promedio fue de 57 más/menos 12, 72,7 % masculinos eran de sexo masculino y 17,4 % presentaban evento coronario previo. El 57,8 % presentaba un síndrome coronario agudo con una incidencia de IAM de 29,8 %, de los cuales requirieron revascularización por ATC el 35 %, y CRM el 9,9 % en el período de seguimiento a 30 días. Como modelo de clasificación se utilizó un Random Forest Classifier que presentó un área bajo la curva ROC de 0,8991, sensibilidad de 0,8552, especificidad de 0,8588 y una precisión de 0,8441. Las variables predictoras más influyentes fueron peso (p = 0,002), edad (p = 5,011e-07), intensidad del dolor (p = 3,0679e-05), tensión arterial sistólica (p = 0,6068) y características subjetivas del dolor (p = 1,590e-04). Este proyecto permite conocer algunas métricas de evaluación de modelos que proporcionan medidas cuantitativas mostrando su rendimiento y las variables que pueden ser consideradas como importantes para el desarrollo del mismo manifestando a su vez el comportamiento de las variables para la problemática a resolver.

*Antecedentes Nacionales*

Según Peres[20], en su estudio titulado "Optimización De Un Modelo De Clasificación De Enfermedades Cardiovasculares Utilizando Técnicas De Aprendizaje Profundo Supervisado Y Despliegue De Dashboard Web" publicado en el año 2021, Cartagena. El desarrollo de este proyecto se basa en la optimización de un modelo previo de predicción de hipertensión, del cual se mejora el proceso de análisis de los datos, ya que el modelo inicial no generaba la predicción exacta de clasificación de enfermedades cardiovasculares debido a la mala interpretación y la falta de limpieza de los datos. Así mismo, la optimización consistió en hacer ajustes en el modelo, realizando mejoras en la construcción de la red neuronal y en el proceso de entrenamiento, identificando técnicas de activación y épocas adecuadas para lograr obtener resultados óptimos. Finalmente, la preparación de los datos permitió formular, entrenar y probar un modelo de clasificación de enfermedades cardiovasculares construido en el entorno de TensorFlow desarrollado por Google el cual se obtuvo una red neuronal recurrente compuesta con una capa de entrada con 18 nodos, dos capas ocultas con 128 nodos cada capa y una capa de salida de 3 nodos en los cuales fueron los más óptimos para la predicción la cual obtuvo una precisión de 97 % en la validación del modelo, y posteriormente se desplegó haciendo uso de una aplicación web para la consulta del personal médico el modelo elaborado por los mencionados anteriormente arrojó como resultados, una precisión superior al 86 %, viendo los porcentajes en las métricas de evaluación que se utilizaron como: Precisión, Recall, F1 Score y accuracy, las cuales arrojaron resultados que superan el 80 % en la clasificación de los riesgos. Esta investigación es valiosa para el desarrollo del proyecto porque al comparar las métricas de evaluación, presentan resultados considerables, que se pueden abordar con la calidad de los datos que se obtengan, así mismo se hace un recorrido sobre el entorno TensorFlow que permite la visualización de datos en distintas áreas del conocimiento.

Como segunda investigación, según Martínez[21], en su investigación titulada "Predicción De Enfermedades Cardiovasculares Mediante Algoritmos De Inteligencia Artificial" publicado en el año 2020, Málaga. Este proyecto se centra en implementar 5 algoritmos distintos de clasificación, analizar cómo se ajustan a los datos disponibles y, posteriormente, crear un algoritmo genético que detecte la combinación de parámetros de cada algoritmo, con la que se obtienen mejores resultados, medidos a través de la exactitud. Los algoritmos fueron implementados con la librería sci-kit learn de Python; para el algoritmo genético no se utilizó ninguna librería

adicional. Los resultados mostraron que algunos algoritmos se adaptan mejor a la evolución, es decir, la exactitud aumenta con el paso de las generaciones. Otros algoritmos mostraron una disminución de este valor, sugiriendo que se necesita estudiar para cada tipo de algoritmo el impacto de cada parámetro, además de los valores que en este proyecto se consideraron constantes: número de generaciones, número de individuos por generación, probabilidad de mutación y cruce y el tamaño del conjunto de datos y de los subconjuntos de entrenamiento, validación y pruebas. Esta investigación es considerable para el proyecto debido a la comparación que hace de algoritmos de clasificación y como estos obtienen mejores resultados a través de la exactitud, aportando a uno de los objetivos específicos con el fin de tener un marco teórico amplio sobre la investigación.

Como tercera investigación, según Florez[22], en su investigación titulado "Modelo de inteligencia artificial como apoyo diagnóstico para la estimación de riesgo cardiovascular en pacientes atendidos bajo la modalidad de telemedicina en una IPS del departamento de Sucre 2021" publicado en el año 2021, Sucre. El presente proyecto busca mediante un estudio experimental la incorporación de Inteligencia Artificial como apoyo diagnóstico en el proceso de atención a pacientes con riesgo cardiovascular. Para lo cual se aprovecharán los datos obtenidos de pacientes atendidos durante el 2018 y 2019 por una IPS del departamento de Sucre. Se presenta la problemática del aumento de mortalidades en personas con enfermedades cardiovasculares que se pueden prevenir si los diagnósticos son oportunos, para ello el proyecto busca mediante un estudio experimental la incorporación de Inteligencia Artificial como apoyo diagnóstico en el proceso de atención de pacientes con riesgo cardiovascular. Entre las variables a monitorear se encuentran edad, sexo, peso, talla, Índice de masa corporal, presión arterial, presencia de diabetes, dislipidemias, entre otras. A su vez, bajo el concepto de Machine Learning, se desarrollará un algoritmo inteligente con la habilidad de aprender sin ser explícitamente programado que permita evaluar los riesgos potenciales del individuo y así, asistir virtualmente al personal médico en las acciones de promoción, prevención y diagnóstico de manera minuciosa y precisa para la modalidad de atención en telemedicina. Esperando que, a partir de la caracterización de las variables, definir los factores de riesgo de mayor impacto en la incidencia del riesgo cardiovascular en la población de estudio. Con la selección del modelo predictivo que más se ajusta a las características de la población y los datos de calidad suministrados, se podrá definir una clasificación de riesgo cardiovascular que podrá ser adaptado al servicio de tele monitoreo en prevención primaria y secundaria. Con el uso de estos algoritmos de manera particular, permite ver la predicción y estos algoritmos desde otra perspectiva, ya que su intención es que el algoritmo aprende sin ser explícitamente programado.

El estudio de este proyecto permite confirmar y corroborar las funciones kernel utilizadas y las que mejor se acondicionen a los datos, dependiente el uso que se le brinde. Además de permitir realizar mejoras en algunos aspectos del desarrollo, para diferentes enfoques.[11]

**Enunciado de supuestos teóricos de la investigación**

En este capítulo se toma en cuenta las bases teóricas del presente trabajo, de tipo conceptual.

