GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Interpretación de electrocardiogramas con técnicas de aprendizaje automático aplicadas a la detección de patologías cardíacas

Autor
Ana Arregui Beltrán

Director
Lucas Novales Peleato

Madrid
Junio 2026

Declaración de originalidad

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título **Interpretación de electrocardiogramas con técnicas de aprendizaje automático aplicadas a la detección de patologías cardíacas** en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Uso de Inteligencia Artificial[1]

Declaro bajo mi responsabilidad que (indicar la opción correcta):

☐ No he utilizado Inteligencia Artificial en la elaboración del presente documento.

☒ He utilizado Inteligencia Artificial en la elaboración del presente documento y/o del Anexo B siempre en las condiciones permitidas por la Universidad Pontificia Comillas, es decir, aplicando el Nivel 2 de la Escala de Evaluación de Perkins et al. (2024): *"La IA puede utilizarse para actividades previas a la tarea, como la lluvia de ideas, la descripción y la investigación inicial. Este nivel se centra en el uso de la IA para la planificación, las síntesis y la generación de ideas, pero las evaluaciones deben hacer hincapié en la capacidad de desarrollar y refinar estas ideas de forma independiente".* En concreto, la Inteligencia Artificial ha sido empleada para:

La inteligencia artificial se ha empleado como apoyo en la generación de ideas iniciales, en el uso de herramientas del proyecto y en la búsqueda de información, la cual ha sido siempre contrastada previamente a su utilización.

[1]Esta declaración se refiere al uso de la Inteligencia Artificial generativa para realizar los documentos del Proyecto (Anexo B y Memoria). No aplica a Proyectos donde, por su naturaleza, deban emplear inteligencia artificial como parte de los mismos (aplicación de técnicas de aprendizaje automático, redes neuronales, análisis de datos...).


Fdo: Ana Arregui Beltrán
Fecha: 21/06/2026

Autorización para la entrega del Proyecto

El Director del Proyecto	El co-Director del Proyecto (si aplica)
NOVALES PELEATO LUCAS FRANCISCO - 50565119X (firma aquí) Firmado digitalmente por NOVALES PELEATO LUCAS FRANCISCO - 50565119X Fecha: 2026.06.22 22:46:32 +02'00'	(firma aquí)
Fdo: Lucas Novales Peleato	Fdo:
Fecha:	

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Interpretación de electrocardiogramas con técnicas de aprendizaje automático aplicadas a la detección de patologías cardíacas

Autor
Ana Arregui Beltrán

Director
Lucas Novales Peleato

Madrid
Junio 2026

Agradecimientos

Este Trabajo Fin de Grado supone el cierre de una etapa muy importante de mi vida, y no habría sido posible sin el cariño y el apoyo de muchas personas que me han acompañado a lo largo de este camino.

Quiero agradecer a mi tutor del TFG, Lucas, por su guía, su paciencia y por haberme ayudado a desarrollar este proyecto. Gracias por su tiempo y disponibilidad incluso a 5000 km de distancia.

A mis padres, por su esfuerzo, su apoyo incondicional y por haberme dado la oportunidad de estudiar esta carrera. Gracias por confiar en mí, incluso cuando yo misma dudaba, y por haber hecho posible que pudiera llegar hasta aquí.

A mis hermanos, por acompañarme siempre, por su apoyo constante y por hacer que este camino haya sido más ligero y divertido.

A mis amigas de toda la carrera, Cris, Eva y Paula, por estos años compartidos que han hecho el camino más fácil y divertido. Gracias por los laboratorios, las horas de estudio, los exámenes, las risas y también los momentos de agobio, que siempre se llevaron mejor con vosotras.

A mis amigas de Boston, Alicia, Fernanda, Laura y Mencía, por una de las experiencias más especiales de mi etapa universitaria. Gracias por convertir una ciudad desconocida en un hogar, por todas las horas compartidas en la cuarta planta del CDS con vistas al M.I.T., por los *Hot Chocolate Deluxe*, las comentadas y por todas las tardes en las que tocaba *tfgear*.

Por último, gracias a todas las personas que, de una forma u otra, han formado parte de esta etapa y han contribuido a que llegue hasta aquí. Este trabajo marca el final de unos años de aprendizaje, esfuerzo y crecimiento que recordaré siempre con mucho cariño.

Interpretación de electrocardiogramas con técnicas de aprendizaje automático aplicadas a la detección de patologías cardíacas

Autor: Arregui Beltrán, Ana
Director: Novales Peleato, Lucas
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

Resumen del proyecto

Este Trabajo Fin de Grado desarrolla un sistema de detección automática de patologías cardíacas mediante aprendizaje automático aplicado a señales ECG de la base de datos PTB-XL. Se evalúan distintos modelos en tres configuraciones experimentales. Los resultados muestran que el uso de 12 derivaciones y la extracción de características mejora el rendimiento, alcanzando hasta 0,93 de AUC en superclases.

Palabras clave: Electrocardiograma, Aprendizaje automático, Clasificación multilabel, Patologías cardíacas, Procesamiento de señales

1. Introducción

Las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad a nivel mundial, por lo que su detección temprana resulta fundamental para mejorar el pronóstico clínico. En este contexto, el electrocardiograma (ECG) es una de las herramientas diagnósticas más utilizadas por su carácter no invasivo, bajo coste y amplia disponibilidad. Sin embargo, su interpretación requiere experiencia clínica y puede verse limitada por el creciente volumen de datos generado en entornos asistenciales. Ante esta situación, el aprendizaje automático surge como una alternativa prometedora para desarrollar sistemas de apoyo al diagnóstico capaces de detectar patrones complejos en señales ECG de forma automática, rápida y consistente.

2. Definición del proyecto

El proyecto utiliza la base de datos PTB-XL para el análisis automático de señales de electrocardiograma mediante técnicas de aprendizaje

automático. Se definen tres configuraciones experimentales de complejidad creciente, que van desde una única derivación con clasificación binaria hasta el uso de las 12 derivaciones con estimación probabilística de 71 patologías.

Se evalúan distintos modelos de aprendizaje automático con el objetivo de comparar su rendimiento y analizar la influencia del tipo de representación de la señal y del número de derivaciones en la capacidad de predicción del sistema.

3. Descripción del modelo/sistema/herramienta

El sistema se basa en un flujo de procesamiento de señales ECG que incluye preprocesado, segmentación de latidos y extracción de características relevantes de tipo temporal, morfológico, estadístico y frecuencial. A partir de estas representaciones, se entrenan y evalúan distintos modelos de aprendizaje automático según la configuración experimental.

El sistema permite abordar desde una clasificación binaria de superclases hasta una estimación probabilística de 71 patologías. Además, se integra un dashboard interactivo que facilita la visualización de las señales, la segmentación de latidos, las predicciones del modelo y la generación de informes clínicos.

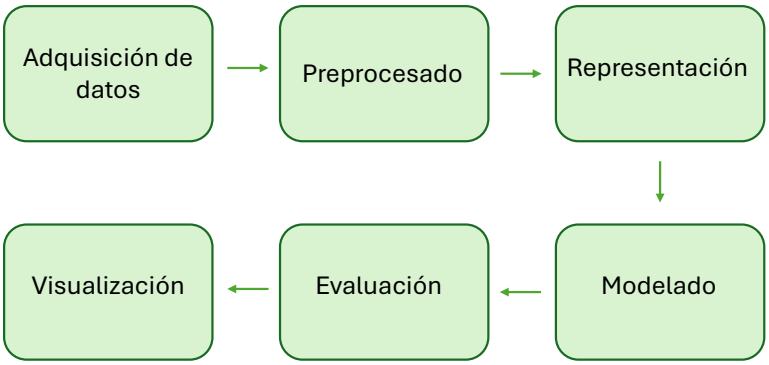

Figura 1: Flujo a seguir del proyecto

4. Resultados

En la primera configuración, la red neuronal multicapa obtiene el mejor rendimiento global, mientras que Random Forest presenta el comportamiento más equilibrado. En la segunda configuración, el uso de las 12 derivaciones mejora de forma consistente todas las métricas, destacando de nuevo la MLP como el modelo más competitivo.

En la tercera configuración se alcanzan los mejores resultados globales del sistema, con AUC de hasta 0,93 en superclases y 0,90 en la estimación probabilística de 71 patologías. En este escenario, la MLP y Random Forest destacan como los modelos más robustos, mientras que la extracción de características mejora de forma consistente el rendimiento frente al uso de la señal en bruto.

En la Tabla 1 se presentan los resultados globales obtenidos en la Configuración 3. Estos resultados se muestran debido a que se trata del escenario más completo del estudio y el que más se aproxima a un entorno clínico real.

Modelo	**MAE**	**MSE**	**Log-loss**	**AUC global**
Ridge Regression	0.0253	0.0097	0.0392	0.8808
ElasticNet Regression	0.0242	0.0098	0.0373	0.8887
Random Forest Regressor	0.0229	0.0098	0.0358	0.9002
HistGradientBoosting	0.0188	0.0085	0.0330	0.8509
MLP Regressor	0.0140	0.0112	0.0558	0.8835

Tabla 1: Resultados globales en test de la configuración 3

5. Conclusiones

El estudio demuestra que la detección automática de patologías cardíacas mediante aprendizaje automático es viable y ofrece un rendimiento elevado sobre señales ECG. El uso de las 12 derivaciones y la extracción de características mejora de forma consistente los resultados frente a representaciones más simples.

Los modelos no lineales, especialmente Random Forest, HistGradientBoosting y la red neuronal multicapa, superan a los enfoques

lineales. La MLP obtiene el mejor rendimiento global, mientras que Random Forest muestra un comportamiento más estable desde una perspectiva clínica, especialmente en patologías de mayor riesgo.

6. Referencias

Ayano, Y. M., Schwenker, F., Dufera, B. D., & Debelee, T. G. (2022). Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review [Accedido: 2026-04-17]. *Diagnostics (Basel)*, *13*(1), 111. https://doi.org/10.3390/diagnostics13010111

European Society of Cardiology. (2024). Artificial intelligence in ECG diagnostics - where are we now? [Accedido: 2026-04-17]. https://www.escardio.org/communities/councils/cardiology-practice/education/cardiopractice/artificial-intelligence-in-ecg-diagnostics-where-are-we-now/

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network [Accedido: 2026-04-17]. *Nature Medicine*, *25*(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

Kalmady, S. V., Salimi, A., Sun, W., Sepehrvand, N., Nademi, Y., Bainey, K., Ezekowitz, J., Hindle, A., McAlister, F., Greiner, R., Sandhu, R., & Kaul, P. (2024). Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level [Accedido: 2026-04-17]. *npj Digital Medicine*, *7*, 133. https://doi.org/10.1038/s41746-024-01133-?

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45

Interpretation of electrocardiograms using Machine Learning techniques applied to the detection of cardiac pathologies

Author: Arregui Beltrán, Ana
Supervisor: Novales Peleato, Lucas
Collaborating entity: ICAI – Universidad Pontificia Comillas

Abstract

This Bachelor's Thesis develops an automatic system for detecting cardiac pathologies using machine learning applied to ECG signals from the PTB-XL database. Different models are evaluated across three experimental configurations. The results show that using 12 leads and feature extraction significantly improves performance, achieving up to 0.93 AUC in superclass classification.

Key words: Electrocardiogram, Machine learning, Multilabel classification, Cardiac pathologies, Signal processing

1. Introduction

Cardiovascular diseases are one of the leading causes of mortality worldwide, making early detection essential to improve clinical outcomes. In this context, the electrocardiogram (ECG) is one of the most widely used diagnostic tools due to its non-invasive nature, low cost, and wide availability. However, its interpretation requires clinical expertise and is increasingly limited by the growing volume of data generated in healthcare environments. In this scenario, machine learning emerges as a promising approach to develop decision-support systems capable of detecting complex patterns in ECG signals in an automatic, fast, and consistent way.

2. Project description

The project uses the PTB-XL database for automatic analysis of ECG signals using machine learning techniques. Three experimental

configurations of increasing complexity are defined, ranging from a single lead with binary classification to the use of 12 leads with probabilistic estimation of 71 pathologies.

Different machine learning models are evaluated in order to compare their performance and analyze the influence of signal representation and number of leads on the system's predictive capability.

3. Model/System/Tool description

The system is based on an ECG signal processing pipeline that includes preprocessing, heartbeat segmentation, and feature extraction of temporal, morphological, statistical, and frequency-based characteristics. Based on these representations, different machine learning models are trained and evaluated according to each experimental configuration.

The system covers both binary superclass classification and probabilistic estimation of 71 pathologies. In addition, an interactive dashboard is integrated to visualize ECG signals, segmented heartbeats, model predictions, and to generate clinical reports.

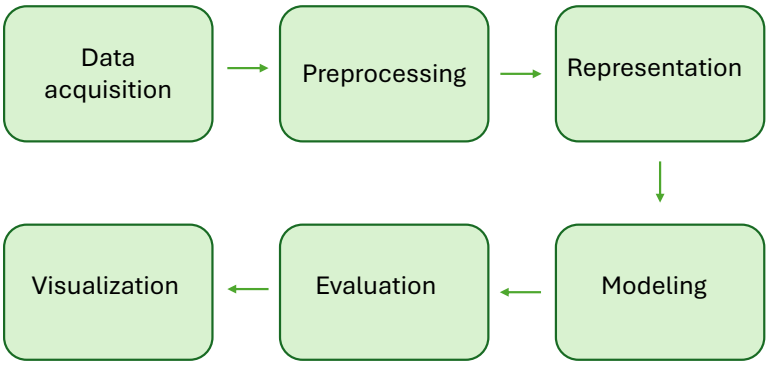

Figura 2: Project's pipeline

4. Results

In the first configuration, the multilayer perceptron achieves the best overall performance, while Random Forest shows the most ba-

lanced behavior. In the second configuration, the use of 12 leads consistently improves all metrics, with the MLP again being the most competitive model.

In the third configuration, the best overall results are obtained, reaching up to 0.93 AUC in superclass classification and 0.90 in probabilistic estimation of 71 pathologies. In this scenario, MLP and Random Forest stand out as the most robust models, while feature extraction consistently improves performance compared to raw signal representation.

Table 2 presents the overall results obtained in Configuration 3. These results are included because this configuration corresponds to the most complete scenario of the study and the one that most closely resembles a real clinical setting.

Model	**MAE**	**MSE**	**Log-loss**	**AUC global**
Ridge Regression	0.0253	0.0097	0.0392	0.8808
ElasticNet Regression	0.0242	0.0098	0.0373	0.8887
Random Forest Regressor	0.0229	0.0098	0.0358	0.9002
HistGradientBoosting	0.0188	0.0085	0.0330	0.8509
MLP Regressor	0.0140	0.0112	0.0558	0.8835

Tabla 2: Global test results for configuration 3

5. Conclusions

The study shows that automatic detection of cardiac pathologies using machine learning is feasible and achieves strong performance on ECG signals. The use of 12 leads and feature extraction consistently improves results compared to simpler representations.

Non-linear models, particularly Random Forest, HistGradientBoosting, and the multilayer perceptron, outperform linear approaches. The MLP achieves the best overall performance, while Random Forest shows more stable behavior from a clinical perspective, especially in high-risk pathologies.

6. References

Ayano, Y. M., Schwenker, F., Dufera, B. D., & Debelee, T. G. (2022). Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review [Accedido: 2026-04-17]. *Diagnostics (Basel)*, *13*(1), 111. https://doi.org/10.3390/diagnostics13010111

European Society of Cardiology. (2024). Artificial intelligence in ECG diagnostics - where are we now? [Accedido: 2026-04-17]. https://www.escardio.org/communities/councils/cardiology-practice/education/cardiopractice/artificial-intelligence-in-ecg-diagnostics-where-are-we-now/

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network [Accedido: 2026-04-17]. *Nature Medicine*, *25*(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

Kalmady, S. V., Salimi, A., Sun, W., Sepehrvand, N., Nademi, Y., Bainey, K., Ezekowitz, J., Hindle, A., McAlister, F., Greiner, R., Sandhu, R., & Kaul, P. (2024). Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level [Accedido: 2026-04-17]. *npj Digital Medicine*, *7*, 133. https://doi.org/10.1038/s41746-024-01133-?

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto del problema

Las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad a nivel mundial, lo que las convierte en un problema sanitario de gran relevancia. Su detección temprana es fundamental para reducir complicaciones y mejorar el pronóstico de los pacientes. En este contexto, el electrocardiograma (ECG) representa una de las herramientas más utilizadas en la práctica clínica para la evaluación de la actividad eléctrica del corazón.

El ECG es una prueba no invasiva, de bajo coste y de fácil acceso, por lo que es una técnica clave en el diagnóstico cardiovascular. Sin embargo, su interpretación requiere formación especializada y experiencia clínica, ya que la señal puede presentar patrones complejos, ruido y variabilidad entre pacientes. Por todo esto y por el incremento del volumen de datos en entornos hospitalarios y sistemas de monitorización continua surge la necesidad de herramientas automáticas que faciliten el análisis de la información de forma eficiente y consistente.

1.2. Motivación

El creciente interés por la aplicación de técnicas de aprendizaje automático en el ámbito médico se debe a su capacidad para analizar

grandes volúmenes de datos y detectar patrones complejos que pueden no ser evidentes mediante métodos tradicionales. En el caso del ECG, esto abre la puerta a sistemas de apoyo al diagnóstico que pueden complementar el trabajo del especialista.

Este trabajo se motiva tanto por su relevancia clínica como por su interés tecnológico. Desde el punto de vista médico, la detección temprana de patologías cardíacas puede tener un impacto directo en la reducción de la mortalidad. Desde el punto de vista técnico, el problema permite aplicar técnicas avanzadas de procesamiento de señales y modelos de aprendizaje automático en un entorno real y altamente complejo.

El objetivo no es sustituir el diagnóstico médico, sino desarrollar herramientas que ayuden a interpretar los registros ECG, proporcionando información adicional que pueda facilitar la toma de decisiones clínicas.

