Diseño de una Herramienta para la Detección de Arritmias Cardíacas en Electrocardiogramas utilizando Técnicas de Aprendizaje Automático

Robinson Álvarez Patiño
Alejandro Ruiz Luna

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Maria Bernarda Salazar Sánchez, Doctor (PhD) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	Álvarez Patiño & Ruiz Luna [1]
Referencia Estilo IEEE (2020)	[1] A. Ruiz Luna y R. Álvarez Patiño “Diseñar una herramienta para la detección de arritmias cardíacas en electrocardiogramas utilizando técnicas de aprendizaje automático”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA)

Universidad de Antioquia - www.udea.edu.co
Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama.

Dedicatoria

"A nuestros apreciados profesores, quienes con su dedicación y conocimientos nos brindaron las herramientas necesarias para abordar los desafíos de esta investigación. Su guía y enseñanza nos permitieron explorar y comprender en profundidad los métodos y procedimientos necesarios para encontrar la mejor técnica frente a esta problemática específica.

A nuestra asesora, la PhD. María Bernarda, agradecemos su valiosa contribución y apoyo constante a lo largo del desarrollo de esta monografía. Su experiencia y perspectiva fueron fundamentales para orientarnos en el camino correcto y alcanzar resultados óptimos.

A Federico y Sara, las inspiraciones e impulsores para crecer profesionalmente.

¡A ellos, nuestro más sincero agradecimiento!"

Agradecimientos

A nuestras familias que sacrificaron horas de compañía y a la Dra. Maria Bernarda Salazar Sánchez quién con su activa participación en el proyecto dirigió y apoyo de manera significativa los resultados del estudio.

1. TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

SVM.	Support Vector Machine
Esp.	Especialista
RF.	Random Forest
RL	Regresión Logistica
KNN.	K Nearest Neighbor
TP.	True Positive
TN.	True Negative
FP.	False Positive
FN.	False Negative
PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia

1. RESUMEN

Esta base de datos de investigación para señales de electrocardiograma (ECG) de 12 derivaciones (permite visualizar la actividad eléctrica del corazón desde la perspectiva frontal y horizontal) fue creada bajo los auspicios de la Universidad Chapman, el Hospital del Pueblo de Shaoxing (Escuela de Medicina de la Universidad de Zhejiang del Hospital Shaoxing) y el Primer Hospital de Ningbo. Su objetivo es permitir a la comunidad científica realizar nuevos estudios sobre arritmia y otras afecciones cardiovasculares. Ciertos tipos de arritmias, como la fibrilación auricular, tienen un impacto negativo pronunciado en la salud pública, la calidad de vida y los gastos médicos. Como prueba no invasiva, el ECG es una herramienta diagnóstica importante y vital para detectar estas condiciones. Esta práctica, sin embargo, genera grandes cantidades de datos, cuyo análisis requiere un tiempo y esfuerzo considerable por parte de expertos humanos. Las herramientas modernas de aprendizaje automático y estadísticas pueden ser entrenadas en datos grandes y de alta calidad para lograr niveles excepcionales de precisión diagnóstica automatizada. Por lo tanto, recopilamos y difundimos esta nueva base de datos que contiene ECGs de 12 derivaciones de 10.646 pacientes con una frecuencia de muestreo de 500 Hz que presenta múltiples ritmos comunes y condiciones cardiovasculares adicionales, todos etiquetados por expertos profesionales. El conjunto de datos se puede utilizar para diseñar, comparar y ajustar nuevas y clásicas técnicas estadísticas y de aprendizaje automático en estudios centrados en arritmia y otras afecciones cardiovasculares.

El presente estudio da a conocer los métodos y procedimientos que utilizamos para encontrar la técnica que tiene mejores resultados frente a este problema en particular. Ilustra los procesos de exploración de datos y tratamiento de datos (especialmente por ser un dataset muy des balanceado), las métricas que usamos para considerar los resultados óptimos y nuestra recomendación final incluyendo hiperparámetros.

2. ABSTRACT

This research database for 12-lead electrocardiogram (ECG) signals (allows viewing of the electrical activity of the heart from frontal and horizontal perspective) was created under the auspices of Chapman University, Shaoxing People's Hospital (School of Zhejiang University Medicine Shaoxing Hospital) and Ningbo First Hospital. Its objective is to allow the scientific community to carry out new studies on arrhythmia and other cardiovascular conditions. Certain types of arrhythmias, such as atrial fibrillation, have a pronounced negative impact on public health, quality of life, and medical costs. As a non-invasive test, the ECG is an important and vital diagnostic tool in detecting these conditions. This practice, however, generates large amounts of data, the analysis of which requires considerable time and effort on the part of human experts. Modern machine learning and statistical tools can be trained on big, high-quality data to achieve exceptional levels of automated diagnostic accuracy. Therefore, we compiled and disseminated this new database containing 12-lead ECGs of 10,646 patients at 500 Hz sampling rate presenting multiple common rhythms and additional cardiovascular conditions, all labeled by professional experts. The data set can be used to design, compare, and adjust new and classic statistical and machine learning techniques in studies focused on arrhythmia and other cardiovascular conditions.

The present study discloses the methods and procedures that we use to find the technique that has the best results against this problem. It illustrates the data exploration and data treatment processes (especially for being a very unbalanced dataset), the metrics we use to consider the optimal results and our final recommendation including hyperparameters.