**Enfermedades Cardiovasculares**

Son enfermedades del sistema circulatorio, de etiología y localización diversas. Se clasifican en cuatro tipos generales: enfermedades isquémicas del corazón, enfermedades cerebrovasculares, en enfermedades vasculares periféricas y otras enfermedades.[23]

Importantes avances en el tratamiento de las ECV han sido facilitados por la identificación de los factores de riesgo tradicionales, pero a pesar de la evidencia clínica acumulada, la implementación de estrategias para prevenir las enfermedades cardiovasculares aún permanecen lejos de ser óptimas.[24]

**El riesgo cardiovascular**

El riesgo cardiovascular se define como la probabilidad de padecer un evento cardiovascular en un determinado período, que habitualmente se establece en 5 o 10 años, especialmente en los pacientes que no padecen enfermedad cardiovascular, es decir, en prevención primaria, es fundamental para establecer la intensidad de la intervención, la necesidad de instaurar tratamiento farmacológico y la periodicidad de las visitas de seguimiento.[24]

**Factores de riesgo cardiovasculares**

Los factores de riesgo son aquellos signos biológicos o hábitos adquiridos que se presentan con mayor frecuencia en los pacientes con una enfermedad concreta. La enfermedad cardiovascular tiene un origen multifactorial, y un factor de riesgo debe ser considerado en el contexto de los otros. Los factores de riesgo cardiovascular, clásicos o tradicionales, se dividen en 2 grandes grupos: no modificables (edad, sexo y antecedentes familiares), y modificables (dislipidemia, tabaquismo, diabetes, hipertensión arterial, obesidad y sedentarismo).

Aunque el impacto de factores de riesgo individuales como la hipertensión arterial, la dislipidemia, el hábito

de fumar y la diabetes, entre otros, está bien establecido y mejora la predicción del riesgo cardiovascular.[24]

Cuanto mayor sea el nivel de cada factor de riesgo, mayor es el riesgo de tener una enfermedad cardiovascular ateroesclerosis como la cardiopatía coronaria.

- Obesidad: la relación entre el peso y las enfermedades del corazón no viene tan sorpresa, ya que una persona obesa tendrá más grasa y, por lo tanto, más posibilidades de sufrir ECV. "Independientemente de la salud metabólica, las personas con sobrepeso y obesas tienen mayor riesgo de enfermedad coronaria que las personas delgadas".[25]
- Actividad física: el ejercicio constante y un estilo de vida saludable pueden, en general, impactar muy positivamente al tratamiento de enfermedades: prevenir o retrasar la aparición el tipo 2 diabetes, reducir la presión arterial y ayudar a reducir el riesgo de ataque cardíaco y accidente cerebrovascular.[25] Cuanto más vigorosa la actividad, mayor el beneficio. Sin embargo, aun las actividades de intensidad moderada ayudan si se realizan de forma habitual y a largo plazo. El ejercicio puede ayudar a controlar el colesterol, la diabetes y la obesidad, así como a reducir la presión arterial en algunas personas. La actividad física debería ser una actividad diaria. Caminar entre 30 a 40 minutos, la mayor cantidad de días por semana posibles, pero no menos de 3 días es un buen ejercicio y tiene pocas contraindicaciones.[26]
- Niveles de colesterol: la acumulación de colesterol es una de las principales causas de la aterosclerosis. Se ha demostrado consistentemente que niveles más altos de colesterol LDL a largo plazo y las concentraciones de colesterol de lipoproteínas que no son de alta densidad están asociadas con un mayor riesgo de ECV.[25]
- Glucosa / Diabetes: la diabetes no es solo una alteración de los niveles de azúcar en sangre, pero afecta al sistema en general. Los estudios reportan una asociación positiva entre hipertensión y resistencia a la insulina.[25]
- Tabaquismo: hay evidencia de que fumar causa aproximadamente 1 de cada 10 muertes por enfermedades cardiovasculares. El humo del tabaco contribuye a las enfermedades cardiovasculares, ya que aumenta la placa aterosclerótica y la posibilidad de trombosis.[25] El humo del tabaco es el principal factor de riesgo para la muerte súbita de origen cardiaco y los fumadores tienen de dos a cuatro veces más riesgo que los no fumadores. El riesgo cardiovascular disminuye rápidamente al dejar de fumar.[26]
- Antecedentes Familiares: los hijos/as de padres con cardiopatía isquémica, especialmente si esta ha sido prematura (padres antes de los 65 años, madres antes de los 55 años) o con hipertensión arterial, tienen mayor probabilidad de desarrollarla. Existen formas minoritarias de colesterol muy elevado (por encima de los 350 mg/dl) llamadas hipercolesterolemia familiar, que son debidas a trastornos hereditarios y que conllevan un riesgo muy elevado, incluso antes de la menopausia. En estos casos son precisos tratamientos médicos agresivos con hipolipemiantes.[26]

Existen 2 métodos de cálculo del riesgo cardiovascular: cualitativos y cuantitativos. Los cualitativos se basan en la suma de factores de riesgo o la medición de su nivel y clasifican al individuo en: riesgo leve, moderado, alto y muy alto; los cuantitativos, por su parte, están basados en ecuaciones de predicción de riesgo que nos dan un número que es la probabilidad de presentar un evento cardiovascular en un determinado tiempo, y la forma de cálculo es a través de programas informáticos o de las llamadas tablas de riesgo cardiovascular, que son unas herramientas de enorme utilidad para la toma de decisiones en la práctica clínica habitual.[24]

- Algoritmo de estudio Framingham: la probabilidad de que ocurra una enfermedad cardiovascular para una variable determinada, así encontramos:
- Hombres y mujeres tienen una probabilidad distinta de llegar a padecer una enfermedad cardiovascular.
- La edad es otro determinante; a mayor edad mayor el riesgo cardiovascular.
- El hábito en el consumo de tabaco es una variable que aumenta el riesgo cardiovascular independientemente de las otras variables.
- Los niveles de colesterol, HDL y LDL son todas variables que aumentan o disminuyen (HDL) el riesgo cardiovascular, independientemente de las demás.
- Los niveles de presión arterial altos y si se tiene o no tratamiento farmacológico para la hipertensión.

Las anteriores son las variables clásicas que analiza el estudio, pero además otros investigadores buscan sumar algunas nuevas:

- Ascendencia, estado civil y educación.
- Tipo de trabajo, ritmo y horario del trabajo, apoyo de los compañeros o del supervisor.
- Perímetro abdominal en personas con diabetes o síndrome metabólico.[27]

La Inteligencia Artificial (IA) podría definirse como la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano, bien sea: pensar, sentir, resolver

problemas, tomar decisiones e inclusive aprender, IA comprende el área del Machine Learning, Deep Learning, Big Data y ciencia de datos.