1.3. Problema a resolver

El problema abordado en este Trabajo Fin de Grado consiste en el desarrollo de un sistema automático capaz de interpretar señales de electrocardiograma y estimar la presencia de distintas patologías cardíacas mediante técnicas de aprendizaje automático.

Para ello, se trabaja con registros ECG multiderivación y se plantean distintos niveles de complejidad del problema, desde la clasificación en superclases clínicas hasta la estimación probabilística de un conjunto amplio de patologías. Esta progresión permite analizar cómo influye la representación de la señal y la formulación del problema en el rendimiento de los modelos.

El objetivo es evaluar la capacidad de distintos algoritmos para capturar patrones relevantes en la señal ECG, así como su comportamiento en escenarios con desbalance de clases y alta complejidad clínica.

1.4. Objetivos

El objetivo general es desarrollar y evaluar distintos modelos de aprendizaje automático para la detección de patologías cardíacas a partir de señales de electrocardiograma, analizando su capacidad de discriminación en distintos escenarios de complejidad.

De forma más concreta, se plantean los siguientes objetivos:

1. Diseñar un sistema de análisis automático de señales electrocardiográficas, capaz de procesar registros ECG multiderivación y extraer información relevante para su posterior modelado.

2. Evaluar distintos modelos de aprendizaje automático y comparar su rendimiento, incluyendo enfoques lineales, basados en ensembles y redes neuronales, con el fin de identificar sus ventajas y limitaciones en este problema.

3. Analizar el impacto de diferentes representaciones de la señal en la capacidad predictiva, estudiando cómo influyen el uso de la señal en bruto frente a características extraídas en el rendimiento de los modelos.

4. Estudiar el comportamiento de los modelos frente a distintas patologías y niveles de complejidad clínica, evaluando su capacidad de generalización en problemas con distinto grado de desbalance y solapamiento entre clases.

5. Desarrollar una herramienta que permita interpretar los resultados de forma accesible, facilitando la visualización de las predicciones y su comprensión desde un punto de vista analítico y clínico.

1.5. Alcance del proyecto

Este trabajo se centra en el desarrollo de un sistema de análisis automático de señales de electrocardiograma mediante técnicas de apren-

dizaje automático. El sistema está diseñado para procesar registros ECG multiderivación, aplicar distintas estrategias de preprocesamiento y entrenamiento, y evaluar la capacidad de diferentes modelos para la detección de patologías cardíacas.

En concreto, el proyecto incluye el desarrollo de varias configuraciones experimentales que permiten analizar el problema desde distintos niveles de complejidad, desde la clasificación en superclases clínicas hasta la predicción probabilística de múltiples patologías. Se realiza un estudio comparativo entre distintos modelos de aprendizaje automático, evaluando su comportamiento mediante métricas de clasificación y regresión.

También se contempla la extracción de características relevantes a partir de la señal ECG y la utilización de estas junto con la señal original, con el objetivo de mejorar la capacidad de representación del problema y analizar su impacto en el rendimiento de los modelos.

Por otro lado, el sistema desarrollado incluye una herramienta de visualización que permite interpretar los resultados obtenidos, facilitando el análisis de los registros ECG y las predicciones del modelo.

El modelo debe entenderse como una herramienta de apoyo al análisis y de estimación preliminar de posibles patologías, basada en patrones aprendidos a partir de los datos. En este sentido, los resultados obtenidos pueden ser útiles para orientar la interpretación de los registros ECG, pero no deben considerarse un diagnóstico definitivo.

Asimismo, el rendimiento de los modelos está condicionado por la distribución de clases y por la presencia de patologías con patrones electrocardiográficos solapados, lo que afecta especialmente a la capacidad de discriminación en las clases más complejas o minoritarias.

1.6. Estructura de la memoria

La memoria se organiza en ocho capítulos que recogen de forma progresiva el desarrollo del Trabajo Fin de Grado, desde la base teórica hasta la implementación y el análisis de resultados.

Después de la introducción, en el Capítulo 2, se presentan los fundamentos teóricos y tecnológicos necesarios para comprender el trabajo. Se incluyen conceptos de electrocardiografía, estructura de la señal ECG, patologías relevantes y nociones básicas de aprendizaje automático y evaluación de modelos.

El Capítulo 3 recoge el estado del arte, donde se revisan diferentes enfoques utilizados previamente para la interpretación automática del ECG. Se analizan tanto métodos clásicos como técnicas basadas en aprendizaje automático, así como sus principales limitaciones.

El Capítulo 4 define el problema planteado y describe la base de datos utilizada, incluyendo su estructura, características principales y limitaciones. También se presentan los distintos escenarios experimentales diseñados para evaluar el sistema.

El Capítulo 5 detalla el desarrollo del modelo, incluyendo las etapas de preprocesamiento de la señal, extracción y selección de características, así como los distintos algoritmos de aprendizaje automático empleados en el estudio.

El Capítulo 6 presenta el análisis de resultados, donde se evalúa el rendimiento de los distintos modelos en cada configuración experimental. Se utilizan métricas de evaluación y se analiza su comportamiento frente a las distintas patologías.

El Capítulo 7 describe el desarrollo del dashboard interactivo, diseñado como herramienta de visualización que permite analizar señales ECG y explorar las predicciones generadas por el modelo de forma intuitiva.

El Capítulo 8 recoge las conclusiones del trabajo, sintetizando los resultados obtenidos y proponiendo posibles líneas de mejora y extensión futura del sistema.

Capítulo 2

Fundamentos Teóricos y Tecnologías

2.1. Anatomía eléctrica del corazón

El corazón es un órgano muscular que funciona como una bomba para impulsar la sangre a través del sistema circulatorio. Esta acción de bombeo está regulada por un sistema eléctrico, encargado de coordinar las contracciones de las distintas cavidades del corazón.

El latido se genera en la aurícula derecha, donde se encuentra el nodo sinoauricular (SA) que actúa como marcapasos natural del corazón. El nodo SA envía pulsos eléctricos que se propagan a través de las aurículas derecha e izquierda, llegando al nodo auriculoventricular (AV). En este nodo, el impulso se frena brevemente, permitiendo que las aurículas se contraigan antes que los ventrículos. Así, la sangre se vacía de las aurículas a los ventrículos antes de su contracción.

Después de pasar por el nodo AV, la corriente eléctrica desciende por el Haz de His hasta los ventrículos. Esta vía se divide en dos ramas, derecha e izquierda, conocidas como ramas de His. Estas ramas estimulan eléctricamente los ventrículos, provocando su contracción y el bombeo de la sangre hacia el exterior del corazón. (Stanford Children's Health, s.f.-a)

La anatomía eléctrica del corazón se muestra en la Figura 2.1

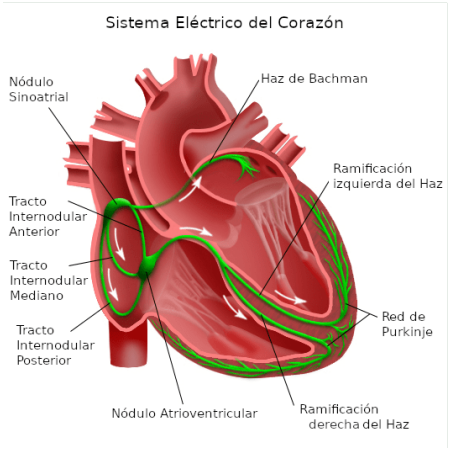

Figura 2.1: Anatomía eléctrica del corazón. (Cigna, s.f.)

2.2. Electrocardiograma

2.2.1. ¿Qué es un ECG?

El electrocardiograma (ECG) es la representación gráfica de la actividad eléctrica del corazón. Se trata de una prueba no invasiva, de bajo coste y fácil de realizar. Esta representación se utiliza para buscar actividad cardíaca anormal, que pueda indicar la presencia de patologías cardíacas.

No solo se utiliza para diagnosticar enfermedades, sino que el ECG puede evaluar el ritmo cardíaco, si este es regular o irregular y la fuerza y sincronización de las señales eléctricas. Además, se puede utilizar para medir el tamaño y la posición de las cavidades del corazón. (MedlinePlus, s.f.)

El ECG es una herramienta fundamental en la práctica clínica, ya que permite la interpretación del ritmo cardíaco y la detección de alteraciones del sistema de conducción eléctrico. Aporta información relevante en la evaluación de patologías cardiovasculares, como valvulopatías, miocardiopatías, pericarditis, isquemia miocárdica o enfermedades hipertensivas. Por otro lado, el ECG se puede utilizar para monitorizar la respuesta a tratamientos farmacológicos, especialmente antiarritmi-

cos, así como en la detección de alteraciones metabólicas. (Prutkin et al., 2025)

2.2.2. Sistema de derivaciones del ECG

Para medir la actividad eléctrica del corazón, se colocan electrodos en cada una de las extremidades y en el pecho, formando un sistema de 12 derivaciones. Cada derivación mide la diferencia de potencial eléctrico entre dos puntos del cuerpo o un punto del cuerpo y una referencia. El electrodo conectado en la pierna derecha sirve como electrodo de referencia a tierra. (Klabunde, Richard E., 2023)

Existen dos tipos de derivaciones: las derivaciones bipolares y las derivaciones unipolares. Las derivaciones bipolares utilizan un electrodo positivo y otro negativo para medir la diferencia de potencial, mientras que las derivaciones unipolares utilizan un único electrodo positivo y una combinación de otros electrodos como electrodo de referencia.

Las doce derivaciones se agrupan en tres categorías según su ubicación: las derivaciones de las extremidades (I, II, III), las derivaciones aumentadas de las extremidades (aVR, aVL, aVF) y las derivaciones precordiales (V1-V6).

- **Derivaciones de las extremidades**

  Son las únicas derivaciones bipolares y se colocan en brazos y piernas. Miden la actividad eléctrica del corazón en el plano frontal.

  - Derivación I: Tiene el electrodo positivo en el brazo izquierdo y el negativo en el brazo derecho, midiendo la diferencia de potencial entre ambos brazos.
  - Derivación II: Tiene el electrodo positivo en la pierna izquierda y el negativo en el brazo derecho, midiendo la diferencia de potencial entre ambos puntos.
  - Derivación III: Tiene el electrodo positivo en la pierna izquierda y el negativo en el brazo izquierdo, midiendo la diferencia de potencial entre ambos puntos.

Estas tres derivaciones forman un triángulo equilátero, conocido como el *triángulo de Einthoven*, mostrado en la Figura 2.2, que representa la relación entre las derivaciones de las extremidades. Esta relación es tal que la suma de los impulsos eléctricos registrados en las derivaciones I y III es equivalente a la suma de los impulsos registrados en la derivación II:

$$DII = DI + DIII$$

Esta ley se conoce como *Ley de Einthoven* y constituye una base fundamental para la interpretación vectorial de la actividad eléctrica cardíaca y permite verificar la coherencia de las señales registradas. (My-EKG, s.f.-a)

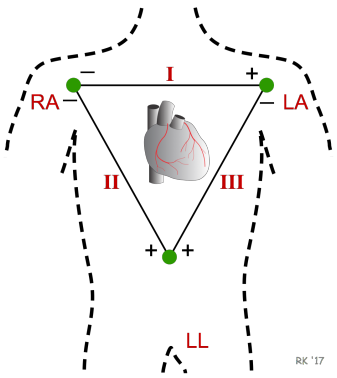

Figura 2.2: Triángulo de Einthoven.(Klabunde, s.f.-c)

- **Derivaciones aumentadas de las extremidades**

  Se trata de derivaciones unipolares, que utilizan como referencia una combinación de los electrodos de las extremidades. Al igual que estas, analizan la actividad eléctrica del corazón en el plano frontal.

  - Derivación aVR: El electrodo positivo está en el brazo derecho y mide la diferencia de potencial entre el brazo derecho y el centro del corazón.

- Derivación aVL: El electrodo positivo está en el brazo izquierdo y mide la diferencia de potencial entre el brazo izquierdo y el centro del corazón.
- Derivación aVF: El electrodo positivo está en la pierna izquierda y mide la diferencia de potencial entre la pierna izquierda y el centro del corazón.

Estas derivaciones complementan a las bipolares y permiten una visión más completa del plano frontal del corazón.

Sistema hexaxial y eje del corazón

Si el triángulo de Einthoven se despliega y se superpone sobre el centro cardíaco, se puede definir un sistema de referencia angular que permite definir la dirección de la actividad eléctrica del plano frontal.

Como se observa en la Figura 2.3 cada derivación bipolar tiene una orientación espacial:

- La derivación I se sitúa a 0^{o}, correspondiente al eje horizontal entre el brazo derecho y el brazo izquierdo.
- La derivación II tiene una orientación aproximada de $+60^{o}$ con respecto a este eje horizontal, entre el brazo derecho y la pierna izquierda.
- La derivación III se orienta a $+120^{o}$, entre el brazo izquierdo y la pierna izquierda.

Estas direcciones permiten interpretar la actividad eléctrica cardíaca como la proyección de un vector sobre distintos ejes espaciales. De este modo, el electrocardiograma no solo representa amplitudes eléctricas, sino también información direccional de la despolarización cardíaca.

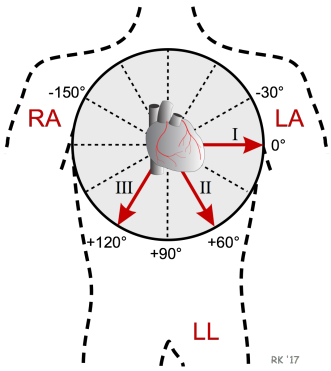

Figura 2.3: Eje de las derivaciones bipolares en el plano frontal. (Klabunde, s.f.-c)

A partir de este sistema de derivaciones se incorporan las derivaciones aumentadas de las extremidades, que completan la cobertura del plano frontal y permiten una explicación más precisa del vector eléctrico. (Figura 2.4) Estas derivaciones se sitúan en los siguiente ángulos:

- Derivación aVL: -30º sobre la horizontal
- Derivación aVR: -150º sobre el eje horizontal
- Derivación aVF: +90º sobre el plano horizontal

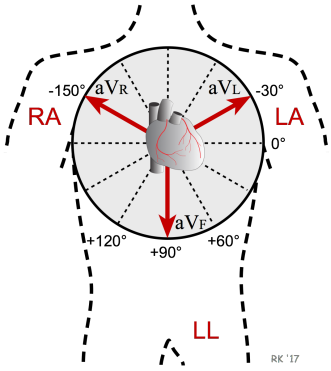

Figura 2.4: Eje de las derivaciones aumentadas de las extremidades en el plano frontal. (Klabunde, s.f.-a)

El conjunto de estas seis derivaciones constituye el sistema hexaxial completo,mostrado en la Figura 2.5, que permite analizar la

dirección del vector eléctrico cardíaco en el plano frontal y constituye la base para la determinación del eje eléctrico del corazón.

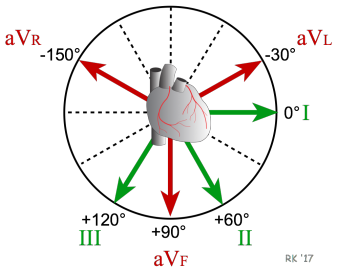

Figura 2.5: Sistema hexaxial. (Klabunde, s.f.-a)

Este eje se determina a partir del complejo QRS (ver apartado 2.2.3), que representa la suma vectorial de todos los vectores eléctricos generados por la despolarización ventricular. Analizando la polaridad del QRS en las derivaciones del plano frontal, es posible determinar la dirección estimada de la actividad eléctrica cardíaca. En la práctica clínica, las derivaciones I y aVF son las más utilizadas para hacer una estimación rápida del eje eléctrico.

La determinación del eje se basa en la polaridad del complejo QRS en estas derivaciones. Cuando el vector de despolarización se dirige hacia el polo positivo de una derivación, la deflexión del complejo QRS es positiva; por el contrario, si el vector se aleja, la deflexión es negativa. («Eje cardíaco en el ECG: cálculo e interpretación», 2026)

El eje cardíaco en condiciones normales se encuentra entre -30º y 90º, haciendo que el vector resultante apunte abajo y a la izquierda. En este caso, el complejo QRS es positivo tanto en DI como en aVF.

En la tabla 2.1 se resume la interpretación del eje cardíaco en función de la polaridad del complejo QRS en las derivaciones I y aVF.

DI	aVF	Cuadrante	Rango del eje	Interpretación
+	+	Inferior izquierdo	$-30°$ a $+90°$	Eje normal
+	-	Superior izquierdo	$-30°$ a $-90°$	Desviación izquierda del eje
-	+	Inferior derecho	$+90°$ a $+180°$	Desviación derecha del eje
-	+	Superior derecho	$-90°$ a $-180°$	Eje extremo derecho o izquierdo

Tabla 2.1: Interpretación del eje cardíaco según las derivaciones DI y aVF (Prutkin et al., 2025)

- **Derivaciones precordiales**

  Estas seis derivaciones unipolares se colocan en el pecho, en puntos bien definidos que permiten valorar diferentes regiones del corazón, y miden la actividad eléctrica en un plano perpendicular al plano frontal. Cada una de las derivaciones precordiales observa una región específica del ventrículo. En la Figura 2.6 se indica que parte del corazón observa cada derivación.

  - Derivaciones V1-V2: Exploran la region septal y anteroseptal del corazón, observando principalmente el septo interventricular.
  - Derivaciones V3-V4: Exploran la región anterior del corazón y el apex cardíaco, observando principalmente la pared anterior del ventrículo izquierdo.
  - Derivaciones V5-V6: Exploran en la región lateral y anterolateral izquierda del corazón, observando principalmente la pared lateral del ventrículo izquierdo.