3. INTRODUCCIÓN

Las arritmias cardiacas son alteraciones del ritmo normal del corazón que pueden causar síntomas como palpitaciones, mareos, falta de aire o desmayos. Algunas arritmias pueden aumentar el riesgo de sufrir complicaciones graves, como accidente cardiovascular, muerte súbita o insuficiencia cardíaca afectado de forma directa la calidad de vida o la salud del paciente [1].

El diagnóstico de arritmias cardiacas es un proceso que requiere evaluación clínica y en algunas ocasiones la realización de pruebas complementarias, con las cuales se busca identificar tipo, frecuencia, duración y causa de las alteraciones del ritmo cardiaco. Algunas de las pruebas más utilizadas son el electrocardiograma (ECG), el holter, el estudio electrofisiológico y el ecocardiograma [2]. Estas permiten registrar la actividad eléctrica del corazón, medir la presión y el flujo sanguíneo en las cámaras cardiacas, y detectar posibles anomalías estructurales o funcionales. Este diagnóstico es importante para establecer el tratamiento más adecuado y prevenir complicaciones posteriores como insuficiencia cardiaca, infarto o accidente cerebrovascular [3].

Usualmente, el diagnóstico se realiza apoyado en el análisis manual del ECG, el cual es un proceso lento, costoso y propenso a errores dado que depende de la experticia de un profesional. Por ello, en el presente trabajo se propone el desarrollo de un modelo que permita detectar automáticamente arritmias cardíacas a partir de la información que se puede extraer de un ECG como herramienta de apoyo al diagnóstico realizado por el profesional experto. Para el desarrollo del modelo se utilizará una base de datos de 10.646 pacientes anonimizados, proveniente del Shaoxing Hospital Zhejiang University School of Medicine, categorizadas de acuerdo con once tipos de arritmias previamente etiquetadas por médicos especialistas.

Se utilizaron técnicas de aprendizaje supervisado y no supervisado, esta última con el fin de experimentar si era posible lograr la agrupación de los datos de manera natural, sin embargo, los resultados no fueron tan contundentes como con las técnicas supervisadas, entre las cuales están: Support Vector Machine, Random Forest y Deep Learning. Para determinar el algoritmo con mejor rendimiento se utilizan las mismas métricas de desempeño para cada una de las iteraciones (Accracy, Positive Predictive Value, Recall y F1-score). Los algoritmos de aprendizaje no supervisado que usamos (kmeans, BDSCAN y Spectral Cluster) no lograron un nivel adecuado de agrupamiento, pero dan ideas importantes de como gran parte de los pacientes que una arritmia particular tienen características comunes.

4. PLANTEAMIENTO DEL PROBLEMA

La detección de arritmias cardiacas es un procedimiento que requiere una alta intervención humana y en contraste se hace difícil su detección debido a que en algunos casos los pacientes pueden presentar síntomas leves, intermitentes o no ocurrir durante la consulta médica.

Según la Organización Mundial de la Salud (OMS), en el mundo se estima que alrededor de 33,5 millones de personas padecen arritmias cardíacas. Estas patologías representan un problema de salud pública importante, ya que pueden aumentar el riesgo de sufrir un accidente cerebrovascular o una insuficiencia cardíaca, entre otras complicaciones graves. En Colombia, aunque no existen estadísticas precisas, se estima que alrededor de 600.000 personas tienen arritmias cardíacas, según la Asociación Colombiana de Cardiología. Además, se ha observado un aumento en la incidencia de estas patologías en los últimos años, lo que hace aún más importante la detección temprana y la implementación de medidas preventivas [4].

La detección de arritmias es en algunos casos un proceso complejo debido a que algunos pacientes pueden presentar síntomas leves o intermitentes, y el proceso de análisis del ECG es tedioso y requiere una gran intervención humana por parte de especialistas. Por esta razón, el uso de algoritmos de inteligencia artificial y dispositivos ECG-enabled puede ser de gran ayuda para facilitar la detección temprana de arritmias y mejorar la atención médica en este campo.

A todo esto, se le suma la necesidad de detección temprana para prevenir complicaciones graves, como accidentes cerebrovasculares y muerte súbita. Por tanto, se pretende eliminar la barrera que separa los datos del ECG del especialista y brindar al médico de primer nivel una herramienta que permita tomar acción con base en los resultados emitidos por el algoritmo. Así mismo, estas tecnologías se abren a un sinfín de posibilidades con los dispositivos ECG-enabled, tales como manillas o relojes inteligentes, que permiten de manera temprana, oportuna y masiva encontrar posibles anomalías en los procesos de despolarización y repolarización del músculo cardiaco.

5. JUSTIFICACIÓN

La detección rápida y temprana de arritmias cardiacas pueden impactar positivamente la calidad de vida de los pacientes y promover la salud pública de la mayoría de los países desarrollados, permitiendo la intervención médica temprana de accidentes cerebrovasculares, insuficiencia cardíaca y muerte súbita. Cuanto antes se detecten y se traten las arritmias, mejor será el pronóstico a largo plazo para el paciente [5]. Así mismo, la reducción de la mortalidad [5] y la realización de un diagnóstico más preciso y una evaluación más completa de la enfermedad subyacente, puede ayudar a los médicos a identificar la causa subyacente de la arritmia y seleccionar el tratamiento más adecuado para el paciente [6].