- Scikit Learn: scikit-learn es una biblioteca de aprendizaje automático gratuita para Python. Cuenta con varios algoritmos como máquina de vectores de soporte, bosques aleatorios y vecinos k, y también admite bibliotecas numéricas y científicas de Python como NumPy y SciPy.[28]

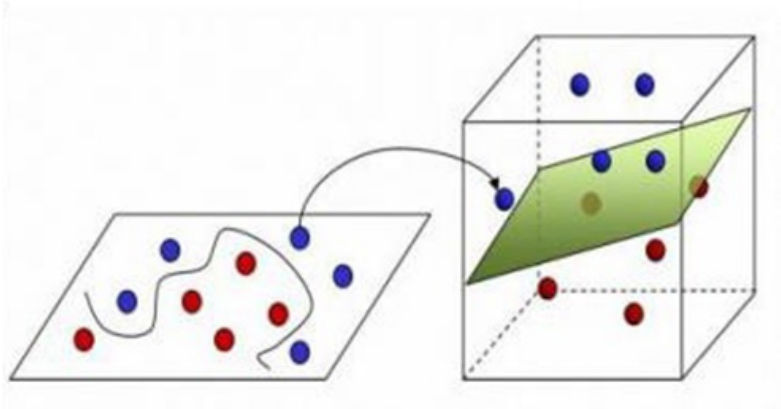
**Máquinas de soporte vectorial, Algoritmo de soporte vectorial (SVM)**

Es un algoritmo de aprendizaje supervisado que se utiliza en muchos problemas de clasificación y regresión, incluidas aplicaciones médicas de procesamiento de señales, procesamiento del lenguaje natural y reconocimiento de imágenes y voz.

El objetivo del algoritmo SVM es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos. "De la mejor forma posible" implica el hiperplano con el margen más amplio entre las dos clases, representado por los signos más y menos en la siguiente figura. El margen se define como la anchura máxima de la región paralela al hiperplano que no tiene puntos de datos interiores. El algoritmo solo puede encontrar este hiperplano en problemas que permiten separación lineal; en la mayoría de los problemas prácticos, el algoritmo maximiza el margen flexible, permitiendo un pequeño número de clasificaciones erróneas.[29,30,31]

**Funciones kernel**

Las Funciones Kernel son funciones matemáticas que se emplean en las Máquinas de Soporte Vectorial. Estas funciones son las que le permiten convertir lo que sería un problema de clasificación no lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio dimensional mayor a este espacio M-dimensional se le conoce como espacio de Hilbert.[32,33,34]



**Fuente:** https://www.researchgate.net/figure/260283043_fig13_Figure-A15-The-non-linear-SVM-classifier-with- the-kernel-trick

**Figura 1.** Truco kernel de dos dimensiones a tres

En la figura 1, se puede observar cómo trabajan las funciones kernel, llevando una distribución de datos de dos dimensiones a tres dimensiones, a este funcionamiento generalmente se lo llama truco kernel. Este truco permite reducir la complejidad de una función que separe las clases de una distribución de datos como lo demuestra Peluffo-Ordóñez, la figura 1 por ejemplo en dos dimensiones se podría separar mediante funciones no lineales o segmentadas (Elipse, Círculo, Rectángulo), mientras que en tres dimensiones se podría separar mediante una función lineal (Hiperplano) como lo demuestra Baudat & Anouar en el documento.[11]

Para que las funciones kernel puedan ser consideradas candidatas a kernels, deben cumplir tres condiciones iniciales fundamentales; deben ser:

- Continuas.
- Simétricas.
- Positivas.

Estos son los requerimientos básicos para poder ser expresadas como un producto escalar en un espacio dimensional alto.[30,35,36]

Existen diversos kernel empleados comúnmente en bibliotecas de aprendizaje automático como lineal, RBF, polinomial y tangente hiperbólico cuyas definiciones se pueden expresar en la figura 2.[11]

| Función kernel | Ecuación | Condición |
|---|---|---|
| Lineal | $k(x,x') = \langle x,x' \rangle$ | $x,x \in \mathbb{R}$ |
| RBF | $k(x,x') = \exp\ (-\sum_{i=1}^{d} \lambda_i (x_i x_i')$ | $\lambda_i > 0, \beta \in (0,2]$ |
| Polinomial | $k(x,x') = (\alpha\langle x,x' \rangle + 1)^m$ | $m \in \mathbb{N}, \alpha > 0$ |
| Tangente hiperbólica | $k(x,x') = tanh(\alpha\langle x,x' \rangle + b)$ | $a > 0, b < 0$ |

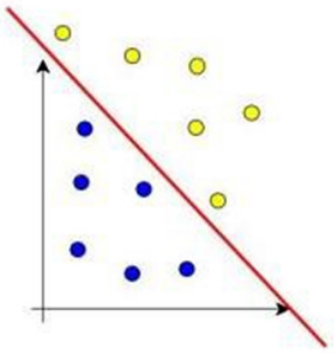
**Fuente:** comparativo de funciones kernel sobre predicción de oferta de fuentes alternativas de energía
**Figura 2.** Definición formal funciones kernel

Aunque las funciones de la figura 2 son funciones de uso común, existen muchas más funciones kernel. Estas funciones son utilizadas en algoritmos supervisados para regresión y clasificación; y no supervisadas para detección de anomalías, análisis, clúster y extracción de características. Dentro de los algoritmos más destacados se encuentran procesos gausianos, Spectral clustering, Kernel Linear Discriminant Analysis, Kernel Principial Components Analysis, Kernel Canonical Correlation Analysis, Kernel Independent Component Analysis, SVM, ANN y muchos más.[11]

**Identificación de patrones mediante modelos lineales**

La detección de patrones es el proceso de encontrar en un conjunto de datos caóticos modelos capaces de generalizar el comportamiento de los datos para la obtención de clasificaciones, predicciones o detección de anomalías. Para la obtención de estos patrones se recurre a la sinergia del conocimiento provista por varias ciencias como las matemáticas, estadística, probabilidad, computación, entre otras. Actualmente, estas técnicas están englobadas en una rama denominada aprendizaje automático, divididas en técnicas de aprendizaje supervisado, no supervisado y por refuerzo.[11]

- Clasificadores lineales: un clasificador lineal es aquel capaz de encontrar la clase (y) discreta a la que pertenece un conjunto de datos basada en una combinación lineal de sus atributos (x) como se muestra en la figura 1.[11]



**Fuente:** https://docplayer.es/192405564-Comparativo-de-kernels-sobre-prediccion-de-oferta-de-fuentes- alternativas-de-energia.html.
**Figure 3.** Ejemplo de clasificador lineal

Como se observa en la figura 3 el clasificador lineal ha trazado un separador, en este caso una línea que permite deducir a que clase pertenece (puntos amarillos o azules) así un nuevo registro, si se posiciona del lado superior de la línea sería clasificado como punto amarillo, de lo contrario como azul.[11]

Si la entrada del clasificador es un vector de características reales $\vec{x}$, entonces el resultado de salida es

$$y = f(\vec{w}.\vec{x}) = \left( f \sum_j w_j x_j \right),$$

**Fuente:** https://docplayer.es/192405564-Comparativo-de-kernels-sobre-prediccion-de-oferta-de-fuentes- alternativas-de-energia.html
**Figura 4.** Clasificadores Lineales

Donde $\vec{w}$ es un vector real de pesos y f es una función que convierte el producto punto a punto de los dos vectores en la salida deseada. El vector de pesos $\vec{w}$ aprende de un conjunto de muestras de entrenamiento. A menudo f es una función simple que mapea todos los valores por encima de un cierto umbral a la primera clase y el resto a la segunda clase.[31,37,38,39]

Algunos de los algoritmos de clasificación lineal más utilizados para encontrar estos patrones de clasificación se encuentran: análisis de discriminante lineal, clasificados de Bayes lineal, regresión