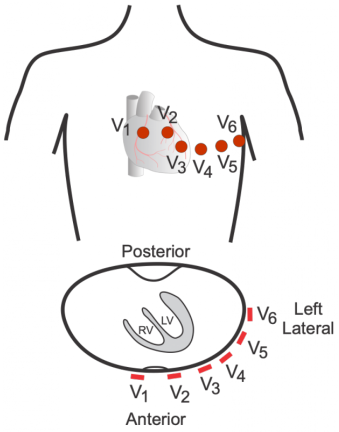

Figura 2.6: Derivaciones precordiales. (Klabunde, s.f.-b)

2.2.3. Partes del ECG

Un ciclo cardíaco típico en el ECG está compuesto por varias ondas y segmentos característicos, cada uno de los cuales se asocia a una fase concreta del proceso eléctrico del corazón. En la Figura 2.7 se muestra un esquema completo de un ciclo cardíaco típico. (Prutkin et al., 2025)

Onda P

La onda P representa la despolarización de las aurículas. Morfológicamente, presenta una onda suave y amplitud baja, siendo positiva en la mayoría de las derivaciones, salvo en aVR.

La aurícula derecha se despolariza primero, seguida de la aurícula izquierda, por lo que esta onda puede presentar una muesca en algunas derivaciones de las extremidades, o ser bifásica en V1. En esta derivación, la parte inicial de la onda P corresponde a la despolarización de la aurícula derecha, mientras que la parte final corresponde a la aurícula izquierda.

La repolarización auricular ocurre simultáneamente a la despolarización ventricular, por que lo que suele quedar oculta por el complejo QRS, por lo que no suele ser visible.

Intervalo PR

El intervalo PR representa el tiempo total que transcurre desde el inicio de la despolarización auricular hasta el incio de la despolarización ventricular. Incluye tanto la onda P como el segmento PR.

Se mide desde el inicio de la onda P hasta el inicio del complejo QRS. En condiciones normales, su duración puede variar en función de la frecuencia cardíaca, ya que a frecuencias elevadas se produce una aceleración de la conducción a través del nodo AV, lo que reduce el intervalo PR. Por el contrario, a frecuencias bajas la conducción se enlentece, alargando el intervalo PR.

Una alteración en la duración del intervalo PR puede indicar bloqueos o cambios en la velocidad de conducción del impulso eléctrico. Un intervalo PR prolongado puede indicar un bloqueo AV de primer grado, y un intervalo PR corto puede indicar la existencia de una vía de conducción adicional, como en el síndrome de Wolff-Parkinson-White.

Complejo QRS

El complejo QRS representa la despolarización ventricular, que es el proceso eléctrico más potente del corazón. Su análisis es fundamental para la interpretación del ECG, ya que proporciona información sobre la conducción eléctrica ventricular, el tamaño y la posición de los ventrículos, así como la presencia de bloqueos o alteraciones en la conducción.

Es una onda de corta duración, lo que indica que la despolarización ventricular es un proceso rápido. Una larga duración del QRS puede indicar alteraciones en la conducción, como bloqueos de rama o ritmos originados fuera del sistema de conducción normal.

El complejo QRS está compuesto por las siguientes deflexiones:

- **Onda Q**: Es la primera deflexión negativa del complejo QRS, que representa la despolarización del septo interventricular. En condiciones normales, esta onda es pequeña y estrecha, y puede no ser visible en todas las derivaciones.

- **Onda R**: Es la primera deflexión positiva del complejo QRS, que representa la despolarización del ventrículo izquierdo. Es la onda más alta del complejo QRS y su amplitud varía según la derivación.
- **Onda S**: Es la segunda deflexión negativa del complejo QRS, que representa las fases finales de la despolarización ventricular.
- **Onda R'**: Es una segunda deflexión positiva que puede aparecer después de la onda S.

Segmento ST

El segmento ST es la porción del ECG que se encuentra entre el final del complejo QRS y el inicio de la onda T. Representa el período en el que los ventrículos están completamente despolarizados, es decir, el tiempo entre la despolarización y la repolarización ventricular.

El punto de unión entre el complejo QRS y el segmento ST se conoce como el punto J, no mostrado en la Figura 2.7. En condiciones normales, el segmento ST es isoeléctrico, es decir, se encuentra al mismo nivel que la línea de base del ECG. Sin embargo, en ciertas patologías, como la isquemia miocárdica o el infarto de miocardio, el segmento ST puede presentar elevación o depresión.

En condiciones normales también pueden observarse ligeras variaciones fisiológicas del segmento ST, especialmente durante episodios de taquicardia sinusal, donde el punto J puede encontrarse ligeramente deprimido acompañado de un ascenso rápido del segmento ST hasta recuperar la línea isoeléctrica.

Onda T

La onda T representa la repolarización ventricular, que es el proceso eléctrico mediante el cual los ventrículos recuperan su estado de reposo después de la contracción. Es una onda de baja amplitud y duración más larga que el complejo QRS.

En condiciones normales, la onda T suele ser asimétrica, con una ascensión más lenta y una bajada más rápida hacia la línea isoeléctrica. Suele ser una onda de amplitud baja y una forma suave. La presencia de muescas, irregularidades o deformaciones puede sugerir la superposición de otras ondas, como la onda P.

Intervalo QT

El intervalo QT representa el tiempo total de actividad eléctrica ventricular. Incluye tanto la despolarización como la repolarización ventricular, desde el inicio del complejo QRS hasta el final de la onda T.

Depende de la frecuencia cardíaca, siendo más corto a frecuencias elevadas y más largo a frecuencias bajas.

Para poder compararlo entre distintos registros con diferentes frecuencias cardíacas, se utiliza el intervalo QT corregido (QTc), que ajusta su valor en función del intervalo RR.

La prolongación del intervalo QT puede asociarse a un mayor riesgo de arritmias ventriculares, por lo que constituye un parámetro de relevancia clínica en la interpretación del ECG.

Onda U

La onda U (no visible en la Figura 2.7) es una pequeña deflexión que puede aparecer en el electrocardiograma después de la onda T. Puede hacerse más evidente en ciertas situaciones, como la hipopotasemia o la bradicardia. La presencia de ondas U prominentes o alteraciones en su morfología puede estar asociada a cambios en la repolarización ventricular y, en algunos casos, a condiciones patológicas subyacentes

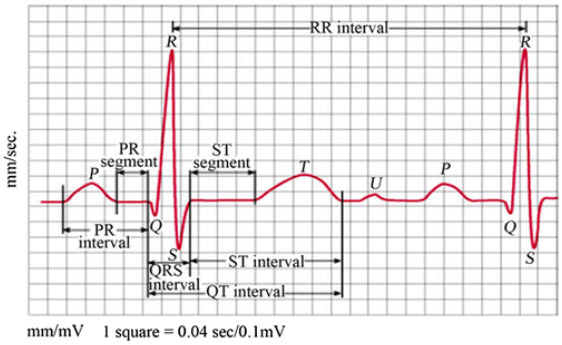

Figura 2.7: Esquema de un ciclo cardíaco típico del ECG. (Adetiba et al., 2017)

2.2.4. Alteraciones posibles en el ECG

Las alteraciones del electrocardiograma corresponden a desviaciones del patrón eléctrico normal, que puede indicar anomalías en la generación del impulso, su propagación o la repolarización del miocardio. Constituyen la base de la interpretación clínica del ECG, permitiendo identificar distintos tipos de patologías a partir del cambio en el ritmo, la conducción o la forma de la señal.

Alteraciones del ritmo cardíaco

Se producen cuando se modifica la generación normal del impulso eléctrico. En condiciones normales, el nodo SA actúa como marcapasos natural, estableciendo un ritmo regular.

Si este mecanismo falla, aparecen las arritmias, que pueden manifestarse como cambios en la frecuencia o en la regularidad del ritmo. Las más comunes son la bradicardia (frecuencia cardíaca baja), la taquicardia (frecuencia cardíaca alta) y ritmos irregulares, como la fibrilación auricular.

Alteraciones de la conducción eléctrica

Aparecen cuando existe un retraso o bloqueo en la propagación del impulso eléctrico a través del sistema de conducción cardíaco.

Pueden localizarse a nivel del nodo AV o del sistema His-Purkinje. Entre las principales se incluyen los bloqueos auriculoventriculares y los bloqueos de rama, que se caracterizan por una prolongación del complejo QRS y alteraciones en su morfología.

Alteraciones morfológicas del ECG

Correponden a cambios en la forma de las ondas, segmentos o intervalos del ECG.

Destacan las alteraciones del segmento ST, que puede aparecer elevado o deprimido y se asocia frecuentemente a procesos isquémicos. También es relevante la inversión de la onda T, relacionada con alteraciones en la repolarización ventricular, así como las modificaciones del intervalo QT, que pueden indicar cambios en la duración global de la actividad eléctrica ventricular.

2.3. Patologías estudiadas

Las alteraciones detectadas en el ECG pueden afectar a la generación, propagación o recuperación del impulso eléctrico, reflejándose en cambios característicos en la señal.

Entre las principales categorías clínicas relevantes en el análisis del ECG se encuentran distintas patologías que afectan a la morfología de las ondas, los intervalos y la conducción cardíaca.

A continuación, se describen las cinco clases en las que se agrupan las patologías estudiadas en este trabajo:

- Normal (NORM – *Normal ECG*)

  Esta clase son electrocardiogramas sin alteraciones y que representan el patrón fisiológico de referencia. La actividad eléctrica del corazón en estos casos sigue una secuencia normal de generación y propagación del impulso.

- Infarto de miocardio (MI – *Myocardial Infarction*)

Agrupa patologías relacionadas con el daño del tejido miocárdico. Se reflejan principalmente en cambios en el segmento ST y en la onda T, así como en modificaciones de la morfología general del electrocardiograma.

- Bloqueos de conducción (CD – *Conduction Disturbance*)

  Incluye alteraciones en la propagación del impulso eléctrico a través del sistema de conducción cardíaco. Pueden afectar a la sincronización entre aurículas y ventrículos o a la activación ventricular, reflejandose en cambios en los intervalos y en el complejo QRS.

- Alteraciones ST-T (STTC – *ST/T change*)

  Engloba alteraciones en la repolarización ventricular. Se manifiesta mediante cambios en el segmento ST y en la onda T y suele estar asociada a procesos isquémicos o alteraciones en la actividad eléctrica del miocardio.

- Hipertrofia (HYP – *Hypertrophy*)

  Incluye alteraciones derivadas de cambios estructurales del corazón, como el aumento del grosor de las paredes cardíacas. Estas modificaciones afectan a la propagación del impulso eléctrico y se reflejan en cambios en la amplitud y morfología de la señal ECG.

2.4. Machine Learning

"El Machine Learning es una rama de la inteligencia artificial que hace posible el aprendizaje autónomo de las máquinas, sin necesidad de ser programadas expresamente para cada tarea. Su objetivo principal es desarrollar algoritmos y modelos capaces de analizar grandes volúmenes de datos, detectar patrones y aprender de ellos." (Repsol, 2026)

Dentro del aprendizaje automático, el enfoque más común en problemas biomédicos es el aprendizaje supervisado, en el que el modelo

se entrena utilizando un conjunto de datos etiquetados. Así, el sistema aprende una relación entre las características de entrada y la salida esperada.

Estos problemas pueden ser de clasificación, donde el objetivo es asignar una etiqueta a cada muestra o de regresión, donde se predice un valor continuo. La clasificación puede ser binaria, cuando solo hay dos clases posibles, o multiclase, cuando existen más de dos clases. En el caso de la clasificación multilabel, cada muestra puede pertenecer a varias clases simultáneamente, como es el caso de este proyecto.

2.5. Clasificación multilabel

La clasificación multilabel es un problema de aprendizaje supervisado en el que cada muestra puede pertenecer a múltiples clases simultáneamente.

Este enfoque es especialmente relevante en el ámbito médico, ya que un mismo paciente puede presentar varias patologías al mismo tiempo. En el caso de los electrocardiogramas, es posible observar alteraciones simultáneas en una misma señal.

En este contexto, las etiquetas suelen representarse mediante vectores binarios, donde cada posición indica la presencia o ausencia de una clase específica. Este tipo de representación permite modelar de forma sencilla la existencia de múltiples patologías en un mismo registro.

2.6. Métricas de evaluación

La evaluación de modelos de clasificación se realiza mediante diferentes métricas que permiten cuantificar su rendimiento desde distintas perspectivas. En problemas médicos, estas métricas resultan especialmente relevantes, ya que no solo interesa la exactitud global del modelo, sino también su capacidad para detectar correctamente cada patología.

Matriz de confusión

Es un método de visualización para los resultados de clasificación. Es la base de la mayoría de métricas de evaluación. (Murel y Kavlakoglu, s.f.-b) Permite comparar las predicciones del modelo con las etiquetas reales y se define en términos de:

- **Verdaderos Positivos (TP)**: Número de muestras correctamente clasificadas como positivas.
- **Falsos Positivos (FP)**: Número de muestras incorrectamente clasificadas como positivas.
- **Verdaderos Negativos (TN)**: Número de muestras correctamente clasificadas como negativas.
- **Falsos Negativos (FN)**: Número de muestras incorrectamente clasificadas como negativas.

En la Figura 2.8 se muestra la estructura general de una matriz de confusión para un problema de clasificación binaria. En problemas multilabel, se puede construir una matriz de confusión para cada clase. Esta matriz es fundamental para calcular otras métricas de evaluación, como la precisión, la sensibilidad o el F1-score, que se definen a partir de los valores de TP, FP, TN y FN.

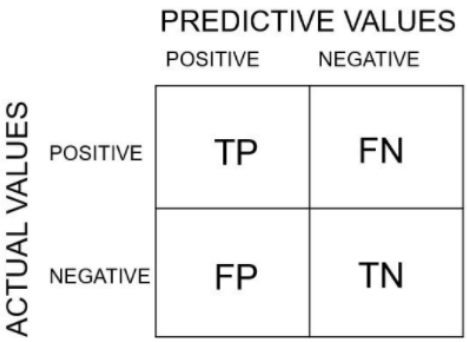

Figura 2.8: Matriz de confusión.

Accuracy

La exactitud o *accuracy* es una métrica de evaluación que mide la proporción de predicciones correctas realizadas por el modelo, respecto al total de muestras evaluadas. (Evidently AI, 2025) Se calcula utilizando la siguiente fórmula:

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

Esta métrica proporciona una visión global del rendimiento del modelo, indicando con qué frecuencia el sistema clasifica correctamente las muestras. Sin embargo, en problemas con clases desbalanceadas o en problemas multilabel, la accuracy puede ser poco representativa, ya que no refleja adecuadamente el rendimiento en cada clase. (Murel y Kavlakoglu, s.f.-b)

Precision

La precisión o *precision* es una métrica que mide la proporción de predicciones positivas que son correctas. Indica que porcentaje de los casos clasificados como positivos por el modelo realmente pertenecen a la clase positiva. Se calcula como:

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

Esta métrica permite evaluar la fiabilidad de las predicciones positivas del modelo. (Evidently AI, 2025) En un contexto médico, una baja precisión implica un aumento de los falsos positivos, lo que puede llevar a diagnósticos erróneos o a la realización de pruebas innecesarias. Sin embargo, en este tipo de problemas es preferible asumir un mayor número de falsos positivos antes que pasar por alto la presencia de una enfermedad grave, dado el mayor coste clínico asociado a un diagnóstico no detectado. En problemas multilabel, la precisión se puede calcular para cada clase de forma individual, lo que permite evaluar el rendimiento del modelo en cada patología específica.

Recall

El *recall* o sensibilidad es una métrica que mide la capacidad del modelo para identificar correctamente los casos positivos reales. Indica que proporción de las muestras que pertenecen a una clase son efectivamente detectadas por el modelo. (Evidently AI, 2025) Se define como:

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

Esta métrica es especialmente relevante en el ámbito médico, ya que una baja sensibilidad implica un aumento de los falsos negativos, es decir, casos en los que una patología existente no es detectada por el modelo. Por esto, se considera una métrica crítica cuando el objetivo es minimizar el riesgo de no identificar enfermedades.

F1-score

El *F1-score* es una métrica que combina la precisión y el recall en un único valor, proporcionando una medida equilibrada del rendimiento del modelo. (Buhl, 2023) Se calcula como la media armónica de la precisión y el recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Esta métrica es especialmente útil en problemas con clases desbalanceadas, ya que penaliza de forma significativa los casos en los que una de las dos métricas es baja. En un contexto médico, el F1-score permite evaluar de manera más equilibrada el rendimiento del modelo cuando es importante minimizar tanto los falsos positivos como los falsos negativos.

Hamming Loss y Hamming Accuracy

En problemas de clasificación multilabel, cada muestra puede pertenecer a varias clases simultáneamente, lo que hace necesario utilizar

métricas específicas para evaluar el rendimiento del modelo en este contexto.

El *Hamming Loss* es una métrica que mide la proporción de etiquetas incorrectamente predichas por el modelo respecto al total de etiquetas evaluadas. (Amit, 2024) Se define como:

$$\text{Hamming Loss} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbb{I}(y_{ij} \neq \hat{y}_{ij})$$

donde N es el número de muestras, L es el número de clases, y_{ij} es la etiqueta real de la muestra i para la clase j, $\hat{y}_{ij}$ es la etiqueta predicha por el modelo, y $\mathbb{I}$ es la función indicadora que devuelve 1 si las etiquetas son diferentes y 0 si son iguales.

Un valor de Hamming Loss cercano a 0 indica un buen rendimiento del modelo ya que implica un menor número de errores a nivel de etiqueta individual.

De forma complementaria, el *Hamming Accuracy* se define como la proporción de etiquetas correctamente clasificadas respecto al total.

$$\text{Hamming Accuracy} = 1 - \text{Hamming Loss}$$

Es especialemnte útil en problemas multilabel, ya que evalúa el rendimiento del modelo considerando cada etiqueta de forma individual, en lugar de considerar únicamente la predicción global de cada muestra.