Apoyarse en nuevas herramientas diagnósticas basadas en tecnología que utilicen técnicas de aprendizaje automático, pueden contribuir a la generación de tratamiento oportunos de las arritmias, que se traducen en mejora de la calidad de vida de los pacientes y sus familias [7].

6. OBJETIVOS

1.1. Objetivo general

Desarrollar un modelo de detección de arritmias cardiacas a partir de electrocardiogramas utilizando técnicas de aprendizaje automático que pueda ser accesible a través de una herramienta informática.

1.2. Objetivos específicos

- Identificar el conjunto de características que más aporten a la clasificación de los diferentes tipos de arritmias cardiacas.
- Proponer y evaluar una estrategia de clasificación temprana de arritmias cardiacas basada en técnicas de aprendizaje automático.
- Validar mediante métricas de desempeño la capacidad predicción del modelo de clasificación de arritmias cardiacas.

7. MARCO TEÓRICO

1.3.Sistema Cardiovascular

1.3.1. Arritmias cardíacas.

Las arritmias cardíacas son alteraciones en el ritmo normal del corazón que pueden manifestarse en el electrocardiograma (ECG) como anomalías en la actividad eléctrica del corazón [8]. Entre los diferentes tipos de arritmias cardiacas se encuentran las siguientes:

A. **Taquicardia.** Caracterizada por una frecuencia cardíaca rápida, superior a 100 latidos por minuto en adultos. Puede ser sinusal (causada por un aumento en la actividad del nodo sinusal) o ventricular (producida por impulsos eléctricos anormales que se originan en los ventrículos).

B. **Bradicardia.** Se presenta cuando la frecuencia cardíaca es anormalmente lenta, inferior a 60 latidos por minuto en adultos. Puede ser sinusal (causada por una disminución en la actividad del nodo sinusal) o ventricular (producida por una interrupción en la conducción de los impulsos eléctricos a través del corazón).

C. **Fibrilación auricular.** Es una de las arritmias más comunes y se caracteriza por una actividad eléctrica caótica en las aurículas, lo que causa una frecuencia cardíaca irregular y rápida. Puede causar síntomas como palpitaciones, fatiga y dificultad para respirar.

D. **Flutter auricular.** Similar a la fibrilación auricular, pero la actividad eléctrica es más organizada y la frecuencia cardíaca es típicamente más rápida. También puede causar síntomas similares.

E. **Bloqueo de rama.** Se produce cuando hay una interrupción en la conducción eléctrica a través de una de las ramas del sistema de conducción del corazón. Puede causar una variedad de síntomas, dependiendo de la gravedad del bloqueo.

1.3.2. Electrocardiograma

El ECG es un registro gráfico de la actividad eléctrica del corazón que se obtienen colocando electrodos en el cuerpo del paciente, constan de ondas características: onda P, complejo QRS y onda T, que son asociados a la actividad eléctrica en diferentes momentos del ciclo cardíaco.

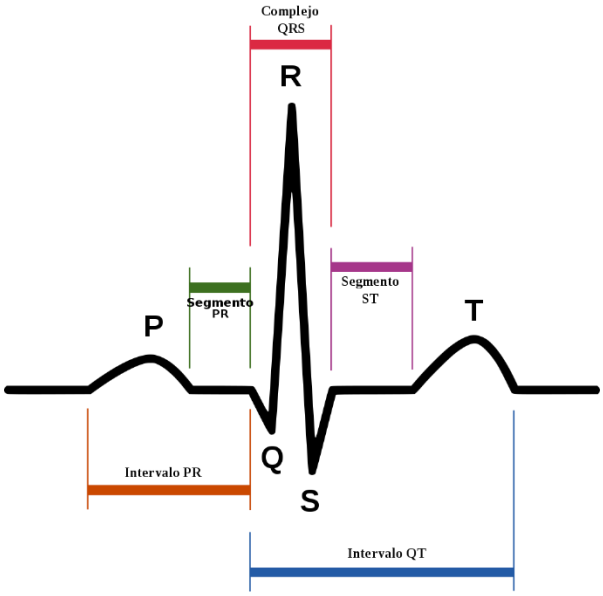

Fig. 1 Onda Electrocardiograma.

Tomada de: https://es.wikipedia.org/wiki/Electrocardiograma

1.4.Técnicas de aprendizaje automático

Existen diferentes modelos de aprendizaje automático que se pueden utilizar en la detección de arritmias cardíacas, como árboles de decisión, regresión logística, redes neuronales, máquinas de vectores de soporte, entre otros. Estos modelos pueden ser entrenados con datos de ECGs etiquetados para aprender a clasificar entre ritmos cardíacos normales y arritmias. Es importante comprender cómo funcionan estos modelos, sus ventajas y desventajas, y cómo se pueden aplicar en el contexto de la detección de arritmias cardíacas.

A. **Regresión Logística.** Este modelo estima la probabilidad de que una observación pertenezca a una de las dos categorías de la variable dependiente, en función de las variables independientes. Esta probabilidad se transforma en una estimación de odds

ratio (OR), que indica la probabilidad relativa de que una observación pertenezca a una categoría de la variable dependiente en comparación con la otra [9].