- Aprendizaje supervisado: es una rama de Machine Learning, un método de análisis de datos que utiliza algoritmos que aprenden iterativamente de los datos para permitir que los ordenadores encuentren información escondida sin tener que programar de manera explícita dónde buscar. El aprendizaje supervisado es uno de los tres métodos de la forma en que las máquinas "aprenden": supervisado, no supervisado y optimización. El aprendizaje supervisado resuelve problemas conocidos y utiliza un conjunto de datos etiquetados para entrenar un algoritmo para realizar tareas específicas.[32,40]
- Aprendizaje no supervisado: es un tipo de Machine Learning que se utiliza para identificar nuevos patrones y detectar anomalías. Los datos que se introducen en los algoritmos de aprendizaje no supervisados no están etiquetados. El algoritmo (o modelos) intentan dar sentido a los datos por sí mismos mediante la búsqueda de características y patrones.[32,41,42] Es importante destacar que para este proyecto se centrara específicamente en las técnicas de aprendizaje supervisado para clasificación y predicción mediante modelos lineales.
- Clasificadores lineales: Un clasificador lineal es aquel capaz de encontrar la clase (y) discreta a la que pertenece un conjunto de datos basados en una combinación lineal de sus atributos (x).[11]

**Base de datos**

Se denomina base de datos a un conjunto de información perteneciente a un mismo contexto, ordenada de modo sistemático para su posterior recuperación, análisis y transmisión. Hoy en día, las bases de datos se presentan de diferentes formas y tamaños, esto de acuerdo al lugar donde son empleados, por ejemplo, en una biblioteca o en cuentas de una empresa. Las bases de datos se originaron para cubrir la necesidad de almacenar grandes cantidades de información, es decir, de preservarla contra el tiempo y el deterioro, para acudir a ella posteriormente. Es ese sentido, la aparición de la electrónica y la computación brindaron el elemento digital indispensable para almacenar enormes cantidades de datos en espacios físicos limitados, gracias a su conversión en señales eléctricas o magnéticas.[33,43,44]

**Métricas de medición**

| MÉTRICAS DE REGRESIÓN EN APRENDIZAJE AUTOMÁTICO | | |
|---|---|---|
| **NOMBRE** | **DEFINICIÓN** | **FÓRMULA** |
| **(MSE) Error cuadrático medio** | Es una métrica utilizada en estadística y aprendizaje automático para evaluar el rendimiento de un modelo de predicción. Se utiliza para medir la diferencia entre los valores predichos por el modelo y los valores reales observados.<br>El MSE se calcula tomando la diferencia entre cada valor predicho y el valor real correspondiente, elevando al cuadrado esta diferencia y luego calculando el promedio de todos estos errores cuadráticos. En otras palabras, se obtiene la media de los errores cuadrados.<br>Cuanto menor sea el valor del MSE, mejor será el modelo, ya que indica que las predicciones se acercan más a los valores reales. Un MSE de cero indicaría un modelo perfecto en el que las predicciones coinciden exactamente con los valores observados [34]. | $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$<br>**Error Cuadrático Medio**<br>Donde n es el número de observaciones en el conjunto de datos, $y_i$ es el valor real observado y $\hat{y}_i$ es el valor predicho por el modelo. |

| | | |
|---|---|---|
| **(RMSE) Raíz del error cuadrático medio** | Es una medida estadística comúnmente utilizada para evaluar la precisión de un modelo de regresión. Es una métrica que se deriva del Error Cuadrático Medio (MSE) y se calcula tomando la raíz cuadrada del MSE [34]. | $$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{\mathrm{MSE}}$$ **RMSE** Donde MSE es el Error Cuadrático Medio. Al igual que el MSE, el RMSE mide la diferencia promedio entre los valores predichos por el modelo y los valores reales observados. La principal diferencia es que el RMSE tiene la misma unidad de medida que la variable objetivo, lo que lo hace más interpretable y fácil de comparar con los valores reales. El RMSE se utiliza ampliamente en problemas de regresión para evaluar qué tan cerca están las predicciones del modelo de los valores reales. Cuanto menor sea el valor del RMSE, mejor será la capacidad predictiva del modelo. Por ejemplo, si se tiene dos conjuntos de predicciones, A y B, y se dice que el MSE de A es mayor que el MSE de B, entonces se puede estar seguro de que RMSE de A es mayor que RMSE de B. Y también funciona en la dirección opuesta. |
| **(MAE) -Error absoluto medio** | Se calcula como la media de las diferencias absolutas entre los valores predichos por el modelo y los valores reales observados. El MAE proporciona una medida de la magnitud promedio de los errores de predicción sin tener en cuenta su dirección. Es una métrica comúnmente utilizada debido a su simplicidad y facilidad de interpretación. El MAE se expresa en las mismas unidades que la variable objetivo y cuanto menor sea su valor, mayor será la precisión del modelo en términos de predicción. Es importante tener en cuenta que el MAE no considera la magnitud relativa de los errores, por lo que todos los errores se tratan por igual en la métrica. Esto puede ser adecuado en ciertos escenarios donde se desea penalizar de manera uniforme todos los errores de predicción [34]. | $$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$$ **Error Absoluto Medio** Donde: • MAE: Mean Absolute Error • N: Número de observaciones en el conjunto de datos. • $\sum$: Sumatoria. • $y_i$: Valor real observado. • $\hat{y}_i$: Valor predicho por el modelo. |
| **($R^2$) – R al cuadrado** | El coeficiente de determinación, o $R^2$ (a veces leído como R-dos), es una medida estadística utilizada en el análisis de regresión para evaluar qué tan bien se ajusta un modelo a los datos observados. El R-cuadrado es un valor que varía entre 0 y 1, y se interpreta como el porcentaje de la variabilidad de la variable dependiente que es explicada por el modelo. | $$R^2 = 1 - \frac{\mathrm{MSE(model)}}{\mathrm{MSE(baseline)}}$$ **R Cuadrado** |

| | | |
|---|---|---|
| | un valor de $R^2$ cercano a 1 indica que el modelo explica una gran proporción de la variabilidad de la variable dependiente y se ajusta bien a los datos. Por otro lado, un valor de $R^2$ cercano a 0 indica que el modelo no explica de manera adecuada la variabilidad de la variable dependiente y no se ajusta bien a los datos [34]. | |
| **R cuadrado ajustado ($R^2$)** | Es una medida estadística relacionada con el coeficiente de determinación (R-cuadrado) que tiene en cuenta la cantidad de variables predictoras en un modelo de regresión y ajusta el R-cuadrado en función del número de | $$R^2_{adj} = 1 - [\frac{(1 - R^2)(n - 1)}{n - k - 1}]$$ **R Cuadrado Ajustado** Donde: • $R^2$ ajustado: Coeficiente de determinación |