Promedios Micro y Macro

En problemas multietiqueta, las métricas de precisión, recall y F1 pueden agregarse mediante diferentes estrategias de promedio. (Qlik, s.f.) En este trabajo se utilizan los promedios macro y micro para evaluar el rendimiento de los modelos de forma complementaria.

El promedio macro calcula la métrica de forma independiente para cada clase y posteriormente obtiene la media aritmética de los resultados, otorgando el mismo peso a todas las clases independientemente

de su frecuencia. Esta medida es especialmente útil en escenarios con desbalance entre clases, ya que permite evaluar el rendimiento del modelo en patologías minoritarias sin que queden enmascaradas por las más frecuentes.

Por otro lado, el promedio micro calcula las métricas globales considerando conjuntamente el número total de verdaderos positivos, falsos positivos y falsos negativos, proporcionando una visión más global del rendimiento del sistema, donde las clases más frecuentes tienen mayor influencia en el resultado final.

MAE y MSE

El MAE (Mean Absolute Error) mide el error medio absoluto entre las predicciones del modelo y los valores reales. En este contexto, indica en qué medida las probabilidades estimadas se desvían, en promedio, de los valores reales. Se trata de una métrica fácil de interpretar y robusta frente a valores extremos.

El MSE (Mean Squared Error) calcula el error cuadrático medio, penalizando de forma más severa los errores grandes. Esto hace que sea especialmente útil para detectar modelos que cometen desviaciones importantes en algunas predicciones, aportando una visión complementaria al MAE. (Chugh, 2020)

Log-loss

La log-loss evalúa la calidad de las probabilidades predichas por el modelo, no solo la decisión final. Esta métrica penaliza de forma especialmente fuerte los errores cometidos con alta confianza, es decir, cuando el modelo asigna una probabilidad elevada a una clase incorrecta. Por este motivo, resulta muy adecuada en problemas de clasificación probabilística, ya que permite evaluar no solo si el modelo acierta, sino también si está bien calibrado en sus estimaciones.

Además, la log-loss es sensible al grado de incertidumbre del modelo: predicciones cercanas a los valores reales reducen significativamente

su valor, mientras que predicciones "sobreconfiadas" pero erróneas lo incrementan de forma notable. Esto la convierte en una métrica especialmente útil en contextos clínicos o multietiqueta desbalanceados, donde una mala estimación de la probabilidad puede ser más informativa que el propio error de clasificación. (KoshurAI, 2024)

AUC

La AUC (Area Under the ROC Curve) se interpreta de forma especialmente útil en contextos clínicos, ya que mide la capacidad del modelo para diferenciar entre pacientes sanos y pacientes con patología, independientemente del umbral que se utilice para tomar la decisión final.

Desde un punto de vista clínico, puede entenderse como la probabilidad de que el modelo asigne una mayor puntuación de riesgo a un paciente realmente enfermo que a uno sano. Por tanto, no depende de un punto de corte concreto, lo que resulta muy relevante en medicina, donde el umbral de decisión puede ajustarse en función del equilibrio entre falsos positivos y falsos negativos.

En escenarios como este, donde existe un fuerte desbalance entre clases, la AUC es especialmente útil porque no se ve tan afectada por la prevalencia de la clase mayoritaria como otras métricas, y permite evaluar de forma más estable la capacidad real del modelo para detectar patologías. (Attal, 2026)

Spearman

La correlación de Spearman es una métrica que se utiliza para medir la relación entre dos variables, pero no a partir de sus valores exactos, sino de su posición relativa u orden.

En el contexto del aprendizaje automático, esta métrica permite evaluar si el modelo es capaz de mantener una coherencia en el ranking de las predicciones respecto a los valores reales. Es decir, no se centra únicamente en acertar el valor exacto, sino en comprobar si los casos

más graves reciben puntuaciones más altas que los menos relevantes, y viceversa.

Esto es especialmente útil en problemas médicos, donde interesa que el modelo ordene correctamente las patologías según su probabilidad o severidad, aunque no siempre acierte con precisión el valor numérico.

En resumen, Spearman mide hasta qué punto el modelo respeta la estructura de orden de los datos reales, siendo una métrica muy adecuada para evaluar la calidad de predicciones probabilísticas en este tipo de aplicaciones clínicas. (Wei, 2024)

2.7. Tecnologías y librerías utilizadas

En este trabajo se ha empleado principalmente Python como lenguaje de programación, debido a su amplio número de librerías orientadas al análisis de datos, el procesamiento de señales y el desarrollo de modelos de Machine Learning.

En el procesamiento y manipulación de datos se han utilizado las librerias Pandas y NumPy. NumPy ha permitido trabajar con arrays multidimensionales de forma eficiente, mientras que Pandas ha facilitado la gestión de datos tabulares, permitiendo realizar operaciones de limpieza, transformación y análisis de manera sencilla, especialmente en lo relativo a las etiquetas y características de los pacientes.

Para el procesamiento de señales y análisis de señales se ha empleado SciPy, que proporciona herramientas específicas para el tratamiento de señales digitales, siendo especialemente relevante en el contexto de electrocardiogramas.

Por otro lado, la lectura y manipulación de los registros de electrocardiogramas se ha realizado mediante la librería WFDB, muy utlizada en el ámbito biomédico para el manejo de bases de datos de señales fisiológicas.

La implementación y validación de los modelos de aprendizaje automático se ha llevado a cabo utilizando Scikit-learn, que incluye una amplia variedad de algoritmos de clasificación, así como herramientas

para la evaluación del rendimiento de los modelos y la validación de los resultados.

Finalmente, para la visualización de datos y la creación de una interfaz interactiva se han utilizado la librerías Matplotlib, Dash y Plotly, que permiten generar gráficos y visualizaciones dinámicas, facilitando la interpretación de los resultados y la exploración de los datos.

Capítulo 3

Estado del Arte

Las enfermedades cardiovasculares son una de las principales causas de mortalidad en el mundo, lo que ha impulsado el desarrollo de herramientas de diagnóstico capaces de detectar estas patologías de manera temprana y precisa. Entre estas herramientas, el electrocardiograma (ECG) se destaca como una técnica no invasiva, rápida y ampliamente disponible para evaluar la actividad eléctrica del corazón.

La interpretación del ECG es una herramienta fundamental en el diagnóstico de un gran número de enfermedades cardiovasculares. Tradicionalmente, este proceso ha sido realizado por especialistas médicos de forma manual mediante la inspección visual de la señal, lo que requiere experiencia clínica y puede verse afectado por la subjetividad del observador o la variabilidad entre especialistas. Además, el aumento del volumen de datos clínicos generados en hospitales y sistemas de monitorización hace cada vez más complejo el análisis manual de grandes cantidades de registros electrocardiográficos.

En este contexto, la inteligencia artificial (IA) y las técnicas de aprendizaje automático han adquirido un papel cada vez más relevante en el ámbito del diagnóstico médico, ofreciendo la posibilidad de automatizar la interpretación del ECG y mejorar la precisión diagnóstica. Estos sistemas buscan ser herramientas de apoyo al diagnóstico, facilitando una interpretación más rápida y consistente de los registros electrocardiográficos, así como la detección de patrones sutiles que po-

drían pasar desapercibidos para el ojo humano.

3.1. Evolución del análisis automático del ECG

Los primeros enfoques para el análisis automático del ECG se basaban en técnicas de procesamiento de señales y extracción manual de características relevantes, como la duración de intervalos, amplitudes de onda o morfología del complejo QRS.

En esta línea, los métodos clásicos de aprendizaje automático han sido ampliamente utilizados. Un ejemplo de esto es el trabajo de Vimal y Sathish (Vimal y Sathish, 2010), donde se propone un sistema de clasificación de arritmias utilizando un modelo de Random Forest, basado en características extraídas manualmente del ECG. Se emplean técnicas como la variabilidad de la frecuencia cardíaca (HRV) y la transformada wavelet discreta (DWT) para generar características, obteniendo una alta capacidad de discriminación entre clases de arritmias.

Este tipo de enfoques dependen en gran medida de la calidad de las características extraídas, lo que puede limitar su capacidad de generalización en escenarios clínicos complejos o con variabilidad significativa en los registros electrocardiográficos.

Dentro de esta etapa de desarrollo de métodos clásicos, la disponibilidad de bases de datos de referencia es fundamental. La base de datos de MIT-BIH Arrhythmia Database (PhysioNet, 2005) se ha convertido en uno de los recursos más importantes en investigación sobre ECG.

La MIT-BIH Arrhythmia Database fue desarrollada conjuntamente por el Beth Israel Hospital de Boston y el MIT con el objetivo de proporcionar un conjunto estándar de señales ECG para la evaluación de algoritmos de detección de arritmias. Está compuesta por 48 registros de aproximadamente 30 minutos, obtenidos de pacientes ambulatorios, muestreados a 360 Hz y anotados por cardiólogos expertos, lo que la convierte en un estándar ampliamente utilizado en la validación de métodos automáticos.

Asimismo, la plataforma PhysioNet ha permitido el acceso a múltiples bases de datos biomédicas utilizadas en investigación. En particular, la base de datos utilizada en este trabajo proviene de este repositorio, lo que garantiza que los registros ECG empleados sean de referencia y ampliamente utilizados en la literatura científica

3.2. Inteligencia artificial aplicada al ECG

Con el avance del aprendizaje profundo, los modelos han evolucionado desde la dependencia de características manuales hasta sistemas capaces de aprender directamente del ECG.

Las redes neuronales convolucionales (CNN) han demostrado ser especialmente efectivas en el análisis de señales ECG, debido a su capacidad para capturar patrones locales y características morfológicas de la señal. En este contexto, Hannun et al. (Hannun et al., 2019) desarrollaron un modelo basado en *deep neural networks* para la clasificación de 12 tipos de ritmos cardíacos a partir de ECG ambulatorios de una sola derivación.

El modelo alcanza un área bajo la curva ROC de 0.97 y un F1-score de 0.837, superando el rendimiento medio de cardiólogos expertos en la misma tarea. Estos resultados evidencian el potencial del aprendizaje profundo en la interpretación automática del ECG en escenarios clínicos reales.

De forma complementaria, otros trabajos han explorado la aplicación de deep learning en patologías distintas a las arritmias. Un ejemplo de esto es estudio de Mayo Clinic (Wysokinski et al., 2024), donde se emplean redes neuronales profundas para la detección de embolia pulmonar a partir de un ECG de 12 derivaciones. El modelo alcanza un área bajo la curva ROC de 0.69 para la detección de la embolia pulmonar aguda, mejorando hasta 0.84 en casos de embolia de alto riesgo. Los autores destacan el potencial clínico del modelo, especialmente por su alto valor predictivo negativo, lo que sugiere su utilidad para descartar casos graves de forma rápida.

El uso del ECG no se limita únicamente a clasificación, sino que también se ha explorado la predicción de eventos cardiovasculares, lo que refuerza el potencial del ECG como herramienta predictiva en medicina.

3.3. Líneas de investigación actuales

En la literatura reciente, el análisis automático del electrocardiograma ha evolucionado hacia líneas de investigación que buscan mejorar no solo la capacidad predictiva de los modelos, sino también su robustez, interpretabilidad y aplicabilidad en entornos clínicos reales. En este sentido, la inteligencia artificial está comenzando a integrarse progresivamente en sistemas de apoyo a la decisión médica, donde se combinan modelos de aprendizaje automático con validación clínica con el objetivo de facilitar su incorporación a la práctica hospitalaria. Aunque estos sistemas muestran resultados prometedores, su uso rutinario aún requiere validación adicional en entornos reales. (European Society of Cardiology, 2024)

Una línea de investigación relevante se centra en el estudio de cómo la información contenida en el electrocardiograma puede variar en función de las derivaciones utilizadas. En este contexto, Bergquist et al. (Bergquist et al., 2023) analizan la detección de disfunción ventricular izquierda utilizando modelos de aprendizaje automático entrenados con derivaciones individuales del ECG, comparándolos con modelos entrenados con las 12 derivaciones completas. Los resultados muestran que incluso con una sola derivación es posible obtener un rendimiento comparable, lo que sugiere que distintas partes del ECG pueden contener información relevante de forma independiente.

El uso de base de datos clínicos a gran escala ha permitido el desarrollo de modelos más robustos y generalizables. Se han utilizado registros de ECG combinados con información clínica adicional para entrenar sistemas de inteligencia artificial capaces de detectar múltiples condiciones cardiovasculares de forma simultánea.

En esta línea, el trabajo de Kalmady et al. (Kalmady et al., 2024) desarrolla y valida algoritmos de aprendizaje automático sobre más de 1.6 millones de ECG. En este estudio se plantean modelos capaces de predecir simultáneamente hasta 15 diagnósticos cardiovasculares diferentes, mostrando un rendimiento elevado en tareas de clasificación múltiple y destacando el potencial del ECG como herramienta de cribado poblacional a gran escala.

Otra línea de investigación relevante es el desarrollo de modelos interpretables de aprendizaje automático. En aplicaciones médicas, la explicabilidad de los modelos es un aspecto crítico, ya que permite aumentar la confianza del personal clínico en las predicciones realizadas por los sistemas automáticos. En este sentido, se han propuesto técnicas de *interpretable machine learning* aplicadas al análisis de ECG con el objetivo de mejorar la transparencia de los modelos y facilitar la comprensión de sus decisiones.

Ayano et al. (Ayano et al., 2022) realizan una revisión sistemática sobre técnicas de aprendizaje automático interpretables aplicadas al ECG, destacando que la falta de interpretabilidad sigue siendo una de las principales barreras para la adopción clínica de estos sistemas. Además, identifican la necesidad de desarrollar modelos más explicables que permitan justificar las decisiones en entornos médicos reales.

De forma complementaria, revisiones recientes han analizado el estado del arte del uso de inteligencia artificial en el ECG, destacando la transición desde métodos basados en ingeniería de características hacia arquitecturas profundas con extracción automática de patrones. Estas revisiones también subrayan problemas recurrentes como el desbalanceo de clases, la variabilidad entre pacientes y la falta de métricas de evaluación estandarizadas.

En este sentido, Prakash et al. (Prakash et al., 2025) revisan más de 100 estudios sobre clasificación de latidos ECG, señalando que, aunque los modelos actuales han alcanzado altos niveles de precisión, todavía existen desafíos importantes relacionados con la generalización y la validación clínica.

Finalmente, distintos centros de investigación han desarrollado líneas de trabajo centradas en la aplicación de machine learning al análisis del ECG para la detección de patologías cardiovasculares. En estos trabajos se destaca la importancia de factores como la selección de datos, la arquitectura del modelo y la validación externa, que influyen de manera determinante en el rendimiento final de los sistemas. (University of Utah - CEG, s.f.)

3.4. Limitaciones de los enfoques actuales

A pesar del avance de las técnicas de inteligencia artificial aplicadas al análisis del electrocardiograma, los enfoques propuestos en la literatura presentan todavía diversas limitaciones que condicionan su aplicabilidad en entornos clínicos reales y su capacidad de generalización.

En primer lugar, una de las principales limitaciones reside en la simplificación del problema de clasificación. En numerosos trabajos, el diagnóstico de patologías cardiovasculares se plantea como un problema binario o como una clasificación multicategoría con un número reducido de clases. Sin embargo, esta formulación no refleja adecuadamente la complejidad del entorno clínico real, donde un mismo paciente puede presentar varias patologías simultáneamente. Esta simplificación facilita el entrenamiento de los modelos, pero reduce su capacidad para representar escenarios médicos más realistas.

Otro aspecto relevante es el uso limitado de la información disponible en las señales de electrocardiograma. En la literatura es frecuente encontrar estudios que utilizan únicamente un subconjunto reducido de derivaciones o representaciones simplificadas de la señal. Esta decisión puede implicar la pérdida de información espacial y temporal relevante, lo que afecta potencialmente a la capacidad diagnóstica de los modelos. Además, existe variabilidad entre trabajos en cuanto al número de derivaciones utilizadas y al tipo de señal considerada, lo que dificulta la comparación directa entre enfoques y limita la estandarización en el

campo.

Relacionado con lo anterior, la representación de los datos constituye otro factor crítico. Algunos enfoques trabajan directamente con la señal ECG en bruto, mientras que otros dependen de la extracción previa de características manuales o estadísticamente diseñadas. Esta diversidad de estrategias implica que el rendimiento de los modelos puede depender fuertemente del tipo de preprocesado aplicado, sin que exista un consenso claro sobre cuál es la representación más adecuada. Esta falta de estandarización refleja una de las principales debilidades metodológicas del área.

Por otro lado, la mayoría de los enfoques existentes asumen una única etiqueta por muestra, lo que implica tratar el problema como una clasificación de clase única. Esta aproximación simplifica considerablemente el diseño del modelo, pero no se ajusta completamente a la realidad clínica, en la que es frecuente la coexistencia de múltiples patologías en un mismo paciente. En este sentido, la ausencia de formulaciones multilabel en gran parte de la literatura limita la capacidad de los sistemas actuales para representar escenarios médicos complejos.

En cuanto a la evaluación de los modelos, es habitual el uso de métricas globales como accuracy, F1-score o AUC. Sin embargo, estas métricas no siempre reflejan adecuadamente el impacto clínico de los errores cometidos, especialmente en escenarios con clases desbalanceadas. En muchos casos, la evaluación se realiza en condiciones controladas que no permiten analizar con suficiente detalle el comportamiento del modelo en situaciones clínicas reales.

El desbalanceo de clases constituye, además, otra limitación importante. Los conjuntos de datos utilizados en este tipo de problemas suelen presentar una distribución desigual de las patologías, con una clara predominancia de las clases más frecuentes. Esto provoca que los modelos tiendan a favorecer dichas clases mayoritarias, afectando especialmente al rendimiento en la detección de enfermedades menos representadas. Aunque existen técnicas para mitigar este problema, su aplicación no es homogénea en la literatura.