B. **Random Forest.** Es un algoritmo de aprendizaje automático que se utiliza para problemas de clasificación y regresión. Consiste en la creación de múltiples árboles de decisión y la combinación de sus resultados para mejorar la precisión de la predicción. Cada árbol se construye de manera independiente, utilizando una muestra aleatoria de los datos y una selección aleatoria de las variables, lo que reduce la correlación entre los árboles y mejora la generalización del modelo. El modelo Random Forest es ampliamente utilizado en la detección de arritmias cardíacas, ya que puede manejar datos con múltiples variables y características complejas. Además, puede identificar las variables más importantes para la clasificación y proporcionar una medida de su importancia relativa [10].

C. **Support Vector Machine** (SVM). Son un algoritmo de aprendizaje supervisado ampliamente utilizado en la clasificación y regresión de datos. Su principal objetivo es encontrar un hiperplano que separe los datos en diferentes clases de forma óptima, maximizando la distancia entre las muestras más cercanas de las diferentes clases, lo que se conoce como margen máximo. Esto las convierte en una técnica muy efectiva en el aprendizaje automático, utilizada en diversos campos, como la clasificación de texto, la detección de rostros, la detección de anomalías y la clasificación de imágenes. Además, tienen la capacidad de trabajar con conjuntos de datos no lineales, grandes y de alta dimensión [11].

D. **K-Nearest Neighbors Algorithm (K-NN).** Es un método de clasificación supervisado utilizado en el aprendizaje automático. La idea principal detrás del algoritmo K-NN es asignar una clase a un punto desconocido basándose en las clases de los puntos conocidos más cercanos a él en el espacio de características. El valor de K se refiere al número de vecinos más cercanos que se tienen en cuenta para tomar la decisión de clasificación. Este algoritmo se utiliza en una amplia variedad de campos, incluyendo la visión por computadora, la bioinformática y la minería de

datos. Es fácil de implementar y entender, y es especialmente útil cuando los datos tienen estructuras no lineales o no se ajustan a un modelo paramétrico específico [12].

1.5.Métricas de evaluación

Las métricas de evaluación de algoritmos son herramientas importantes para evaluar la precisión y el rendimiento de los modelos de aprendizaje automático. A continuación, se describen algunas de las métricas de evaluación más comunes.

A. **Exactitud.** Es una medida de la proporción de muestras clasificadas correctamente. Se define como el número de verdaderos positivos (TP) y verdaderos negativos (TN) dividido por el número total de muestras. Su fórmula matemática es la siguiente:

$$\text{precisión} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

.

B. **Sensibilidad.** Mide la proporción de verdaderos positivos (TP) y se define como el número de verdaderos positivos dividido por el número total de muestras positivas. Su fórmula matemática es la siguiente:

$$\text{sensibilidad} = TP / (TP + FN) \quad (2)$$

C. **Especificidad.** Mide la proporción de verdaderos negativos (TN) y se define como el número de verdaderos negativos dividido por el número total de muestras negativas. Su fórmula matemática es la siguiente:

$$\text{especificidad} = TN / (TN + FP) \quad (3)$$

D. **Precisión ponderada.** Es considerada la precisión de cada clase y la frecuencia de esa clase en los datos. Se define como la suma de la precisión de cada clase multiplicada por su frecuencia, dividido por el número total de muestras. Su fórmula matemática es la siguiente:

$$\text{precisión ponderada} = (\text{precisión de clase 1} * \text{frecuencia de clase 1} + \text{precisión de clase 2} * \text{frecuencia de clase 2} + \ldots + \text{precisión de clase n} * \text{frecuencia de clase n}) / \text{número total de muestras} \quad (4)$$

E. **Área bajo la curva (AUC).** Es una medida de la capacidad del modelo para distinguir entre las clases. Se puede calcular trazando la curva ROC (Receiver Operating Characteristic) y calculando el área debajo de la curva. Su valor varía entre 0 y 1, siendo 1 el valor óptimo. Cuanto mayor sea el valor de AUC, mejor será el rendimiento del modelo. Es importante tener en cuenta que no hay una métrica única que sea adecuada para todos los casos de uso. La elección de la métrica dependerá de los objetivos del proyecto y de las características específicas del conjunto de datos [13].

1.6.Técnicas de reducción de dimensión

Las técnicas de reducción de dimensión son métodos que se utilizan para reducir el número de características de un conjunto de datos mientras se conserva la mayor cantidad posible de información relevante. Algunas de las técnicas de reducción de dimensión más comunes son las siguientes:

A. **Análisis de componentes principales (PCA).** Es una técnica de reducción de dimensión lineal que se utiliza para transformar un conjunto de datos de alta dimensión en un conjunto de datos de baja dimensión. PCA encuentra la combinación lineal de características que explica la mayor varianza en los datos y la proyecta en un nuevo espacio de características de menor dimensión.

B. **Análisis discriminante lineal (LDA).** Es una técnica de reducción de dimensión lineal que se utiliza para maximizar la separación entre clases en un conjunto de datos. LDA encuentra la combinación lineal de características que maximiza la relación entre la varianza inter-clases y la varianza intra-clases y la proyecta en un nuevo espacio de características de menor dimensión.

C. **T-SNE (t-Distributed Stochastic Neighbor Embedding).** Es una técnica de reducción de dimensión no lineal que se utiliza para visualizar datos en un espacio de baja dimensión. T-SNE conserva las relaciones locales entre las muestras y reduce la

dimensión de manera que las muestras similares se mapeen a puntos cercanos en el espacio de baja dimensión.