| | | |
|---|---|---|
| | predictores utilizados. El R-cuadrado ajustado se utiliza para evaluar y comparar modelos de regresión que tienen diferentes números de variables predictoras.<br>El R-cuadrado ajustado penaliza el uso de variables predictoras adicionales que no aportan información significativa al modelo. A medida que se agregan más variables predictoras al modelo, el R-cuadrado ajustado disminuirá si esas variables no mejoran de manera sustancial la capacidad de explicación del modelo. Por lo tanto, el R-cuadrado ajustado tiende a ser más conservador que el R-cuadrado estándar y proporciona una medida más realista del ajuste del modelo [34]. | ajustado.<br>• $R^2$: Coeficiente de determinación (R-cuadrado).<br>• n: es el número total de observaciones<br>• k: es el número de regresores independientes, es decir, el número de variables en su modelo, excluyendo la constante. |
| **(MSPCE) – Error porcentual cuadrático medio** | Es una medida de error utilizada para evaluar la precisión de un modelo de predicción en relación con los valores reales. Es una métrica que combina la magnitud de los errores y la proporción relativa de los errores en relación con los valores reales.<br>El MSPE se expresa como un porcentaje, lo que facilita la interpretación de la magnitud del error en relación con los valores reales. Un valor más bajo de MSPE indica una mejor precisión del modelo, mientras que un valor más alto indica una mayor discrepancia entre las predicciones y los valores reales [34]. | $$MSPE = \frac{100\%}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{y_i})^2$$ **Error de Porcentaje Cuadrático Medio (MSPCE)** Donde: • n es el número total de muestras o datos. • $y_i$: son los valores reales o verdaderos. • $\hat{y}_i$: son los valores predichos por el modelo. |
| **(MAPE) – Error porcentual** | Es una medida de error comúnmente utilizada para evaluar la precisión de un | |

| | | |
|---|---|---|
| **absoluto medio** | modelo de predicción en relación con los valores reales. El MAPE calcula el promedio del porcentaje absoluto de error entre las predicciones y los valores reales.<br>El MAPE se expresa como un porcentaje, lo que facilita la interpretación de la magnitud del error en relación con los valores reales. Un valor más bajo de MAPE indica una mejor precisión del modelo, mientras que un valor más alto indica una mayor discrepancia entre las predicciones y los valores reales [34]. | $$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$$<br>**Error Porcentual Absoluto Medio (MAPE)**<br>Donde:<br>• n es el número total de muestras o datos.<br>• $y_i$: son los valores reales o verdaderos.<br>• $\hat{y}_i$: son los valores predichos por el modelo. |
| **(RMSLE) – Error logarítmico cuadrático medio** | Es solo un RMSE calculado en escala logarítmica. Es una métrica útil cuando los valores tienen una amplia gama y hay una gran variabilidad en los datos.<br>El RMSLE toma el logaritmo de los valores reales y los valores predichos antes de calcular el error cuadrático medio. Esto es útil cuando los valores objetivo abarcan un rango amplio y se desea penalizar de manera más equitativa los errores en diferentes magnitudes.<br>El RMSLE se calcula como la raíz cuadrada del promedio de los errores cuadráticos de los logaritmos. Al tomar la raíz cuadrada, se obtiene una medida de error en la misma escala que los valores originales [34]. | $$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2} = \text{RMSE}(\log(y_i + 1), \log(\hat{y}_i + 1)) = \sqrt{\text{MSE}(\log(y_i + 1), \log(\hat{y}_i + 1))} =$$<br>**Error Logarítmico Cuadrático Medio (RMSLE)**<br>Por lo tanto, esta métrica se usa generalmente en la misma situación que MSPCE y MAPE, ya que también conlleva errores relativos más que errores absolutos.<br>RMSLE penaliza una estimación poco predicha mayor que una estimación sobre pronosticada.<br>RMSLE se puede calcular sin la operación raíz, pero la versión rooteada se usa más ampliamente. |

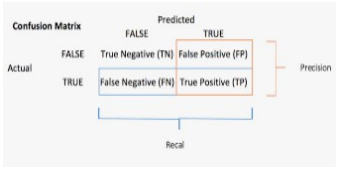
**Figura 5.** Métricas de regresión en aprendizaje automático

## Métricas de clasificación

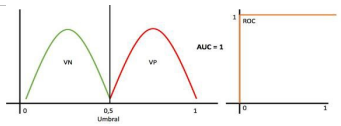
| MÉTRICAS DE EVALUACIÓN | | |
|---|---|---|
| **NOMBRE** | **DEFINICIÓN** | **FÓRMULA** |
| **Precision (Precisión)** | Es una medida de evaluación utilizada en problemas de clasificación para medir la exactitud de un modelo al predecir correctamente las instancias positivas. Representa la proporción de predicciones positivas que son verdaderamente positivas en comparación con todas las predicciones positivas realizadas por el modelo.<br>La precisión se expresa como un valor entre 0 y 1, donde 1 representa una precisión perfecta y 0 indica una precisión nula. | $$\text{precision} = \frac{TP}{TP + FP}$$<br>**Fórmula de Precisión (Presicion)**<br>Donde:<br>• Verdaderos positivos (TP) son los casos positivos que fueron correctamente identificados por el modelo.<br>• Falsos positivos (FP) son los casos negativos que fueron incorrectamente clasificados como positivos por el modelo. |

| | | |
|---|---|---|
| | La precisión es una métrica útil cuando el objetivo principal es minimizar los falsos positivos, es decir, evitar clasificar incorrectamente instancias negativas como positivas. Es especialmente importante en casos donde los falsos positivos tienen un impacto significativo o costoso.<br>Es importante tener en cuenta que la precisión no tiene en cuenta los casos negativos que fueron correctamente clasificados como negativos (verdaderos negativos) [35]. | |

| | | |
|---|---|---|
| **RECALL (Exhaustividad)** | El recall, también conocido como sensibilidad o tasa de verdaderos positivos, es una medida de evaluación utilizada en problemas de clasificación para medir la capacidad de un modelo de identificar correctamente las instancias positivas.<br>El recall se expresa como un valor entre 0 y 1, donde 1 representa un recall perfecto y 0 indica un recall nulo.<br>El recall es una métrica importante cuando el objetivo principal es minimizar los falsos negativos, es decir, evitar clasificar incorrectamente instancias positivas como negativas. Es especialmente relevante en casos donde los falsos negativos tienen un impacto significativo o costoso [35]. | $$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$<br>**Exhaustividad (recall)**<br>Donde:<br>• Verdaderos positivos (TP): son los casos positivos que fueron correctamente identificados por el modelo.<br>• Falsos negativos (FN): son los casos positivos que fueron incorrectamente clasificados como negativos por el modelo. |
| **F1-SCORE (Valor-F)** | El F1-score también conocido como puntuación F1, es una medida de evaluación utilizada en problemas de clasificación que combina la precisión y el recall en una sola métrica. Representa el equilibrio entre la precisión y el recall de un modelo.<br>El F1-score es una medida que varía entre 0 y 1, donde 1 representa un F1-score perfecto y 0 indica un rendimiento nulo. Un F1-score más alto indica un mejor equilibrio entre la precisión y el | $$\text{F1} = 2.\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$<br>**F1-Score**<br>• Precision es la proporción de verdaderos positivos (TP) sobre el total de instancias clasificadas como positivas por el modelo, es decir, la capacidad del modelo para evitar falsos positivos.<br>• Recall es la proporción de verdaderos positivos (TP) sobre el total de instancias positivas reales, es decir, la capacidad del modelo para evitar falsos negativos. |