Finalmente, la interpretabilidad de los modelos representa una barrera fundamental para su adopción clínica. Muchos de los enfoques basados en aprendizaje profundo funcionan como modelos de tipo "caja negra", lo que dificulta la comprensión de las decisiones que generan. Esta falta de explicabilidad reduce la confianza de los profesionales sanitarios en los sistemas automáticos y limita su integración en la práctica médica. Aunque existen líneas de investigación orientadas al desarrollo de modelos más interpretables, todavía existe un compromiso claro entre rendimiento y capacidad de explicación.

En conjunto, todas estas limitaciones evidencian que, a pesar del alto rendimiento alcanzado por los modelos actuales en entornos experimentales, todavía existen importantes retos abiertos en cuanto a la representación del problema, la calidad y utilización de los datos, la evaluación de los sistemas y su aplicabilidad clínica. Estas limitaciones justifican la necesidad de desarrollar enfoques más completos, capaces de abordar el problema desde una perspectiva más realista y alineada con la práctica médica.

Capítulo 4

Definición del Trabajo

4.1. Justificación del proyecto

A partir de las limitaciones analizadas en el estado del arte, el presente trabajo plantea un análisis comparativo de distintos escenarios de clasificación de señales ECG, evaluando el impacto que tienen distintos factores sobre el rendimiento de los modelos de aprendizaje automático. Se analiza la influencia del número de derivaciones utilizadas, comparando los resultados obtenidos con una sola derivación frente a los obtenidos utilizando el ECG completo de 12 derivaciones. Del mismo modo, se comparan distintos formas de representar la información, desde la señal ECG en bruto hasta la extracción de características diseñadas manualmente o estadísticas, con el objetivo de evaluar su impacto en el rendimiento de los modelos.

El trabajo utiliza distintos enfoques del problema de clasificación. Se considera un enfoque binario que detecta grupos generales de patologías y otro enfoque más complejo basado en múltiples enfermedades y salidas probabilísticas. Permite analizar escenarios más próximos a la realidad clínica, donde un mismo registro puede puede presentar varias alteraciones simultáneamente y donde el nivel de confianza asociado a cada predicción puede resultar relevante desde el punto de vista médico.

Además, el trabajo no se limita únicamente al estudio del rendimien-

to global de los modelos, sino que también analiza el comportamiento de distintas métricas de evaluación adaptadas a problemas con clases desbalanceadas y escenarios multilabel. De esta forma, se busca obtener una evaluación más representativa del comportamiento real de los clasificadores en tareas de diagnóstico automático.

En conjunto, el objetivo del trabajo es estudiar distintas estrategias de clasificación sobre registros electrocardiográficos reales, analizando las ventajas y limitaciones de cada configuración experimental y explorando enfoques más flexibles y representativos de situaciones clínicas reales.

4.2. Base de datos utilizada

Para el desarrollo del presente trabajo se ha utilizado una base de datos de registros electrocardiográficos ampliamente utilizada en investigación biomédica. La selección de una base de datos adecuada resulta especialmente relevante en problemas de clasificación automática de ECG, ya que las características de los registros, el etiquetado clínico y la diversidad de patologías condicionan de forma directa el rendimiento y la capacidad de generalización de los modelos desarrollados.

En este trabajo se emplea la base de datos PTB-XL, disponible públicamente a través de la plataforma PhysioNet. (Wagner et al., 2022) Esta base de datos ha sido diseñada específicamente para facilitar tareas de investigación relacionadas con el análisis automático de señales ECG y el desarrollo de modelos de aprendizaje automático aplicados al diagnóstico cardiovascular.

4.2.1. Origen del dataset

La base de datos PTB-XL fue desarrollada con el objetivo de proporcionar un conjunto de datos electrocardiográficos de gran tamaño y alta calidad, que permita el entrenamiento y la evaluación de modelos de inteligencia artificial en tareas de clasificación de patologías

cardíacas.

El dataset cuenta con más de 21000 registros de ECG de 12 derivaciones, obtenidos de aproximadamente 19000 pacientes diferentes. Proceden de práctica clínica real, y fueron anotados por especialistas médicos, lo que garantiza la calidad y relevancia de las etiquetas asociadas a cada registro.

4.2.2. Estructura del dataset

Cada registro de la base de datos corresponde a un ECG de 12 derivaciones con una duración aproximada de 10 segundos. Las señales incluyen las derivaciones estándar (I, II, III, aVR, aVL, aVF) y las precordiales (V1-V6).

Los registros se encuentran almacenados en formato WFDB, muy utilizado en bases de datos biomédicas y compatible con herramientas de procesamiento de señales ECG en Python. Además de las señales electrocardiográficas, el dataset incorpora información adicional asociada a cada registro, incluyendo identificadores de paciente, características demográficas y anotaciones diagnósticas.

PTB-XL proporciona dos versiones de las señales ECG con distintas frecuencias de muestreo: una versión de 100 Hz y otra de 500 Hz. Esta característica permite adaptar el procesamiento de las señales a distintos tipos de aplicaciones y modelos de aprendizaje automático. En este trabajo se ha utilizado la versión de 500 Hz, ya que proporciona una mayor resolución temporal, lo que puede ser beneficioso para la detección de patrones sutiles en la señal ECG. Es por esto que la señal se presenta como una matriz de 12 x 5000, donde las 12 filas corresponden a las derivaciones y las 5000 columnas representan las muestras de la señal a lo largo del tiempo.

La estructura del dataset permite trabajar tanto con las señales originales completas como con características extraídas posteriormente mediante técnicas de procesamiento y análisis de señal.

4.2.3. Etiquetado original

El sistema de etiquetado de la base de datos se basa en un enfoque multilabel, en el que cada registro electrocardiográfico puede estar asociado simultáneamente a múltiples patologías. A diferencia de un esquema de clasificación convencional con una única clase por muestra, este tipo de anotación permite representar escenarios clínicos más realistas, en los que pueden coexistir distintas alteraciones cardíacas en un mismo paciente.

En este trabajo, las etiquetas se encuentran estructuradas en forma de diccionario, donde cada clave corresponde a una de las 71 patologías consideradas en la base de datos. El valor asociado a cada clave representa una probabilidad (expresada en escala de 0 a 100) que indica el grado de presencia o compatibilidad de dicha patología en el registro ECG correspondiente.

Esta representación transforma el problema en un escenario de clasificación multilabel con salidas probabilísticas, lo que permite desarrollar modelos capaces de predecir no solo la presencia o ausencia de cada patología, sino también el nivel de confianza asociado a cada predicción.

A partir de este etiquetado original, el trabajo plantea posteriormente distintos escenarios de clasificación, incluyendo formulaciones binarias y agrupaciones de patologías en clases más generales, con el objetivo de analizar el comportamiento de los modelos bajo diferentes niveles de complejidad.

4.2.4. Problemas del dataset

A pesar de tratarse de una base de datos ampliamente utilizada en investigación y de gran valor para el desarrollo de modelos de aprendizaje automático, el conjunto de datos presenta diversas limitaciones que condicionan el diseño de los modelos y la interpretación de los resultados.

En primer lugar, se observa un desbalanceo significativo entre patologías. Algunas enfermedades aparecen con una frecuencia elevada dentro del conjunto de registros, mientras que otras están representadas por un número reducido de casos. Esta distribución desigual puede influir en el proceso de aprendizaje, favoreciendo la detección de las clases mayoritarias y dificultando la identificación de patologías menos representadas.

En segundo lugar, la propia naturaleza del etiquetado introduce complejidad adicional. Al tratarse de un sistema multilabel probabilístico, en el que cada patología se asocia a un valor de probabilidad entre 0 y 100, existe una dependencia directa del criterio del umbral que se utilice posteriormente para transformar estas probabilidades en decisiones discretas. Este aspecto puede afectar de forma relevante al rendimiento final del sistema.

Otro factor importante es la coexistencia de múltiples patologías en un mismo registro, lo que incrementa la complejidad del problema respecto a enfoques de clasificación tradicional. Esta situación refleja mejor la realidad clínica, pero dificulta el aprendizaje de fronteras de decisión claras entre clases.

Además, la variabilidad intrínseca de las señales electrocardiográficas, derivada de diferencias entre pacientes, condiciones de adquisición y posibles interferencias en la señal, introduce ruido que puede afectar a la robustez de los modelos.

Finalmente, aunque la base de datos es un referente en el ámbito del análisis automático de ECG, los modelos entrenados sobre este conjunto pueden no generalizar de forma directa a otros entornos clínicos o poblaciones distintas, lo que pone de manifiesto la necesidad de validaciones adicionales en escenarios externos.

4.3. Agrupación de patologías

La base de datos utilizada en este trabajo contiene un total de 71 patologías diferentes, definidas mediante el sistema de anotación SCP-

ECG. Este nivel de precisión introduce una elevada complejidad en el problema de clasificación, especialmente en escenarios multilabel y con distribución desbalanceada de clases.

Con el objetivo de simplificar el problema y estructurar el análisis, en este trabajo se emplea una agrupación de las patologías en cinco superclases clínicas generales, previamente definidas en el apartado 2.3. Esta agrupación permite reducir la complejidad del espacio de salida y facilita la interpretación de los resultados obtenidos por los modelos.

Este enfoque permite trabajar simultáneamente con dos niveles de análisis: por un lado, el nivel detallado basado en las 71 patologías originales, y por otro, un nivel más general basado en las cinco superclases. Esta dualidad resulta especialmente útil para evaluar el impacto del nivel de detalle en el rendimiento de los modelos de clasificación.

Además, esta estructura facilita la definición de distintos escenarios experimentales, permitiendo comparar el comportamiento de los modelos bajo formulaciones más complejas frente a aproximaciones más simplificadas.

4.4. Escenarios experimentales

Con el objetivo de analizar el impacto de distintos factores en el rendimiento de los modelos de clasificación automática de ECG, se plantean tres enfoques experimentales diferentes, que varían en función del número de derivaciones utilizadas, el nivel de detalle en la representación de las patologías y el tipo de salida generada por el modelo.

4.4.1. Configuración 1: señal reducida y clasificación binaria

En este primer escenario plantea un enfoque simplificado, en el que se utiliza únicamente la derivación I del ECG, y se agrupan las patologías en las cincos superclases definidas previamente. El objetivo de esta configuración es evaluar el rendimiento de los modelos en condi-

ciones de información reducida, simulando escenarios en los que solo se dispone de dispositivos de monitorización con capacidad limitada.

Este modelo se entrena en un problema de clasificación multilabel, en el que se estima de forma independiente la presencia o ausencia de cada una de las cinco superclases de patologías. Esto permite que un mismo registro pueda presentar simultáneamente varias condiciones clínicas.

Esta configuración se entrenará tanto con la señal ECG en bruto como con características extraídas, lo que permitirá analizar el impacto de la representación de los datos en el rendimiento del modelo.

4.4.2. Configuración 2: señal completa y clasificación binaria

En el segundo escenario se utiliza la señal completa de 12 derivaciones pero manteniendo la misma agrupación de patologías en cinco superclases.

De esta forma, este escenario permite aislar el efecto del aumento en la información de entrada, manteniendo constante la complejidad del problema de clasificación.

En este caso, debido a la enorme cantidad de datos, el modelo se entrenará únicamente con las características extraídas de la señal, lo que permitirá evaluar el impacto de la representación de los datos en un escenario con información completa.

4.4.3. Configuración 3: señal completa y clasificación multilabel probabilística

En este tercer escenario se plantea el enfoque más complejo, en el que se utiliza la señal completa de 12 derivaciones y se trabaja con las 71 patologías originales, sin realizar agrupaciones.

Además, en este caso se mantiene la naturaleza multilabel y probabilística del etiquetado original, lo que implica que el modelo debe

ser capaz de predecir simultáneamente la presencia o ausencia de cada patología, así como el nivel de confianza asociado a cada predicción.

Este escenario permite evaluar el comportamiento de los modelos en condiciones lo más cercanas posible a la complejidad clínica real, donde pueden coexistir múltiples patologías simultáneamente y donde la salida no se limita a una decisión binaria.

Al igual que en el escenario anterior, debido a la gran cantidad de datos y la complejidad del problema, el modelo se entrenará únicamente con características extraídas de la señal ECG.

4.5. Objetivos técnicos

A partir de los escenarios experimentales definidos anteriormente, este trabajo plantea una serie de objetivos técnicos orientados a analizar el impacto de diferentes decisiones de modelado en el rendimiento de los sistemas de clasificación de señales electrocardiográficas.

En primer lugar, se estudia el efecto del número de derivaciones utilizadas en la entrada del modelo. Se compara el rendimiento obtenido al emplear una única derivación frente al uso completo de las 12 derivaciones disponibles, con el objetivo de evaluar hasta qué punto la información espacial del electrocardiograma influye en la capacidad de discriminación de las distintas patologías.

En segundo lugar, se analiza la influencia de la formulación del problema de clasificación. En particular, se comparan enfoques binarios basados en la detección simplificada de presencia o ausencia de patología con un enfoque multilabel probabilístico, en el que cada salida del modelo representa la probabilidad de pertenencia a una determinada clase.

Otro objetivo consiste en el análisis del nivel de detalle considerado en las etiquetas. Para ello, se contraponen escenarios basados en un conjunto reducido de 5 clases frente al uso completo de 71 patologías. Esta comparación permite evaluar cómo afecta la complejidad del espacio de salida al rendimiento global del modelo.

Por otro lado, se aborda la transformación de las probabilidades asociadas a las etiquetas en decisiones discretas en los escenarios binarios. Para ello, se emplean umbrales óptimos, determinados de forma empírica con el objetivo de maximizar el rendimiento del modelo en cada caso. El análisis de estos umbrales resulta clave para equilibrar la sensibilidad frente a distintas clases y mejorar la capacidad de generalización.

Finalmente, se pretende estudiar la interacción entre todos estos factores, evaluando cómo la combinación de la representación de la señal, la formulación del problema y la estructura de las etiquetas condiciona el comportamiento de los modelos de clasificación automática de ECG.

4.6. Metodología general

A partir de los escenarios experimentales y objetivos definidos anteriormente, se presenta a continuación la metodología general empleada en este trabajo, que integra el procesamiento de los datos, el diseño de los modelos y la evaluación de los resultados.

Cada registro de la base de datos es una matriz de 12 x 5000, donde las filas corresponden a las derivaciones del ECG y las columnas representan las muestras de la señal a lo largo del tiempo. También incluye información demográfica como la edad y el sexo del paciente, así como un sistema de etiquetado basado en un diccionario que asocia cada patología a una probabilidad de presencia. En función del escenario experimental considerado, se ha transformado la etiqueta original para adaptarla a cada caso.

Antes de comenzar con la limpieza y preprocesado de las señales se eliminaron todos los registros de aquellos pacientes con menos de 18 años y más de 89, con el objetivo de enfocar el análisis en la población adulta.

Posteriormente, las señales pasan por un proceso de preprocesado con el objetivo de detectar los ciclos cardíacos principales, identificando los picos R como referencia dentro del ECG. A partir de esta informa-

ción, se eliminan los registros que contienen menos de 10 latidos, y de los restantes se seleccionan únicamente los 10 primeros latidos para cada paciente. Tras este filtrado, el conjunto final queda compuesto por 16984 pacientes. Se segmenta la señal en latidos individuales, lo que permite estrucutrar los datos de forma homogénea.

A partir de esta representación, se consideran distintas formas de entrada para los modelos. En un primer enfoque, se construye un megavector concatenando los latidos de cada paciente, de manera que el modelo pueda analizar la señal completa de forma global. En un segundo enfoque, se extraen características temporales, morfológicas y estadísticas de cada latido, con el objetivo de proporcionar a los modelos una representación más compacta y representativa de la señal.

Una vez preparados los datos, se procede al entrenamiento de los distintos modelos de aprendizaje automático definidos en el estudio, evaluando su comportamiento bajo las diferentes configuraciones experimentales.

Finalmente, la evaluación de los resultados se realiza mediante métricas adecuadas a cada escenario experimental, permitiendo comparar de forma consistente el rendimiento de los modelos. Adicionalmente, se ha desarrollado una herramienta de visualización tipo dashboard con el objetivo de facilitar el análisis de los resultados obtenidos.

Capítulo 5

Modelo Desarrollado

En este capítulo se describe la estructura del sistema desarrollado para el análisis automático de señales electrocardiográficas. Se detallan los distintos bloques que componen el pipeline de procesamiento, desde la entrada de los datos hasta la obtención de las predicciones finales.

Además, se presentan las principales etapas de análisis de los datos, así como los procesos de preprocesado y transformación de las señales ECG utilizados como base para los modelos de aprendizaje automático.

Finalmente, se describen los distintos enfoques experimentales planteados y la metodología seguida para el desarrollo y evaluación de los modelos, con el objetivo de analizar el impacto de las distintas configuraciones propuestas.

5.1. Arquitectura general del sistema

El sistema desarrollado en este trabajo se basa en un pipeline de aprendizaje automático aplicado al análisis de señales ECG, cuyo objetivo es la detección y clasificación automática de patologías cardíacas a partir de registros electrocardiográficos.

Como se muestra en la Figura 5.1, el sistema está compuesto por varias etapas secuenciales que permiten transformar la señal original en predicciones clínicas relevantes. El flujo completo empieza con la entrada de los registros electrocardiográficos, junto con la información

asociada a cada paciente, y las etiquetas diagnósticas correspondientes.

A continuación, las señales pasan por una etapa de preprocesado, donde los datos se limpian y se preparan con el objetivo de mejorar la representación para el posterior análisis. Esta fase permite homogeneizar los registros y eliminar posibles problemas derivados de la variabilidad intrínseca de las señales ECG.

Posteriormente, los datos procesados son transformados para su utilización en los modelos de aprendizaje automático. Dependiendo del escenario experimental considerado, esta transformación puede realizarse mediante un megavector que concatena los latidos de cada paciente o mediante la extracción de características representativas de la señal ECG.