D. **Autoencoders.** Son una técnica de reducción de dimensión no lineal que se utiliza para aprender una representación latente de los datos, consisten en una red neuronal que se entrena para reconstruir los datos de entrada a través de una capa oculta de baja dimensión.

1.7.Técnicas de selección de características

La selección de características o feature selection es un paso importante en el diseño de una herramienta de detección de arritmias cardíacas. Se utilizan diferentes métodos de selección de características, como el análisis de componentes principales, selección basada en información mutua, selección por regularización, entre otros, para identificar las características más relevantes que contribuyen a la detección precisa de arritmias en un ECG.

1.8.Técnicas de validación

Las técnicas de validación son métodos que se utilizan para evaluar el rendimiento de un modelo de aprendizaje automático en un conjunto de datos. Algunas de las técnicas de validación más comunes son las siguientes:

A. **Validación cruzada.** Es una técnica de validación que se utiliza para evaluar la capacidad de generalización de un modelo. La validación cruzada divide el conjunto de datos en varios subconjuntos y entrena el modelo en cada subconjunto mientras se evalúa en el resto de los datos.

B. **Hold-out**. Es una técnica de validación que se utiliza para dividir el conjunto de datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. El modelo se entrena en el conjunto de entrenamiento y se evalúa en el conjunto de prueba.

C. **Validación en conjunto.** Es una técnica de validación que se utiliza para combinar los resultados de varios modelos entrenados en subconjuntos diferentes del conjunto de datos

D. **Validación por bootstrapping**. Es una técnica de validación que se utiliza para evaluar el rendimiento de un modelo utilizando múltiples subconjuntos de datos de entrenamiento generados por muestreo con reemplazo.

8. ESTADO DEL ARTE

El electrocardiograma (EKG o ECG, por sus siglas en inglés) fue inventado por Willem Einthoven (1860-192), un fisiólogo holandés, en el año 1901. Einthoven desarrolló el primer electrocardiógrafo, que era una máquina que podía registrar la actividad eléctrica del corazón en forma de gráficos. Su invención del electrocardiograma permitió a los médicos estudiar la actividad eléctrica del corazón y obtener información valiosa sobre su funcionamiento, lo que ha sido fundamental en el diagnóstico y tratamiento de enfermedades cardiovasculares. Por su trabajo pionero en el campo de la electrocardiografía, Willem Einthoven fue galardonado con el Premio Nobel de Medicina en 1924.

Willem Einthoven desarrolló el electrocardiograma con el objetivo de estudiar y comprender la actividad eléctrica del corazón, con la esperanza de mejorar la comprensión de las enfermedades cardiovasculares y su diagnóstico. En esa época, se sabía que el corazón generaba electricidad durante su funcionamiento, pero no se tenía una forma precisa de medir y registrar esta actividad eléctrica. Gracias al electrocardiograma, se pueden detectar y diagnosticar una amplia variedad de condiciones cardíacas, como arritmias, enfermedades del músculo cardíaco, obstrucciones en las arterias coronarias y otros trastornos cardiovasculares. El invento de Einthoven ha tenido un impacto significativo en la medicina cardiovascular, proporcionando una herramienta diagnóstica invaluable y ayudando a salvar vidas al facilitar la detección temprana y el tratamiento adecuado de enfermedades del corazón.

En la década de 1990, se utilizaron algoritmos basados en reglas para la detección de arritmias cardiacas. Estos algoritmos requerían la identificación manual de características en las señales de ECG y la definición de reglas para detectar arritmias. A pesar de su simplicidad, estos algoritmos no eran muy precisos y tenían dificultades para detectar arritmias complejas.

En la década de 2000, se empezaron a utilizar técnicas de Machine Learning para la detección de arritmias cardiacas. Se utilizaron diferentes técnicas de clasificación, como Redes Neuronales Artificiales, Máquinas de Vectores de Soporte, Árboles de Decisión, etc. Estas técnicas fueron capaces de aprender automáticamente las características de las señales de ECG y clasificarlas en diferentes categorías.

En la última década, se ha producido un gran avance en la detección de arritmias cardiacas a través de Machine Learning gracias a la disponibilidad de grandes conjuntos de datos y al

desarrollo de algoritmos más sofisticados. En particular, el uso de técnicas de Deep Learning, como las Redes Neuronales Convolucionales y las Redes Neuronales Recurrentes, ha permitido la detección de arritmias con una precisión sin precedentes.

Hoy en día, la detección de arritmias cardiacas a través de Machine Learning es un área de investigación muy activa, y se están desarrollando nuevos algoritmos y técnicas para mejorar aún más la precisión y la eficiencia de la detección. La detección temprana y precisa de arritmias cardiacas es esencial para prevenir complicaciones graves y mejorar la calidad de vida de los pacientes.

Ahora existen varios enfoques diferentes para utilizar el Machine Learning en la detección de arritmias cardíacas. Uno de los enfoques más comunes es utilizar redes neuronales convolucionales (CNN), que son una clase de redes neuronales diseñadas específicamente para analizar datos de imágenes o señales. En este caso, los datos de ECG se tratan como una imagen y la CNN busca patrones en los datos que puedan indicar la presencia de una arritmia.

Otro enfoque común es utilizar redes neuronales recurrentes (RNN), que son una clase de redes neuronales diseñadas para analizar datos de secuencia, como el habla o el texto. En este caso, los datos de ECG se tratan como una secuencia de puntos en el tiempo y la RNN busca patrones en los datos que puedan indicar la presencia de una arritmia.