| | | |
|---|---|---|
| | recall.<br>El F1-score es especialmente útil cuando hay un desequilibrio entre las clases en los datos de entrenamiento [35]. | |
| **Accuracy (Exactitud)** | La exactitud (accuracy) representa la proporción de instancias clasificadas correctamente sobre el total de instancias en el conjunto de datos [35].<br>La exactitud se calcula dividiendo el número de predicciones correctas (verdaderos positivos y verdaderos negativos) entre el número total de instancias en el conjunto de datos.<br>La exactitud es la cantidad de aciertos del modelo sobre el total de instancias. Una exactitud de 1 indica que el modelo clasifica todas las instancias correctamente, mientras que una exactitud de 0 indica que el modelo no acierta ninguna instancia [35]. | $\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$<br>**Fórmula Accuracy** |
| **CONFUSION MATRIX (Matriz de confusión)** | Confusión o error Matrix es una tabla que describe el rendimiento de un modelo supervisado de Machine Learning en los datos de prueba, donde se desconocen los verdaderos valores. Se llama "matriz de confusión" porque hace que sea fácil detectar dónde el sistema está confundiendo dos clases.<br>• True Positives (TP): cuando la clase real del punto de datos |  <br>Fig. 4. Matriz de Confusión |

| | | |
|---|---|---|
| | era 1 (Verdadero) y la predicha es también 1 (Verdadero).<br>• Verdaderos Negativos (TN): cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).<br>• False Positives (FP): cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True).<br>• False Negatives (FN): Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso) [34]. | |

| | | |
|---|---|---|
| **ESPECIFICIDAD O TNR (Tasa negativa real)** | Es el número de ítems correctamente identificados como negativos fuera del total de negativos.<br>La especificidad se calcula dividiendo el número de verdaderos negativos (TN) entre la suma de los verdaderos negativos (TN) y los falsos positivos (FP).<br>La especificidad proporciona información sobre la capacidad del modelo para evitar clasificar incorrectamente los casos negativos. Un valor alto de especificidad indica que el modelo tiene una alta tasa de aciertos en la identificación de los casos negativos [34]. | $\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$<br>**Fórmula Especificidad Área** |
| **ÁREA BAJO LA CURVA DE FUNCIONAMIENTO DEL RECEPTOR (ROC) (AUC)** | Es una métrica utilizada en problemas de clasificación binaria para evaluar la capacidad de un modelo para discriminar entre clases positivas y negativas.<br>La curva ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación. El área bajo esta curva (AUC) proporciona una medida de la capacidad de discriminación del modelo: cuanto mayor sea el valor del AUC, mejor será la capacidad del modelo para distinguir entre las clases.<br>El AUC se calcula integrando el área bajo la curva ROC, que varía de 0 a 1. Un valor de AUC de 0.5 indica un rendimiento aleatorio o equivalente al azar, mientras que un valor de AUC cercano a 1 indica un rendimiento excelente del modelo en la clasificación.<br>El AUC es una métrica popular en la evaluación de modelos de clasificación, especialmente cuando los conjuntos de datos están desequilibrados. Proporciona una medida agregada de la precisión del modelo en todos los umbrales de clasificación posibles y es independiente del umbral de decisión específico utilizado [34]. |  <br>Fig. 5. Área Bajo la Curva |
| **PÉRDIDA LOGARÍTMICA** | Es una métrica utilizada en problemas de clasificación para evaluar la calidad de las predicciones de un modelo de | |

| | | |
|---|---|---|
| | aprendizaje automático.<br>La pérdida logarítmica mide la discrepancia entre las probabilidades predichas por el modelo y las etiquetas verdaderas de los ejemplos de entrenamiento. Cuanto más baja sea la pérdida logarítmica, mejor será el modelo en términos de su capacidad para asignar probabilidades adecuadas a las clases correctas.<br>La pérdida logarítmica penaliza de manera más pronunciada las predicciones incorrectas con probabilidades cercanas a 0 o 1. En problemas de clasificación binaria, se busca minimizar la pérdida logarítmica para obtener un modelo con mejores predicciones [36]. | $-(y\log(p)+(1-y)\log(1-p))$<br>**Pérdida Logarítmica**<br>Donde:<br>• y es la etiqueta verdadera del ejemplo (0 o 1).<br>• p es la probabilidad predicha por el modelo de que el ejemplo pertenezca a la clase positiva. |

**Figura 6.** Métricas de evaluación

**Metodología en Espiral**

El desarrollo en espiral es un modelo de ciclo de vida del software definido por primera vez por Barry Boehm en 1986,1 utilizados generalmente en la ingeniería de software.[47,48,49]

Las actividades de este modelo consisten en la confirmación en un espiral, en la que cada bucle o iteración representa un conjunto de actividades. Las actividades no están fijadas a ninguna prioridad, sino que las siguientes se eligen en función del análisis de riesgo, comenzando por el bucle interior. Un modelo de ciclo de vida en espiral tiene en cuenta fuertemente el riesgo que aparece a la hora de desarrollar software.[50,51] Para ello, se comienza mirando las posibles alternativas de desarrollo, se opta por la de riesgo más asumible y se hace un ciclo de la espiral.[52,53,54] Si el cliente quiere seguir haciendo mejoras en el software, se vuelve a evaluar las distintas nuevas alternativas y riesgos y se realiza otra vuelta de la espiral, así hasta que llegue un momento en el que el producto software desarrollado sea aceptado y no necesite seguir mejorándose con otro nuevo ciclo.[37]

**Variables de estudio**

*Definición nominal de las variables*

En las siguientes figuras se muestran las variables dependientes e independientes presentes en el estudio de la predicción.

| Nombre | Descripción | Tipo de Variable |
|---|---|---|
| **Riesgo Cardiovascular** | El riesgo cardiovascular es la probabilidad que tiene cada persona de sufrir una enfermedad cardiovascular, es decir, un infarto de miocardio, hemorragias cerebrales, embolias, etcétera. | Dependiente |
| **Tiempo de entrenamiento** | Es el tiempo que se toma un algoritmo en detectar los patrones de un conjunto de datos. | Dependiente |

**Figura 7.** Variables dependientes de estudio

| Nombre | Descripción | Unidad de Medida | Tipo de Variable | Valores Normales |
|---|---|---|---|---|
| **Edad** | Edad de la persona | Años | Independiente | 0-120 años |
| **Peso** | El peso es una medida de la fuerza gravitatoria que actúa sobre un objeto. | kilogramos | Independiente | IMC es entre 18.5 y 24.9, está dentro de los valores "normales" o de peso saludable. Si su IMC es entre 25.0 y 29.9, está dentro de los valores correspondientes a "sobrepeso". Si su IMC es 30.0 o superior, está dentro de los valores de "obesidad". |
| **Talla** | Medida de la persona en estatura | Centímetros | Independiente | 1-200 centímetros |
| **Género** | El género se refiere a los roles, las características y oportunidades definidos por la sociedad que se consideran apropiados para los hombres, las mujeres, los niños, las niñas y las personas con identidades no binarias. | No aplica | Independiente | No Aplica |
| **Tensión Arterial** | Medida de la fuerza o presión de la sangre sobre las arterias cuando el corazón bombea. | Milímetro de mercurio | Independiente | Presión sistólica de menos de 120 y una presión diastólica de menos de 80. |
| **Nivel de colesterol en sangre** | Mide la cantidad colesterol y de ciertos lípidos en la sangre. | Miligramos de colesterol por decilitro de sangre | Independiente | Menos de 200 mg/dL |
| **Nivel de glucosa en sangre** | Mide los niveles normales de azúcar en sangre | Miligramos de glucosa por decilitro de sangre | Independiente | Menor que 100 mg/dl |

**Figura 8.** Variables independientes de estudio

*Definición operativa de las variables*

Las variables independientes corresponden a variables que ya han sido medidas en grupo de datos proporcionados por investigadores expertos de importancia académica y de investigación, en la cual se confirma en el documento SEA 2022 para el control global del riesgo cardiovascular.[55,56,57] Para obtener los resultados de este estudio, se ha analizado un dataset (conjunto de datos) obtenido a través de la plataforma Kaggle, plataforma de Data Science que permite a los usuarios adquirir y publicar conjuntos de datos.