Una vez obtenida la representación adecuada de los datos, se procede al entrenamiento de los modelos de aprendizaje automático definidos en el estudio, evaluando su rendimiento bajo las distintas configuraciones experimentales planteadas. La salida del sistema puede representarse de forma binaria o probabilística, dependiendo del enfoque utilizado en cada caso.

Finalmente, las predicciones obtenidas son evaluadas mediante distintas métricas de rendimiento, permitiendo comparar el comportamiento de los modelos en los diferentes escenarios planteados.

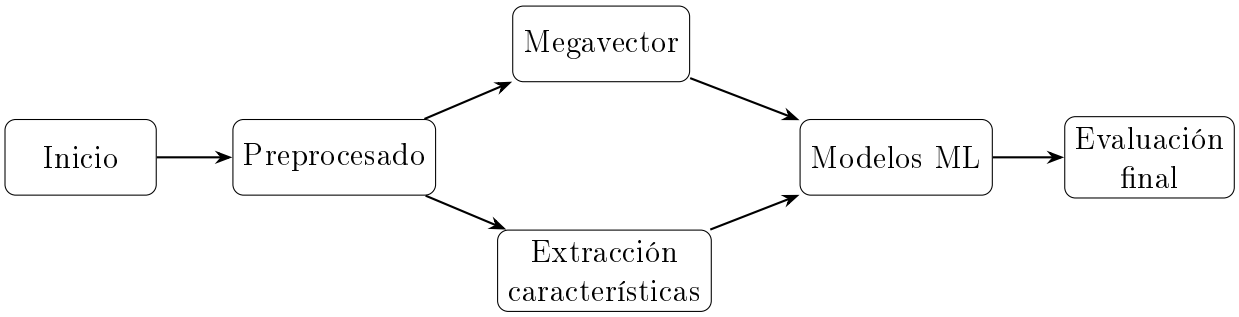

Figura 5.1: Pipeline general del sistema. Elaboración propia.

5.2. Análisis exploratorio de datos

En este apartado se realiza un análisis exploratorio de los datos de la bases de datos utilizada, con el objetivo de comprender la estructura del conjunto utilizado, así como las caracterísitcas demográficas de la población estudiada. Este análisis permite contextualizar los datos antes de su utilización en los modelos de aprendizaje automático y detectar posibles patrones relevantes o desbalances en la población.

En primer lugar, se analiza la distribución de los pacientes por edad en el conjunto de datos. La Figura 5.2 muestra el histograma de edades correspondiente, que revela una distribución relativamente equilibrada entre los distintos grupos de edad, con una ligera predominancia de pacientes en el rango de 50 a 70 años, lo que es consistente con la prevalencia de patologías cardiovasculares en la población adulta. Como se ha mencionado anteriormente, se han eliminado los registros de pacientes menores de 18 años y mayores de 89 años, por lo que en el gráfico solo se muestran los pacientes comprendidos en este rango de edad.

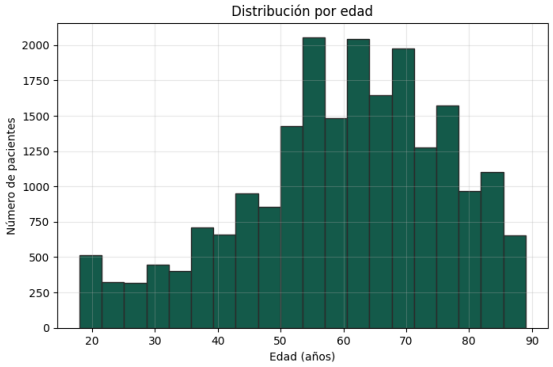

Figura 5.2: Distribución de los pacientes por edad

A continuación, se analiza la distribución de los pacientes por sexo. La Figura 5.3 muestra que existe una ligera predominancia de pacientes masculinos en el conjunto de datos, con aproximadamente un 55 % de hombres frente a un 45 % de mujeres. Esta distribución es coherente

con la epidemiología de las enfermedades cardiovasculares, que suelen presentar una mayor prevalencia en hombres (Salvavidas, 2025), aunque también es importante destacar la presencia significativa de mujeres en el conjunto de datos, lo que permite desarrollar modelos que puedan generalizar a ambos sexos. El 0 en el gráfico corresponde a pacientes masculinos y el 1 a pacientes femeninos.

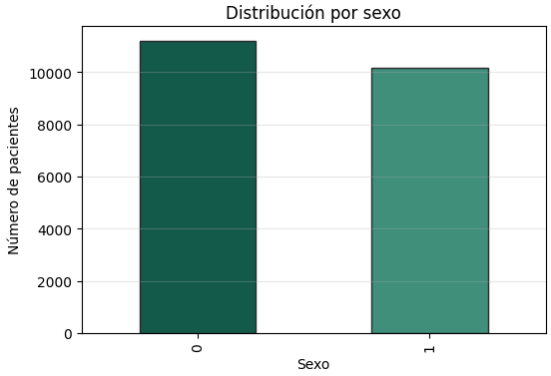

Figura 5.3: Distribución de los pacientes por sexo

De forma complementaria, se analiza la relación entre edad y sexo, con el fin de obtener una visión más detallada de la composición de la población. Este análisis permite observar cómo se distribuyen ambas variables de forma conjunta y si existen diferencias relevantes entre grupos. La Figura 5.4 representa esta relación mediante una visualización combinada de ambas variables.

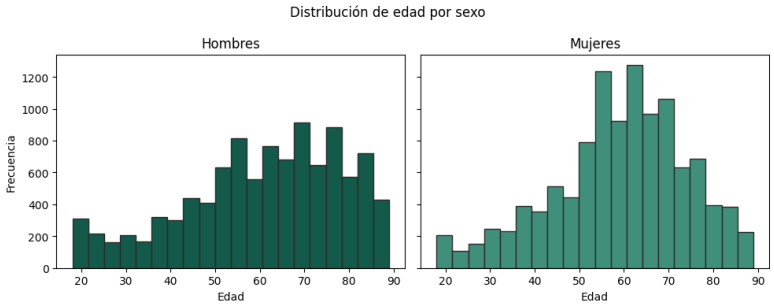

Figura 5.4: Distribución de los pacientes por edad y sexo

En conjunto, este análisis proporciona una visión general de la es-

tructura demográfica del conjunto de datos, lo cual resulta fundamental para interpretar adecuadamente los resultados obtenidos en las etapas posteriores del estudio.

Una vez analizada la distribución demográfica del conjunto de datos, se estudia la distribución de las etiquetas diagnósticas con el objetivo de evaluar el balance entre las distintas patologías presentes en el estudio.

En primer lugar, se analiza la distribución de las clases originales del conjunto de datos. Dado el elevado número de categorías (71 patologías), se presenta una visualización alternativa basada en las 10 clases más frecuentes, agrupando el resto en una categoría adicional. Esta representación, mostrada en la Figura 5.5, permite observar de forma más clara las patologías predominantes en el conjunto de datos y facilita la interpretación del desbalance existente.

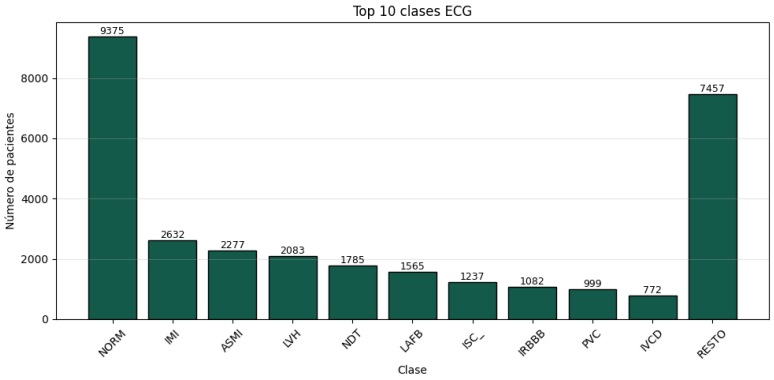

Figura 5.5: Distribución de las patologías más frecuentes en el conjunto de datos.

Además, se analiza la distribución del conjunto de datos en función de las superclases en las que se agrupan las patologías. La Figura 5.6 muestra esta agrupación, en la que se observa una reducción de la especificidad del problema original, permitiendo una visión más general del tipo de patologías presentes en el dataset.

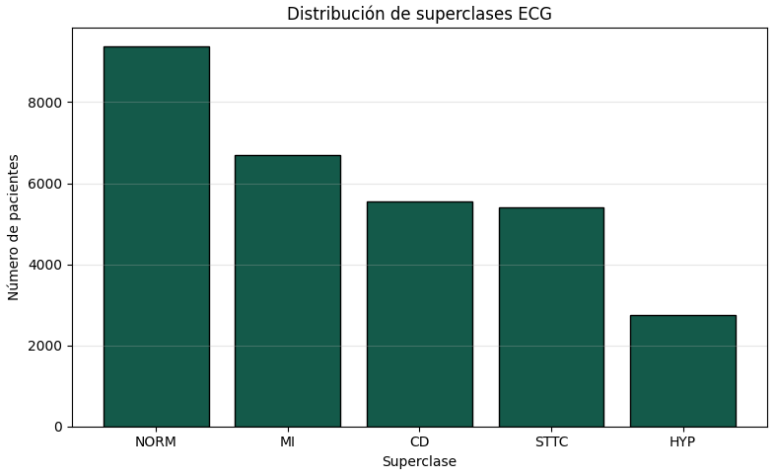

Figura 5.6: Distribución de las patologías por superclases.

En conjunto, este análisis evidencia la existencia de un importante desbalance en las etiquetas, tanto a nivel de clases individuales como en su agrupación por superclases, lo cual constituye un aspecto relevante a tener en cuenta en el desarrollo de los modelos de clasificación.

Cabe destacar que, en todos los casos analizados, la clase "NORM", correspondiente a registros electrocardiográficos normales, se mantiene como la clase mayoritaria. Esta predominancia es consistente tanto en la distribución de clases individuales como en las representaciones agrupadas, lo que refuerza el carácter desbalanceado del conjunto de datos y la importancia de considerar este efecto en el diseño y evaluación de los modelos.

Por otro lado, se analiza la coexistencia de patologías mediante una matriz de co-ocurrencia entre las 20 clases más frecuentes, ya que representar las 71 enfermedades dificulta la interpretación visual.

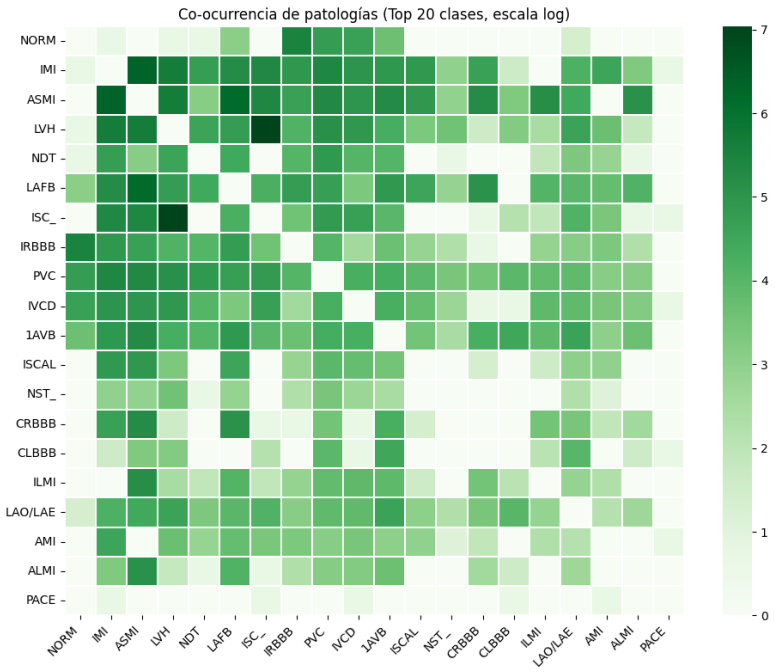

Figura 5.7: Matriz de co-ocurrencia de las 20 clases más frecuentes.

En la Figura 5.7 se observa que existen ciertas patologías que tienden a coexistir con mayor frecuencia. Destaca especialmente la relación entre infartos como IMI (infarto inferior), ASMI (infarto anteroseptal) y ALMI (infarto lateral alto) con alteraciones de conducción como LAFB (bloqueo fascicular anterior izquierdo), CRBBB (bloqueo de rama derecha completo) e IRBBB (bloqueo de rama derecha incompleto). Estas asociaciones reflejan la frecuente coexistencia entre alteraciones estructurales del corazón y trastornos del sistema de conducción.

De igual forma, se observan co-ocurrencias relevantes entre patologías isquémicas como ISC_ (isquemia general), ISCAL (isquemia anterolateral) e ISCIN (isquemia inferior) y eventos como IMI, lo que pone de manifiesto la superposición de manifestaciones electrocardiográficas en distintos estadios de enfermedad coronaria.

Por otro lado, patologías como LVH (hipertrofia ventricular izquierda) presentan múltiples asociaciones con distintas clases, especialmente con ISC_ y con alteraciones de conducción como IVCD (trastorno

inespecífico de conducción intraventricular), lo que sugiere que aparece frecuentemente asociada a otras patologías dentro del conjunto de datos. También se identifican combinaciones poco frecuentes entre determinadas patologías, especialmente en el caso de arritmias aisladas, lo que refleja la heterogeneidad del dataset y el marcado desbalance entre clases.

Estos factores justifican la necesidad de enfoques de modelado capaces de manejar múltiples etiquetas y capturar relaciones entre clases, aspectos que se abordarán en las siguientes secciones del trabajo.

Además de analizar la frecuencia individual de cada patología, es interesante analizar cuantas patologías diferentes aparecen asociadas a un mismo registro ECG. Este análisis permite evaluar la complejidad de la base de datos y calcular la frecuencia con la que varias alteraciones cardíacas coexisten en un mismo paciente.

Como se observa en la Figura 5.8, la mayoría de los registros presentan una o ninguna patología asociada, lo que indica una predominancia de electrocardiogramas normales. Sin embargo, también existe una proporción significativa de registros que presentan múltiples patologías simultáneamente, lo que refleja la complejidad clínica del conjunto de datos y la necesidad de desarrollar modelos capaces de manejar escenarios multilabel en los que coexisten múltiples alteraciones.

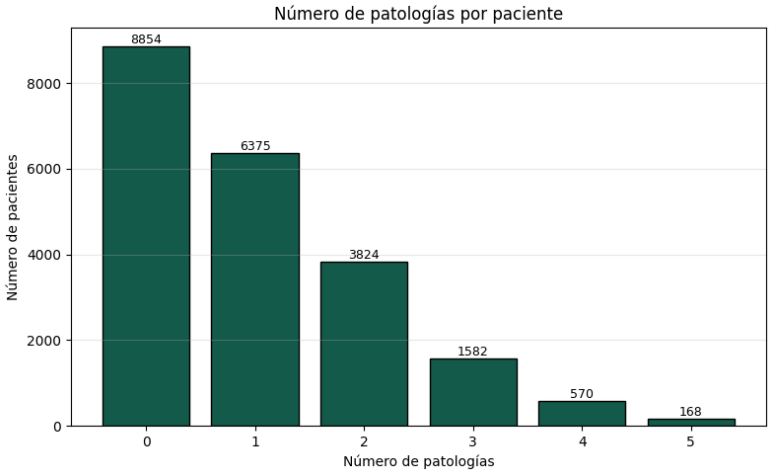

Figura 5.8: Distribución del número de patologías por paciente.

Para la elaboración de este gráfico se ha considerado que un paciente tuviese 0 patologías si como único elemento en la etiqueta presenta la categoría "NORM" con una probabilidad mayor que 0, mientras que si presentaba esta categoría y otra más con una probabilidad mayor que 0 se ha considerado que el paciente presenta 1 patología, y así sucesivamente.

5.3. Preprocesado de datos

El preprocesado de datos es esencial para mejorar la calidad de las señales ECG y facilitar el aprendizaje de los modelos de clasificación. Permite reducir la influencia de registros poco representativos, homogeneizar las entradas del sistema y facilitar la extracción de patrones relevantes para el diagnóstico automático de patologías cardíacas.

En este trabajo, el preprocesado de las señales se centra en la detección de los ciclos cardíacos principales, utilizando los picos R como referencia para segmentar la señal en latidos individuales.

Antes de proceder a la detección de los picos R, se aplican distintas etapas de procesamiento sobre la señal ECG con el objetivo de resaltar las características asociadas a los complejos QRS y facilitar

su identificación. Estas operaciones permiten reducir la influencia del ruido y mejorar la precisión del algoritmo de detección, especialmente en registros con variaciones de amplitud o interferencias.

En primer lugar, se realiza una descomposición de la señal mediante una transformada wavelet discreta, que permite separar la señal en diferentes bandas de frecuencia. (Talebi, 2025) Se utiliza la transformada wavelet de Daubechies de orden 6 (db6) como base para la descomposición de la señal, mostrada en la Figura 5.9. Las wavelets de esta familia presentan una forma similar a los complejos QRS y ofrecen un buen equilibrio entre resolución temporal y frecuencial, lo que las hace especialmente adecuadas para el análisis de señales electrocardiográficas.

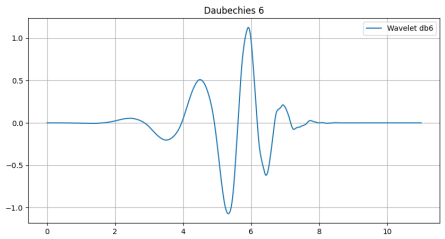

Figura 5.9: Morfología de la wavelet db6.

La señal se descompone en 8 niveles, permitiendo dividir el espectro en distintas bandas de frecuencia. Esta profundidad es adecuada para la frecuencia de muestreo utilizada (500 Hz) y permite aislar con mayor precisión las componentes asociadas a las distintas ondas del ECG. Los coeficientes d4 y d5 contienen la información correspondiente al rango de frecuencias típico del complejo QRS (aproximadamente entre 5 y 25 Hz). Por este motivo, serán estos coeficientes los que se utilicen para reconstruir la señal que se utilizará para detectar los latidos.