9. METODOLOGÍA

La metodología usada durante el proyecto se esquematiza en la Fig. 2. Cada una de las fases de describen a continuación.

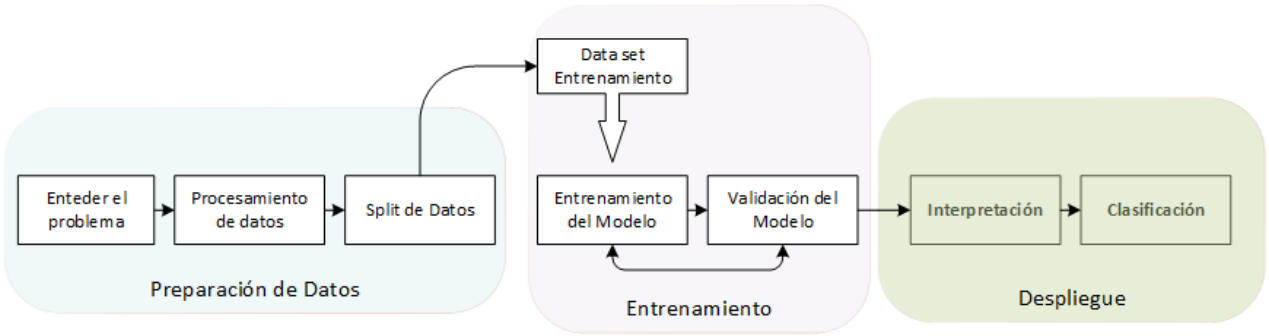

Fig. 2 Fases de la metodología usada

1. **Entender el problema:** Conocer las características entregadas por el dataset, los tipos de arritmias y sus posibles agrupaciones.
2. **Pre-Procesamiento de datos:** Se colectaron datos públicos expuestos. Un conjunto de 10.646 individuos todos de origen asiático. Se evalúan las características, en búsqueda de atípicos y variables con alta correlación. Además, al ser un problema multi-clases con número de muestras muy diferentes, se requiere considerar diferentes técnicas de ajuste de entre dos pares de variables (VentricularRate-QRSCount y QTInterval-Toffset). También se evidencio un pronunciado desbalance de clases con el objetivo de determinar si este comportamiento afectaba drásticamente el comportamiento de los modelos.
3. **Entrenamiento del Modelo:** Para esta fase se utiliza un 80 % de los datos para entrenamiento y el 20 % restante serán los datos de prueba. Los modelos desarrollados, considerando lo encontrado en el estado del arte serán Maquinas de Soporte Vectorial, Arboles Aleatorios, Regresión Logística y Redes Neuronales. Dado que el problema

abordado en este trabajo es de clasificación se harán pruebas con Kmeans, MiniBatch Kmeans, DBScan y Spectral Cluster.

4. **Validación del Modelo:** En esta etapa, el análisis y selección del mejor modelo se realiza de acuerdo a las métricas de desempeño generadas a partir de la matriz de confusión.
5. **Despliegue del Modelo:** Es de interés que este modelo pueda ser accesible en una plataforma abierta, se dejará disponible en GitHub en el repositorio: https://github.com/Alejoruizl/Monografia

10. RESULTADOS Y DISCUSIÓN

1.9. Dataset

Este dataset contiene 10.646 registros de pacientes con diagnósticos de arritmia, con sus correspondientes variables de escala de predicción de arritmia. Contiene 14 atributos, de los cual 11 son valores númericos y el resto categóricas. El propósito es clasificarlo en uno de los 11 tipos de arritmias.

La base de datos está disponible en: https://physionet.org/content/ecg-arrhythmia/1.0.0/

Tabla I. Tipos de arritmias

Acrónimo	**Nombre en el Dataset**	**Nombre Español**
SB	Sinus bradycardia	Bradicardia Sinusual
SR	Sinus rhythm	Ritmo sinusual
AFIB	Atrial fibrillation	Ribrilación auricular
ST	Sinus tachycardia	Taquicardia sinusal
AF	Atrial flutter	Aleteo auricular
SI	Sinus irregularity	Irregularidad sinusal
SVT	Supraventricular tachycardia	Taquicardia supraventricular
AT	Atrial tachycardia	Taquicardia auricular
AVNRT	Atrioventricular node reentrant tachycardia	Taquicardia por reentrada del nódulo auriculoventricular
AVRT	Atrioventricular reentrant tachycardia	Taquicardia por reentrada auriculoventricular
SAAWR	Sinus atrium to atrial wandering rhythm	Aurícula sinusal a ritmo errante auricular