Las variables dependientes se pueden medir después de que se entrene cada modelo, partiendo de ello la exactitud (accuracy) representa el porcentaje de predicciones correctas frente al total. Por tanto, es el

cociente entre los casos bien clasificados por el modelo (verdaderos positivos y verdaderos negativos), y la suma de todos los casos.[20]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} - \text{TN} + \text{FP} + \text{FN}}$$

Ecuación 1. Accuracy

La precisión (precision) se refiere a lo cerca que está el resultado de una predicción del valor verdadero. Por tanto, es el cociente entre los casos positivos bien clasificados por el modelo y el total de predicciones positivas. Esta es priorizada en los casos donde tener una gran cantidad de falsos positivos tiene un mayor coste.[20,58]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Ecuación 2. Precision

La sensibilidad (recall) representa la tasa de verdaderos positivos (TP). Es la proporción entre los casos positivos bien clasificados por el modelo, respecto al total de positivos.[20,59]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Ecuación 3. Recall

El valor-F (F1-score) se utiliza para combinar las medidas de precision y recall en un solo valor, donde 0 es la peor puntuación y 1 la mejor. Se utiliza cuando es mejor reducir tanto falsos positivos como falsos negativos.[20,60]

$$\text{F1score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ecuación 4.F1-Score

$$\text{F1score} = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Ecuación 5. Modelo matemático de la evaluación de la medida-F

**Formulación de hipótesis**

*Hipótesis de investigación*

Las funciones kernel alternativas mejoran la relación error tiempo en la predicción de enfermedades cardiovasculares.

*Hipótesis nula*

Las funciones kernel alternativas no mejoran la relación error tiempo en la predicción de enfermedades cardiovasculares.

*Hipótesis alterna*

Las funciones kernel disminuyen el error, pero tardan más en dar una predicción.

## CONCLUSIONES

Las enfermedades cardiovasculares representan una de las principales causas de muerte a nivel mundial, siendo un problema crítico de salud pública debido a su alta incidencia y dificultad de diagnóstico temprano. A pesar de los avances en la comprensión de los factores de riesgo, aún persisten desafíos en la prevención, detección oportuna y tratamiento eficaz. En este contexto, el uso de tecnologías emergentes como la Inteligencia

Artificial (IA), en particular el aprendizaje automático y las funciones kernel, se posiciona como una alternativa prometedora para apoyar el diagnóstico y monitoreo de estas patologías.

Este estudio propone el uso de algoritmos de clasificación como las Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN), aplicando nuevas funciones kernel no tradicionales, con el fin de mejorar la precisión, eficiencia y utilidad de los modelos predictivos en la identificación de riesgo cardiovascular. La validación de esta propuesta, basada en bases de datos confiables y métricas rigurosas, demuestra que es posible optimizar el desempeño de los modelos mediante ajustes adecuados en los hiperparámetros y la elección de kernels adecuados, lo cual puede traducirse en herramientas útiles para el personal médico en escenarios reales.

Así, se concluye que la integración de modelos de aprendizaje automático en el campo de la salud no solo aporta una ventaja técnica en el análisis de grandes volúmenes de datos clínicos, sino que también tiene un valor social al facilitar diagnósticos más precisos, promover la prevención y contribuir a una atención médica más oportuna y personalizada. La implementación efectiva de estas tecnologías puede impactar positivamente en la reducción de la morbilidad y mortalidad asociada a las enfermedades cardiovasculares, así como en la promoción de estilos de vida saludables y la concienciación pública.

## REFERENCIAS BIBLIOGRÁFICAS

1. Patiño Zambrano CF. Dispositivo vestible inteligente para la generación de alertas tempranas de eventos cardiovasculares de riesgo. Envigado; 2022.

2. Sanofi Campus. Machine Learning y predicción de enfermedades cardiovasculares. 2022. https://campus.sanofi.es/es/noticias/machine-learning-prediccion-enfermedades-cardiovasculares

3. Gómez LA. Las enfermedades cardiovasculares: un problema de salud pública y un reto global. SciELO. 2011;1.

4. Pérez Leal LE, Buitrago Cárdenas JA. Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático. Bogotá: Universidad Antonio Nariño; 2021.

5. Gallego Valcárcel DA, Lucas Monsalve DF. Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardíaca con datos clínicos. Bogotá: Universidad Antonio Nariño; 2021.

6. Álvarez Vega M, Quirós Mora LM, Cortés Badilla MV. Inteligencia artificial y aprendizaje automático en medicina. Rev Méd Sinergia. 2020;5(8):12.

7. Mora Paz HA. Comparativo de Kernels sobre predicción de oferta de fuentes alternativas de energía. Pasto: UNIR La Universidad en Internet; 2019.

8. Friedman PA, Kapa S, López Jiménez F, Noseworthy PA. Inteligencia artificial en cardiología. Mayo Clinic; 2023. https://www.mayoclinic.org/es-es/departments-centers/ai-cardiology/overview/ovc-20486648

9. Aprende IA. Kernel y máquinas de vectores de soporte. https://aprendeia.com/kernel-maquinas-vectores-de-soporte-clasificacion-regresion/

10. Sowmya V, Sanjana K, Gopalakrishnan E, Soman KP. Inteligencia artificial explicable para la variabilidad de la frecuencia cardíaca en la señal de ECG. Health Technol Lett. 2020;7(6):146.

11. Wu H, Yang L, Jin X, Zheng P. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020;10(1):5245.

12. Chavez Olivera O, Galindo Honores L, Barrientos Padilla A, Cuadros Galvez M. Aplicación móvil para predecir la probabilidad de pertenecer al grupo de riesgo cardiovascular utilizando machine learning. En: XII Conf Iberoamericana de Complejidad, Informática y Cibernética. Lima; 2022.

13. Scavino M, Castrillejo A, Estragó Mérola VS, Luraghi López LE, Muñoz M, Álvarez Vaz R. Informe final publicable del proyecto de creación de algoritmos utilizando técnicas de clasificación supervisada y no supervisada para el diagnóstico de enfermedades cardiovasculares. Uruguay; 2022.

14. Polero LD, Garmendia CM, Echegoyen RE, Alves de Lima A, Bertón F, Lambardi F, et al. Predicción de

riesgo de sufrir un síndrome coronario agudo mediante un algoritmo de Machine Learning (ANGINA). Rev Argent Cardiol. 2020;88(1).

15. Perez Tatis JD. Optimización de un modelo de clasificación de enfermedades cardiovasculares utilizando técnicas de aprendizaje profundo supervisado y despliegue de dashboard web. Cartagena; 2021.