En la Figura 5.10 se muestra la señal original junto con la señal reconstruida a partir de los coeficientes d4 y d5 de la transformada wavelet de la derivación I del registro 00010. Es importante señalar que la transformada wavelet discreta implica un proceso de submuestreo en cada nivel de descomposición, lo que modifica la representación de la señal en el dominio transformado. Sin embargo, en este trabajo el

objetivo no es preservar la amplitud de la señal filtrada, sino la correcta localización temporal de los picos R. Por este motivo, las posibles variaciones en la escala de la señal no afectan al propósito principal del algoritmo, ya que la detección se basa en la posición de los máximos y no en su valor absoluto.

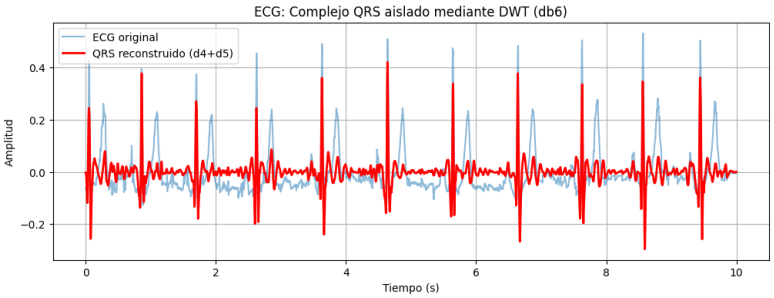

Figura 5.10: Señal original y señal reconstruida a partir de los coeficientes d4 y d5

Después de reconstruir la señal a partir de los coeficientes d4 y d5, se aplica la transformada Hilbert para obtener la envolvente de la señal. Esta envolvente es una versión suavizada que sigue la amplitud de la señal original, destancando sus máximos locales. De este modo, los picos R pueden identificarse de forma más precisa como lo máximos de dicha envolvente, reduciendo el impacto del ruido y de posibles variaciones en la forma de la señal.

En la Figura 5.11 se muestra la envolvente de la señal reconstruida la derivación I del registro 00010, donde se pueden observar claramente los picos correspondientes a los latidos cardíacos, lo que facilita su detección y posterior segmentación de la señal en latidos individuales.

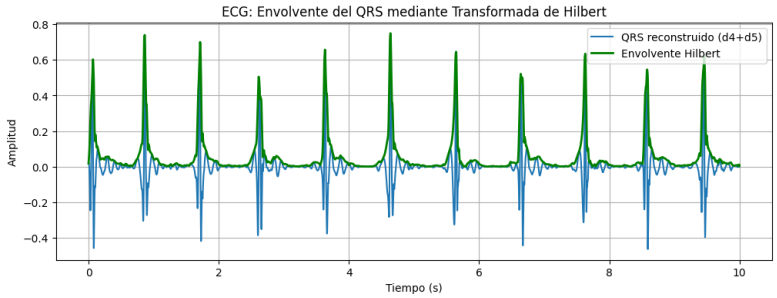

Figura 5.11: Envolvente de la señal reconstruida a partir de los coeficientes d4 y d5.

La detección de los picos R se realiza mediante un algoritmo de búsqueda de máximos locales en la envolvente obtenida. Para ello, se utiliza un umbral adaptativo que permite identificar los latidos incluso en registros con variaciones de amplitud o presencia de ruido. Se define como la suma de la media de la envolvente y su desviación estándar, lo que permite ajustar el umbral de detección en función de las características específicas de cada registro. Este enfoque evita la necesidad de un umbral fijo y mejora la robustez frente a la variabilidad entre pacientes y niveles de ruido. Adicionalmente, se impone una distancia mínima entre picos consecutivos con el fin de evitar la detección múltiple de un mismo latido y garantizar la coherencia fisiológica de los resultados. (Chauhan et al., 2021)

Como se observa en la Figura 5.12, los picos R detectados se corresponden con los máximos de la envolvente, lo que confirma la eficacia del proceso de preprocesado para resaltar las características relevantes de la señal ECG y facilitar la identificación de los latidos cardíacos.

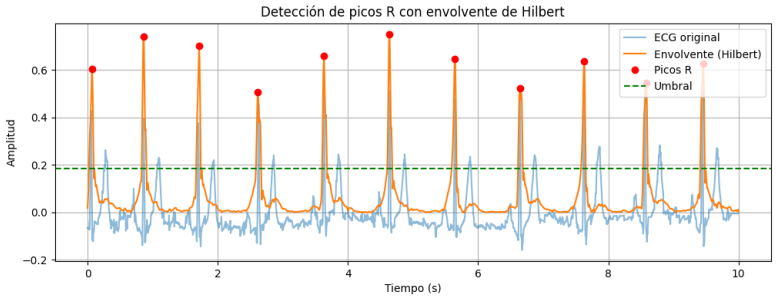

Figura 5.12: Picos R detectados en la envolvente.

Una vez detectados los picos R, se segmenta la señal ECG en latidos individuales, utilizando cada pico como referencia para definir el inicio y el final de cada latido. Por cada pico R, se extrae un segmento de 600 ms, comprendido entre 200 ms antes y 400 ms después del pico.

Esta elección se fundamenta en la morfología temporal del ECG, en el que el complejo QRS se localiza en torno al pico R y actúa como referencia estable para la alineación de los latidos. La inclusión de una ventana previa al pico R permite incorporar la transición desde la onda P y asegurar la captura completa del inicio del complejo QRS, teniendo en cuenta posibles variaciones en la detección del pico. Por otro lado, la extensión posterior al pico R garantiza la inclusión del segmento ST y de la onda T, que presentan una duración mayor y contienen información relevante sobre la repolarización ventricular.

De este modo, la ventana definida permite capturar la morfología completa del latido, asegurando un equilibrio entre información fisiológica y consistencia en la segmentación, lo que resulta adecuado para su posterior análisis y clasificación mediante técnicas de aprendizaje automático.

En la Figura 5.13 se muestran superpuestos los latidos segmentados a partir de la derivación I del registro 00010, lo que permite visualizar la consistencia en la segmentación y la variabilidad morfológica entre los distintos latidos del mismo paciente. Los latidos están correctamente alineados en torno al pico R, lo que confirma la fiabilidad del método de detección y segmentación empleado.

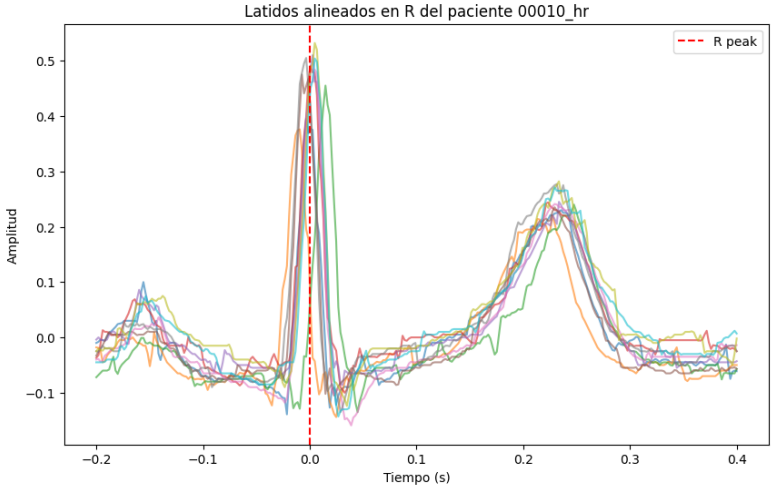

Figura 5.13: Latidos cardíacos segmentados.

Para mantener la homogeneidad en la representación de los datos, a la vez que se detectan los picos R, se eliminan aquellos registros que contengan menos de 10 latidos detectados, ya que se considera que no proporcionan información suficiente para el análisis. De los registros restantes, se seleccionan únicamente los 10 primeros latidos para cada paciente, con el objetivo de limitar la cantidad de datos y evitar un posible sesgo derivado de la variabilidad en el número de latidos entre pacientes.

Por último, en función del escenario experimental considerado, se procede a la construcción de las entradas del modelo. En el caso de utilizar únicamente la derivación I, se genera un megavector mediante la concatenación de los latidos segmentados de cada paciente. De la misma forma, se extraen características temporales, morfológicas y estadísticas de cada latido, con el objetivo de obtener una representación más compacta y discriminativa de la señal.

También se adapta la etiqueta asociada a cada paciente en función del escenario experimental. Para ello, la formulación original del problema se transforma, ya sea mediante la agrupación de patologías en superclases o manteniendo el enfoque multilabel probabilístico, en el que cada patología se representa mediante su probabilidad de ocurren-

cia.

5.4. Selección y extracción de características

Aunque la señal ECG contiene una gran cantidad de información fisiológica, no toda resulta igualmente relevante para la tarea de clasificación. Además, trabajar directamente con la señal completa implica manejar vectores de gran dimensionalidad, lo que puede incrementar la complejidad computacional y dificultar el aprendizaje de determinados modelos.

Por este motivo, se realiza una extracción explícita de características sobre los latidos previamente segmentados. El objetivo es obtener una representación compacta de la señal que preserve la información más relevante para la identificación de patologías cardíacas.

Las características seleccionadas pueden agruparse en cuatro categorías principales:

- **Características temporales:** incluyen medidas relacionadas con la variabilidad de los intervalos cardíacos.
- **Características morfológicas:** asociadas a la forma y estructura de los latidos.
- **Características estadísticas:** derivadas de la distribución de los valores de la señal.
- **Características de frecuencia:** derivadas de la descomposición wavelet.

Adicionalmente, se incluye una estimación del eje eléctrico del corazón como variable global del paciente.

5.4.1. Características temporales

Las características temporales se obtienen a partir de los intervalos RR, definidos como el tiempo transcurrido entre dos picos R consecu-

tivos. Constituyen una de las medidas fundamentales en el análisis del ritmo cardíaco, ya que reflejan directamente la regularidad de los latidos y permiten caracterizar su comportamiento a lo largo del tiempo.

Los intervalos RR se calculan a partir de la secuencia de picos R detectados previamente en cada registro. A partir de esta secuencia se construyen diferentes medidas que permiten resumir tanto el comportamiento global del ritmo cardíaco como su variabilidad. Esta característica resulta útil como indicador general del estado del sistema cardíaco.

En primer lugar, se calcula la media de los intervalos RR, que proporciona una estimación global de la duración media de los ciclos cardíacos del paciente. Esta medida es especialmente relevante ya que está directamente relacionada con la frecuencia cardíaca media, que se obtiene posteriormente mediante la transformación de los intervalos RR a latidos por minuto. Esta variable es uno de los indicadores fisiológicos más utilizados en cardiología, ya que permite una interpretación directa del ritmo cardíaco.

De forma complementaria, se calcula la desviación estándar de los intervalos RR, que cuantifica la variabilidad global del ritmo. Esta medida permite evaluar hasta qué punto los intervalos entre latidos se mantienen constantes o presentan fluctuaciones significativas, lo cual puede ser indicativo de alteraciones en la regulación del ritmo cardíaco.

Además de estas medidas globales, se incorporan también los valores extremos de la distribución de intervalos RR, es decir, el intervalo RR mínimo y el intervalo RR máximo observados en cada registro. Estas medidas permiten capturar episodios puntuales de aceleración o desaceleración del ritmo cardíaco, que podrían no ser reflejados adecuadamente por medidas promedio.

A partir de estos valores se calcula el rango de los intervalos RR, definido como la diferencia entre el valor máximo y el valor mínimo. Esta característica proporciona una medida directa de la amplitud total de variación del ritmo cardíaco durante el registro.

Por último, se incluye el índice RMSSD (*Root Mean Square of Suc-*

cessive Differences), que se calcula como la raíz cuadrada de la media de las diferencias al cuadrado entre intervalos RR consecutivos. Esta métrica es ampliamente utilizada en estudios de variabilidad cardíaca, ya que es especialmente sensible a variaciones de corto plazo entre latidos consecutivos y permite caracterizar de forma más fina la irregularidad del ritmo.

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2}$$

En conjunto, estas características permiten describir tanto el ritmo cardíaco medio como su variabilidad temporal, proporcionando una representación robusta del comportamiento dinámico del corazón.

5.4.2. Características morfológicas

Las características morfológicas tienen como objetivo describir la forma de cada latido cardíaco, capturando información relacionada con la estructura de las distintas ondas que componen el ciclo cardíaco. Este tipo de información es especialmente relevante en electrocardiografía, ya que muchas patologías no se reflejan únicamente en cambios del ritmo, sino también en deformaciones de la forma del latido. Permiten analizar cómo es cada latido de forma individual, evaluando posibles alteraciones en la despolarización y repolarización ventricular.

En primer lugar, se considera la amplitud del pico R, que representa el punto de máxima activación ventricular. Esta medida está relacionada con la intensidad de la despolarización del ventrículo, y puede verse alterada en presencia de cambios estructurales o eléctricos en el corazón.

De la misma forma, se calcula la amplitud de la onda T, asociada al proceso de repolarización ventricular. Variaciones en esta onda pueden estar relacionadas con alteraciones en la recuperación eléctrica del miocardio o con procesos isquémicos.

De forma complementaria, se estima la amplitud mínima previa al pico R, utilizada como aproximación a la onda Q. Esta componente permite analizar la fase inicial del complejo ventricular, siendo relevante en la detección de posibles alteraciones asociadas a daño previo en el tejido cardíaco.

El complejo QRS se caracteriza además mediante su área, que representa la energía total del latido dentro de esa ventana, y mediante su duración, que aproxima el tiempo de activación ventricular. Ambas medidas son especialmente relevantes, ya que alteraciones en la conducción eléctrica del corazón suelen reflejarse directamente en el ensanchamiento o modificación del complejo QRS.

Finalmente, se calcula el nivel medio del segmento ST, una región clave del ciclo cardíaco situada entre la despolarización y la repolarización ventricular. Variaciones en este segmento están directamente asociadas a procesos isquémicos y constituyen uno de los indicadores clínicos más importantes en cardiología.

Para todas estas características se calculan posteriormente la media y la desviación estándar a nivel de paciente, para capturar tanto el valor típico de cada medida como su variabilidad entre latidos, obteniendo así una representación más robusta de la morfología cardíaca.

5.4.3. Características estadísticas

Las características estadísticas permiten describir la señal desde una perspectiva global, sin depender explícitamente de su interpretación fisiológica. Complementan la información temporal y morfológica, aportando una representación más general del comportamiento de la señal ECG.

Para cada latido se calcula la media y la desviación estándar de la señal, que permiten caracterizar el nivel medio de amplitud y su dispersión. Se incluyen la amplitud máxima y mínima, que permiten identificar los valores extremos de la señal dentro de cada latido. Así, se consigue una representación más robusta y menos sensible a varia-

ciones puntuales o posibles errores de segmentación. Adicionalmente, se incorporan medidas estadísticas de orden superior como la asimetría (*skewness*) y la curtosis (*kurtosis*).

En primer lugar, se calcula la media de la señal, que proporciona información sobre el nivel medio de la amplitud del latido, lo que puede ser indicativo de la intensidad general de la actividad eléctrica cardíaca. Pequeñas variaciones pueden estar relacionadas con cambios en la morfología de la señal.

La desviación estándar permite cuantificar la dispersión de los valores de amplitud de la señal. Un valor elevado de esta medida indica una mayor variabilidad en la amplitud del latido, lo que podría estar asociado a alteraciones en la conducción eléctrica o a la presencia de ruido en el registro, mientras que un valor bajo sugiere una señal más homogénea. Esta característica resulta muy útil para diferenciar señales con distinta complejidad morfológica, ya que latidos con formas más irregulares suelen presentar una mayor dispersión de valores.

Los valores máximo y mínimo de la señal permiten caracterizar los extremos de amplitud de cada latido y complementan la información obtenida mediante las características morfológicas, proporcionando una visión más completa de la forma del latido. Estas medidas pueden ser especialmente relevantes en la detección de alteraciones isquémicas o de daño miocárdico, que suelen manifestarse mediante cambios en la amplitud de las ondas del ECG.

Además, se calcula la asimetría (skewness) de la distribución de amplitudes. Esta medida permite evaluar si los valores de la señal se distribuyen de forma simétrica alrededor de su media o si existe una mayor concentración de valores hacia amplitudes positivas o negativas. Cambios en la forma del complejo QRS o en la presencia de ondas más pronunciadas pueden reflejarse en esta característica.

Por último, se incorpora la curtosis (kurtosis), que describe el grado de concentración de los valores alrededor de la media y la presencia de picos pronunciados en la distribución. Esta medida permite diferenciar señales con morfologías más suaves de aquellas que presentan cambios

bruscos o componentes más agudos.

Dado que cada registro tiene múltiples latidos, todas estas medidas se agregan posteriormente mediante su media y desviación estándar, lo que permite obtener una representación global estable y menos sensible a variaciones puntuales entre latidos.

En conjunto, estas características aportan una descripción global de la señal ECG y complementan la información proporcionada por las características temporales y morfológicas. Aunque individualmente pueden parecer menos interpretables desde un punto de vista clínico, resultan especialmente útiles para los algoritmos de aprendizaje automático, ya que permiten capturar patrones generales presentes en la señal que no siempre son evidentes mediante el análisis visual.

5.4.4. Características de frecuencia

Las características frecuenciales permiten analizar la señal en el dominio tiempo-frecuencia, capturando información que no es directamente observable en el dominio temporal. Se emplea la transformada wavelet discreta, ya que es especialmente adecuada para señales no estacionarias como el ECG.

Para obtener estas características se utiliza de nuevo la transformada wavelet discreta utilizando la wavelet db6 y una profundidad de 8 niveles. A partir de los coeficientes obtenidos se calcula la energía asociada a diferentes bandas de frecuencia.