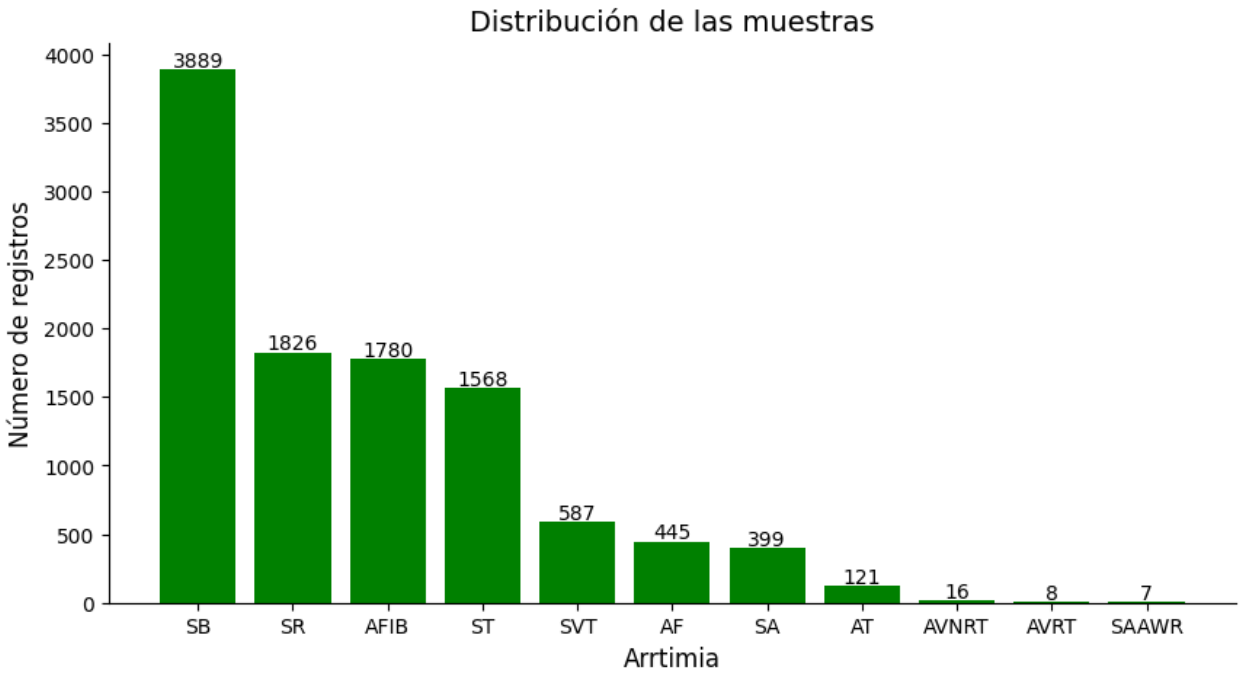

Fig. 3 Distribución de las Muestras

1.10. Preprocesamiento

1.10.1. Desbalance de clases

El desbalanceo se presenta cuando para el entrenamiento de un modelo de aprendizaje automático en un problema de clasificación, una de las clases tiene significativamente más muestras que las otras.

1.10.2. Correlación (Pearson)

Análisis valores de correlación altos son considerados aquellos mayores a 0.8 (ver Fig. 6).

- **VentricularRate y QRSCount**. Ambos son parámetros que están relacionados con la contracción del ventrículo, y es común que a mayor frecuencia cardíaca, mayor cantidad de complejos QRS se produzcan en un cierto período de tiempo. Por lo tanto, hay una correlación esperable entre estos dos parámetros.

- **QTInterval y Toffset.** Existe una correlación entre el QTInterval y TOffset porque ambos son medidas que reflejan la duración de diferentes etapas de la actividad eléctrica del corazón. El intervalo QT es la medida desde el inicio del QRS hasta el final de la onda T, lo cual refleja el tiempo total que tarda el corazón en contraerse y luego relajarse durante un latido. Por otro lado, el TOffset es la medición desde el final de la onda T hasta el final del complejo QRS, lo que indica la duración del período de la repolarización ventricular. Cuando se produce una alteración en la duración de la repolarización (como en el caso de ciertas arritmias o trastornos metabólicos), esto puede afectar tanto al QTInterval como al TOffset. Debido a que estas dos mediciones están estrechamente relacionadas y ambas pueden ser marcadores de cambios en la actividad eléctrica del corazón, es posible que presenten correlación. Además, otros factores (como la edad, el género y las anomalías cardíacas) también pueden influir en la duración del QTInterval y TOffset, lo que podría contribuir aún más a la correlación entre ellos.