16. Martínez EJ. Predicción de enfermedades cardiovasculares mediante algoritmos de inteligencia artificial. Málaga; 2020.

17. Carrascal Porras FL, Florez Prias LA. Modelo de inteligencia artificial como apoyo diagnóstico para la estimación de riesgo cardiovascular en pacientes atendidos bajo la modalidad de telemedicina. Sucre: UNAD; 2021.

18. Dolores C, Ordovás J. Genes, dieta y enfermedades cardiovasculares. Genética. 2007;5:71–118.

19. Martínez EJ. Predicción de enfermedades cardiovasculares. Málaga: Universidad de Málaga; 2020.

20. Sans Menéndez S. Enfermedades cardiovasculares. Barcelona: Institut d'Estudis de la Salut; 2011.

21. Perez Tatis JD. Optimización de un modelo de clasificación de enfermedades cardiovasculares utilizando técnicas de aprendizaje profundo supervisado y despliegue de dashboard web. Cartagena: Universidad del Sinú; 2021.

22. MathWorks. Máquina de soporte vectorial (SVM). https://es.mathworks.com/discovery/support-vector-machine.html

23. Ecured. Función kernel. https://www.ecured.cu/Funci%C3%B3n_Kernel

24. Wikipedia. Clasificador lineal. 2019. https://es.wikipedia.org/wiki/Clasificador_lineal

25. Tibco Data Science. ¿Qué es el aprendizaje supervisado? https://www.tibco.com/es/reference-center/what-is-supervised-learning

26. Montiel de Jesús A. Desarrollo de una aplicación para dispositivos móviles para la detección temprana de enfermedades cardiovasculares. Orizaba, México: TECNM; 2022.

27. Sitio BigData. Modelos de Machine Learning: Métricas de regresión. 2019. https://sitiobigdata.com/2019/05/27/modelos-de-machine-learning-metricas-de-regresion-mse-parte-2/

28. Martínez Heras J. Métricas de clasificación: precisión, recall, F1, accuracy. IArtificial; 2020. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

29. Quintanilla L, Warren G, Kirsch S, Youssef V, Kershaw N, Killeen S, et al. Learn Microsoft: métricas de ML. 2023. https://learn.microsoft.com/es-es/dotnet/machine-learning/resources/metrics

30. Boehm B. Desarrollo en espiral. Wikipedia; 2012. https://es.wikipedia.org/wiki/Desarrollo_en_espiral

31. Ballina Ríos F. Paradigmas y perspectivas teórico-metodológicas en el estudio de la administración. UVMX; 2013.

32. Radrigán M. Método empírico-analítico. Wikipedia; 2022. https://es.wikipedia.org/wiki/M%C3%A9todo_emp%C3%ADrico-anal%C3%ADtico

33. IBM. Conceptos básicos de ayuda de CRISP-DM. 2021. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

34. Vallalta Rueda JF. CRISP-DM: una metodología para minería de datos en salud. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

35. Toral Barrera JA. Redes neuronales. España: CUCEI; 2019.

36. Asunción A, Newman D. Repositorios. Rexa.info: Massachusetts Amherst; 2007.

37. Mora Paz H, Riascos JA, Salazar Castro JA, Mora G, Pantoja A. Comparación de funciones kernel para la predicción de la oferta energética fotovoltaica. RISTI. 2020;(E38):310-24.

38. Belanche LA, Villegas MA. Kernel functions for categorical variables with application to problems in the life sciences. 2023;(08034):1-3.

39. Esri. Spatial analysis in ArcGIS Pro. https://pro.arcgis.com/es/pro-app/latest/help/analysis/introduction/spatial-analysis-in-arcgis-pro.htm

40. Scriptología. Tutorial de Flask: desarrollando aplicaciones web en Python. 2024. https://scriptologia.com/tutorial-de-flask-desarrollando-aplicaciones-web-en-python/

41. Hunter J, Dale D, Firing E, Droettboom M. Introducción a pyplot. Matplotlib; 2012. https://es.matplotlib.net/stable/tutorials/introductory/pyplot.html

42. Python. Biblioteca Pickle. 2001. https://docs.python.org/es/3/library/pickle.html

43. DataScientest. Uso de pandas en Python. 2023. https://datascientest.com/es/pandas-python

44. Manav N. Escribir bytes a archivo en Python. 2023. https://www.delftstack.com/es/howto/python/write-bytes-to-file-python/

45. Python. Biblioteca base64. https://docs.python.org/es/dev/library/base64.html

46. Navarro S. ¿Para qué sirve el train-test split? KeepCoding; 2024. https://keepcoding.io/blog/para-que-sirve-el-train-test-split/

47. Scikit Learn. Nystroem Kernel Approximation. 2007. https://scikit-learn.org/stable/modules/generated/sklearn.kernel_approximation.Nystroem.html

48. Li B, Lu P, chmccl v. Multiclass Neural Network. Microsoft Learn; 2023. https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2

49. Imbert A, Lemaitre G. sklearn.svm.SVC. Scikit-Learn; 2024. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

50. Ruiz M. ¿Qué es Redux? OpenWebinars; 2018. https://openwebinars.net/blog/que-es-redux/

51. Moes T. ¿Qué es Windows? SoftwareLab; 2023. https://softwarelab.org/es/blog/que-es-windows/

52. D. A. Qué es Bootstrap. Hostinger; 2023. https://www.hostinger.mx/tutoriales/que-es-bootstrap

53. Monk R. What is Python used for. Coursera; 2023. https://www.coursera.org/mx/articles/what-is-python-used-for-a-beginners-guide-to-using-python

54. Valiente FT. Aprendizaje por refuerzo. IIC UAM. https://www.iic.uam.es/inteligencia-artificial/aprendizaje-por-refuerzo

55. Arceo Vilas AM. Estado nutricional y adherencia a la dieta mediterránea en población mayor de 40 años: IA vs estadística clásica. A Coruña: Universidad de Coruña; 2020.

56. Sánchez Santos JM, Sánchez Fernández PL. Predicción de eventos cardiovasculares y hemorrágicos en pacientes con doble antiagregación con modelos ML. CREDOS. Salamanca; 2020.

57. Gallego Valcárcel DA, Lucas Monsalve DF. Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardiaca con datos clínicos. Bogotá: Universidad Antonio Nariño; 2021.

58. Lozada J. Investigación aplicada. Dialnet UNIRIOJA. 2014;3(1):47-50.

59. Núñez Cárdenas FJ, Zavaleta Chi IDC, Felipe Redondo AM, Meléndez Hernández J. Aplicación de minería de datos para tipificación de ECV en alumnos universitarios. México; 2018.

60. Martínez J. Más allá del accuracy: precision, recall y F1. Datasmarts; 2019. https://datasmarts.net/es/mas-alla-del-accuracy-precision-recall-y-f1/

## FINANCIACIÓN

Ninguna.

## CONFLICTO DE INTERESES

Ninguno.

## CONTRIBUCIÓN DE AUTORÍA

*Conceptualización:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Curación de datos:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Análisis formal:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Investigación:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Metodología:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Administración del proyecto:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Recursos:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Software:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Supervisión:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Validación:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Visualización:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Redacción – borrador original:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

*Redacción – revisión y edición:* Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.