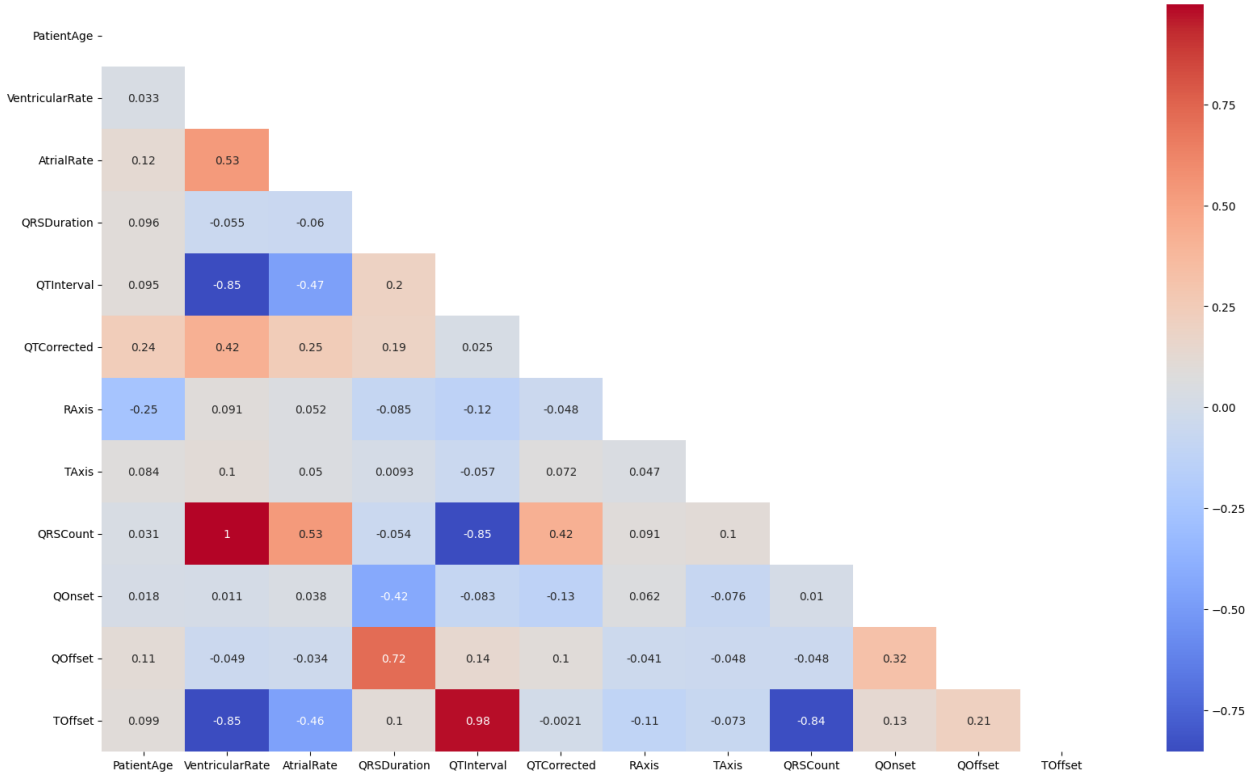

Fig. 6 Análisis de colinealidad entre las características de la base de datos

1.11. Desarrollo del modelo

1.12. Estructura del repositorio

El repositorio asociado al detalle del desarrollo del modelo se encuentra disponible en: https://github.com/Alejoruizl/Monografia

Los notebooks están organizados en el siguiente orden:

- Deteccion_Arritmias_Eleccion_Modelo
- Deteccion_Arritmias_Otros_Modelos

Se recomienda seguir el orden en el que se presentan los notebooks para una mejor comprensión del proceso completo. Cada notebook contiene comentarios detallados que explican el código y los resultados obtenidos.

11. CONCLUSIONES

- Emplear técnicas de búsqueda exhaustiva de hiperparámetros como validación cruzada y grid search permiten encontrar una mejor combinación de hiperparámetros que maximiza en más del 30% el rendimiento del modelo en un conjunto de datos con desbalance de clases como el utilizado en este proyecto.

- Los modelos obtenidos con accuracy de 96% contribuyen a la generación de herramientas de apoyo diagnóstico de arritmias a partir de señales de ECG para pacientes con características similares a los utilizados para entrenar el modelo de clasificación.

12. Referencias

[1] Personal de Mayo Clinic, «Mayo Clinic,» 21 Abril 2023. [En línea]. Available: https://www.mayoclinic.org/es-es/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668#:~:text=Las%20complicaciones%20dependen%20del%20tipo,mayor%20riesgo%20de%20co%C3%A1gulos%20sangu%C3%ADneos. [Último acceso: 22 Mayo 2023].

[2] MedlinePlus, «medlineplus.gov,» 5 Agosto 2022. [En línea]. Available: https://medlineplus.gov/spanish/ency/article/001101.htm#:~:text=Es%20un%20trastorno%20de%20la,peligro%20inmediato%20para%20su%20salud. [Último acceso: 4 Abril 2023].

[3] Personal Mayo Clinic, «mayoclinic.org,» 21 Abril 2023. [En línea]. Available: https://www.mayoclinic.org/es-es/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668#:~:text=Las%20complicaciones%20dependen%20del%20tipo,mayor%20riesgo%20de%20co%C3%A1gulos%20sangu%C3%ADneos. [Último acceso: 4 Mayo 2023].

[4] Organizacion Mundial de la Salud, «who.int,» 17 Mayo 2017. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Último acceso: 17 Mayo 2023].

[5] H. C. L. Samuel Wann, «AHA Journals,» 28 Enero 2019. [En línea]. Available: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000665. [Último acceso: 23 Abril 2023].

[6] M. J. A. William G. Stevenson, «AHA Journal,» 1 Agosto 2018. [En línea]. Available: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000549. [Último acceso: 24 Marzo 2023].

[7] D. L. D. C. Graham Thrall, «National Library of Medicine (USA),» 5 Mayo 2006. [En línea]. Available: https://pubmed.ncbi.nlm.nih.gov/16651058/. [Último acceso: 19 Febrero 2023].

[8] National Heart, Lung and Blood Institute, «NIH,» 24 Marzo 2022. [En línea]. Available: https://www.nhlbi.nih.gov/health/arrhythmias. [Último acceso: 24 Mayo 2023].

[9] S. L. R. X. S. David W. Hosmer Jr., Applied Logistic Regression, Hoboken, NJ: John Wiley and Sons, 2013.

[10] L. Breima, Random forests. Machine Learning, 2001.

[11] C. V. Cortes, Support Vector Networks, Springer, 1995.

[12] E. Alpaydin, Introduction to machine learning (2nd Ed.), Cambridge: MIT Press, 2010.

[13] A. Geron,). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.), O'Reilly Media, 2019.

[14] S. G. Andreas C. M¨uller, Introduction to Machine Learning with Python: A Guide, O'Reilly, 2016.