En concreto, se consideran los coeficientes d4, d5 y d6, que se corresponden con diferentes rangos de frecuencia asociados a la actividad eléctrica cardíaca. La energía se calcula como la suma del cuadrado de los coeficientes, lo que proporciona una medida de la potencia de la señal en cada banda frecuencial.

Las energías correspondientes a d4 y d5 contienen gran parte de la información asociada al complejo QRS. Es por esto que estas características proporcionan una medida indirecta de la intensidad y complejidad de los procesos de despolarización ventricular presentes en cada

latido.

Por su parte, la energía del coeficiente d6 se asocia a componentes de frecuencia más bajas, que pueden estar relacionadas con la actividad de las ondas P y T, así como con posibles alteraciones en la repolarización ventricular. Permite capturar variaciones más lentas de la señal que no siempre se reflejan en las bandas dominadas por el complejo QRS.

Al igual que en el resto de características, las energías calculadas para cada latido se agregan posteriormente a nivel de paciente mediante su media y desviación estándar. La media permite caracterizar el comportamiento frecuencial típico del paciente, mientras que la desviación estándar proporciona información sobre la estabilidad o variabilidad de dicho comportamiento entre latidos.

Finalmente, se calcula una relación entre las energías de la banda de frecuencia más alta y la banda de frecuencia más baja. Esta medida permite describir cómo se distribuye la energía de la señal entre componentes de frecuencia más alta y más baja. Dos señales pueden presentar una morfología aparentemente similar en el dominio temporal, pero diferir significativamente en su composición espectral. Por este motivo, esta característica aporta información complementaria que puede resultar útil para mejorar la capacidad discriminativa de los modelos de clasificación.

5.4.5. Estimación del eje eléctrico

Finalmente, se incluye una estimación del eje eléctrico cardíaco, obtenido a partir de las derivaciones I y aVF del electrocardiograma. Este eje representa la dirección predominante de la actividad eléctrica durante la despolarización ventricular y constituye un indicador clínico relevante en la evaluación de patologías cardíacas.

Para su estimación, se calcula la amplitud media del complejo QRS en ambas derivaciones, considerando la diferencia entre los valores máximos y mínimos en la región del complejo. A partir de estas amplitudes se obtiene el ángulo del eje eléctrico mediante una relación trigono-

métrica basada en la función `arctan2`, lo que permite determinar la dirección del vector eléctrico resultante.

Desviaciones significativas del eje eléctrico pueden estar asociadas a alteraciones estructurales del corazón, bloqueos de conducción o hipertrofias ventriculares, por lo que esta variable aporta información global complementaria a las características previamente descritas.

El conjunto de características extraídas permite construir una representación multidimensional de la señal electrocardiográfica, integrando información temporal, morfológica, estadística y frecuencial. Esta combinación es fundamental para proporcionar a los modelos de clasificación una base informativa capaz de capturar patrones complejos asociados a diferentes patologías cardiovasculares.

5.5. Desarrollo experimental y optimización del sistema

En esta sección se describe todo el proceso experimental llevado a cabo para la construcción y optimización del sistema de clasificación automática de patologías cardíacas a partir de señales ECG. Dado que el problema trata una tarea compleja de clasificación multilabel y multietiqueta y probabilística en algunos casos, se ha seguido un enfoque iterativo basado en la evolución progresiva de distintos modelos y configuraciones.

El objetivo principal de esta fase no es únicamente obtener un modelo final con un rendimiento competitivo, sino analizar en profundidad el comportamiento de diferentes algoritmos bajo distintos planteamientos del problema, con el fin de identificar la estrategia más adecuada en términos de capacidad predictiva y estabilidad del sistema.

Para ello, el desarrollo experimental se ha estructurado en tres escenarios principales. En primer lugar, un enfoque de clasificación binaria utilizando una única derivación del ECG, en el que el problema se redefine mediante la agrupación de patologías en superclases. Así, el

objetivo consiste en la clasificación binaria de patologías agrupadas y la clase normal, lo que permite evaluar la capacidad del modelo para diferenciar entre registros normales y patológicos, sin entrar en la especificidad de cada diagnóstico individual. Cada superclase se modela de forma independiente como una variable binaria, permitiendo la coexistencia de múltiples condiciones en un mismo paciente.

En segundo lugar, se ha considerado también un enfoque de clasificación binaria basado en la utilización de las 12 derivaciones, con el objetivo de evaluar el impacto de incorporar información espacial más completa del registro electrocardiográfico.

Finalmente, se ha abordado un tercer escenario más complejo, basado en las 12 derivaciones, orientado a la clasificación probabilística de las 71 patologías, donde cada paciente puede presentar simultáneamente múltiples diagnósticos, lo que incrementa significativamente la dificultad del problema debido al desbalance de clases y a la naturaleza multietiqueta del diagnóstico.

Adicionalmente, en los dos primeros escenarios de clasificación binaria, se han explorado distintas estrategias de decisión mediante la modificación de los umbrales de clasificación, con el objetivo de ajustar el equilibrio entre sensibilidad y especificidad del sistema. En concreto, se han evaluado tanto umbrales inferiores al valor por defecto como configuraciones de umbral fijo personalizdo a cada superclase, con el fin de analizar su impacto en la capacidad del modelo para detectar correctamente la presencia de patología. Este ajuste resulta especialmente relevante en aplicaciones clínicas, donde los falsos negativos tienen un impacto crítico.

En conjunto, este proceso experimental permite comparar de forma sistemática diferentes formulaciones del problema y analizar cómo la complejidad del escenario afecta al comportamiento de los modelos y a la capacidad de generalización del sistema.

5.5.1. Modelos de clasificación binaria

En el escenario 1 y 2 se utiliza un enfoque de clasificación binaria, por lo que se han evaluado distintos modelos de aprendizaje supervisado con el objetivo de analizar su capacidad para diferenciar la presencia o ausencia de patologías agrupadas en superclases. Trabajan sobre dos configuraciones de entrada distintas: la primera basada en la utilización de una única derivación del ECG, mientras que la segunda incorpora las 12 derivaciones disponibles.

En este trabajo los modelos se han adaptado explícitamente a un problema multietiqueta en el que un mismo paciente puede activar simultáneamente varias superclases. Esto condiciona tanto el diseño de los clasificadores como la interpretación de sus salidas, que se basan en probabilidades independientes por etiqueta.

Regresión logística

La regresión logística es un modelo de clasificación lineal muy utilizado en problemas biomédicos debido a su simplicidad, interpretabilidad y estabilidad. (Lera et al., 2025) Se basa en modelar la probabilidad de pertenencia a una clase mediante una combinación lineal de las características de entrada, seguida de la aplicación de una función sigmoide que acota la salida entre 0 y 1. (Lee, 2025)

Este modelo se usa como referencia base, ya que permite establecer un punto de comparación frente a métodos más complejos. Su principal ventaja es la interpretabilidad de sus coeficientes, lo que facilita entender que características del ECG influyen en la predicción de cada superclase.

En este problema, la regresión logística se adapta a un escenario multietiqueta mediante un enfoque One-vs-Rest. Consiste en descomponer el problema original, que involucra múltiples clases, en varios problemas de clasificación binaria independientes, uno por cada superclase. Cada modelo aprende modelos específicos asociados a una condición concreta sin verse condicionado por la presencia de otras pa-

tologías, lo que permite una mayor flexibilidad en la interpretación de los resultados. (Scikit-learn developers, s.f.-e)

Por otro lado, se ha incorporado el parámetro `class_weight = "balanced"`, que ajusta automáticamente el peso de cada clase de acuerdo con su frecuencia en el conjunto de entrenamiento. Este mecanismo es especialmente relevante en el contexto de datos clínicos, donde existe un fuerte desbalance entre clases, como se ve en la Figura 5.6, ya que permite mejorar la capacidad del modelo para detectar clases minoritarias sin sacrificar excesivamente la precisión en las clases mayoritarias.

Sin este ajuste, el modelo tendería a favorecer las clases mayoritarias, obteniendo buenos resultados globales pero con un rendimiento pobre en las patologías minoritarias, que son precisamente las de mayor interés clínico. La ponderación de clases permite mitigar este efecto, penalizando en mayor medida los errores cometidos sobre clases infrarrepresentadas y favoreciendo un comportamiento más equilibrado del modelo.

Random Forest

Random Forest es un algoritmo de clasificación basado en la construcción de múltiples árboles de decisión. La idea principal consiste en entrenar un conjunto de árboles de decisión independientes sobre diferentes subconjuntos aleatorios de los datos y combinar posteriormente sus predicciones. De esta forma, en lugar de depender de un único árbol, que podría producir overfitting, el modelo final se beneficia de la diversidad entre los árboles, lo que mejora su capacidad de generalización. (IBM, s.f.)

Una de las principales ventajas del Random Forest es su capacidad para modelar relaciones complejas y no lineales entre las variables de entrada y las clases objetivo. Esta característica resulta especialmente interesante en el análisis de señales ECG, donde las relaciones entre las características temporales, morfológicas y frecuenciales y la presencia

de una determinada patología suelen ser altamente complejas y difíciles de describir mediante modelos lineales. (Zou et al., 2022)

A diferencia de la regresión logística, que requiere una adaptación explícita para manejar problemas multietiqueta, Random Forest admite directamente múltiples variables objetivo. El modelo recibe directamente la matriz de etiquetas de entrenamiento, donde cada columna corresponde a una superclase diferente. De esta forma, el clasificador aprende simultáneamente a predecir la presencia o ausencia de cada una de ellas, permitiendo que un mismo paciente pueda presentar varias patologías de forma simultánea.

Se implementa un bosque de 300 árboles, lo que proporciona un buen equilibrio entre capacidad de modelado y eficiencia computacional. Inicialmente, se realizaron pruebas con un número menor de árboles, pero se observó que el rendimiento del modelo mejoraba significativamente al aumentar la cantidad de árboles, hasta alcanzar un punto de saturación a partir de aproximadamente 300 árboles, donde las mejoras adicionales eran marginales.

También se limitó la profundidad máxima de cada árbol a 10 niveles, con el objetivo de evitar el sobreajuste. Esta restricción impide que los árboles memorizen los patrones del conjunto de entrenamiento, lo que podría comprometer su capacidad de generalización a nuevos datos.

Se aumentó el número mínimo de muestras por hoja hasta 10. Este parámetro impide que los árboles generen ramas muy específicas para casos individuales, lo que reduce la influencia de ruido y mejora el comportamiento sobre el conjunto de prueba.

También se utilizó la configuración `max_features = "sqrt"`, que limita el número de características consideradas en cada división a la raíz cuadrada del total de características disponibles. Esta estrategia introduce una mayor aleatoriedad en la construcción de los árboles, lo que contribuye a mejorar la diversidad del conjunto y, por tanto, su capacidad de generalización.

Por otro lado, se incorporó el parámetro `class_weight="balanced"` para compensar el desbalance entre clases, de forma similar al plan-

teamiento de la regresión logística. Si todas las muestras recibieran el mismo peso durante el entrenamiento, el modelo tendería a favorecer las clases mayoritarias, obteniendo aparentemente buenos resultados globales pero un rendimiento deficiente en patologías menos frecuentes. La opción `balanced` ajusta automáticamente el peso de cada clase de forma inversamente proporcional a su frecuencia, penalizando más los errores cometidos sobre las clases minoritarias y favoreciendo un aprendizaje más equilibrado.

Una característica especialmente relevante de Random Forest es su capacidad para proporcionar probabilidades de pertenencia a cada clase. En lugar de generar únicamente una decisión binaria, cada árbol emite una predicción y el conjunto del bosque permite estimar la probabilidad de que una muestra pertenezca a una determinada categoría.

En conjunto, Random Forest constituye un modelo especialmente adecuado para este trabajo debido a su capacidad para capturar relaciones no lineales, su robustez frente al ruido, su tolerancia a datos heterogéneos y su facilidad para generar estimaciones probabilísticas que posteriormente pueden ajustarse mediante estrategias de decisión específicas para cada patología. Estas características lo convierten en una alternativa muy competitiva para el análisis automático de señales electrocardiográficas y la detección de patologías cardíacas.

Red Neuronal Multicapa

Las redes neuronales artificiales son modelos de aprendizaje supervisado inspirados en la estructura y funcionamiento del cerebro humano. Constituyen una de las técnicas más utilizadas para modelar relaciones complejas entre variables de entrada y clases objetivo, especialmente en problemas con alta dimensionalidad y no linealidad, como es el caso del análisis de señales ECG. (Lee, Fangfang, 2021)

En el ámbito de señales electrocardiográficas, esta capacidad es especialmente relevante debido a la complejidad de las relaciones entre las características extraídas y la presencia de diferentes patologías, que

a menudo no pueden ser capturadas adecuadamente mediante modelos lineales o basados en reglas.

En este trabajo se ha implementado una red neuronal multicapa (MLP) completamente conectada, diseñada para abordar el problema multietiqueta. A diferencia de los modelos anteriores, donde cada etiqueta se predice de forma relativamente independiente, la red neuronal aprende una representación conjunta de los datos que posteriormente se utiliza para estimar la probabilidad de pertenencia a cada superclase de forma simultánea, lo que permite capturar interacciones entre las diferentes patologías presentes en un mismo paciente.

La arquitectura de la red está compuesta por tres capas ocultas de tamaño decreciente (256, 128 y 64 neuronas respectivamente) seguida de una capa de salida con 5 neuronas, cada una correspondiente a una superclase diferente. Esta estructura permite reducir de forma progresiva la dimensionalidad de la información, extrayendo características cada vez más abstractas y relevantes para la tarea de clasificación.

Esta arquitectura decreciente se ha seleccionado para evitar modelos excesivamente complejos que podrían sobreajustar el conjunto de entrenamiento. Durante las primeras fases del desarrollo se evaluaron configuraciones con capas de mayor tamaño, observándose una mayor tendencia al sobreajuste. Por este motivo se optó por una arquitectura más compacta, capaz de mantener una buena capacidad de representación sin comprometer la generalización.

La función de activación utilizada en las capas ocultas es ReLU (*Rectified Linear Unit*), una de las más utilizadas en redes neuronales modernas debido a su simplicidad computacional y a su capacidad para mitigar problemas asociados al desvanecimiento del gradiente durante el entrenamiento. Gracias a esta función, la red puede modelar relaciones no lineales complejas entre las características de entrada y las patologías objetivo. (Codificando Bits, s.f.)

Para el entrenamiento del modelo se utiliza la función de pérdida `BCEWithLogitsLoss`, especialmente adecuada en problemas de clasificación multietiqueta, ya que combina en una única operación la función

sigmoide y la entropía cruzada binaria. Esta formulación permite una mayor estabilidad numérica durante el entrenamiento en comparación con la aplicación separada de ambas funciones. (PyTorch contributors, s.f.)

Al igual que en los modelos anteriores, se incorpora el ajuste de pesos para compensar el desbalance entre clases. En este caso, se utiliza el parámetro `pos_weight`, que asigna un peso específico a las clases positivas en función de su frecuencia relativa en el conjunto de entrenamiento. Este peso se calcula a partir de la proporción entre muestras positivas y negativas en el conjunto de entrenamiento, de forma que las clases menos frecuentes tengan una mayor penalización cuando son clasificadas incorrectamente. Este ajuste resulta especialmente relevante en este problema, ya que sin él el modelo tendería a favorecer las clases mayoritarias, degradando el rendimiento en patologías minoritarias. (PyTorch contributors, s.f.)

Para controlar el sobreajuste durante el entrenamiento, se reserva un 20 % del conjunto de entrenamiento como validación independiente, utilizada únicamente para monitorizar el rendimiento del modelo. Esta partición permite evaluar la capacidad de generalización durante el entrenamiento y ajustar la complejidad de los modelos cuando es necesario. De este modo, se evita que el modelo se optimice exclusivamente sobre el conjunto de entrenamiento, favoreciendo un comportamiento más estable en datos no vistos. (Durán, 2019)

Una vez entrenado el modelo, la salida de la red consiste en una probabilidad independiente por cada superclase, obtenida mediante la función sigmoide aplicada a la capa de salida. Estas probabilidades no se transforman directamente en decisiones mediante un umbral fijo estándar, sino que se utiliza un conjunto de umbrales específicos por clase.

Este diseño permite ajustar de forma independiente la sensibilidad de cada patología, lo cual resulta especialmente relevante en aplicaciones médicas donde el coste de un falso negativo es significativamente mayor que el de un falso positivo. De este modo, el sistema prioriza la

detección de patologías incluso a costa de incrementar ligeramente la tasa de falsas alarmas.

5.5.2. Estrategias de decisión y ajuste de umbrales

Los modelos desarrollados en este trabajo no generan directamente una predicción binaria, sino una probabilidad asociada a cada una de las superclases consideradas. Estas probabilidades deben transformarse en decisiones finales de presencia o ausencia de cada patología mediante un proceso de binarización.

Este proceso se realiza mediante la aplicación de un umbral de decisión, que determina a partir de qué valor de probabilidad se considera que una clase está presente (valor 1) o ausente (valor 0). De forma general, este umbral se aplica sobre cada una de las salidas del modelo tras la función sigmoide (en el caso de la red neuronal) o sobre las probabilidades estimadas por los distintos clasificadores.

Inicialmente, se utilizó el valor estándar de 0.5 como umbral de referencia, que es el criterio habitual en problemas de clasificación binaria. Sin embargo, este enfoque mostró limitaciones en el contexto del problema abordado, especialmente debido al desbalance entre clases y a la importancia clínica de reducir los falsos negativos.

Posteriormente, se probó una estrategia más sensible consistente en reducir de forma uniforme el umbral a 0.3 para todas las clases. Esta modificación incrementó la sensibilidad del sistema, favoreciendo la detección de patologías, aunque también introdujo un aumento de falsos positivos.

Finalmente, se optó por una estrategia más ajustada basada en umbrales específicos por superclase, definidos manualmente en función de la relevancia clínica de cada patología y del impacto relativo de los falsos negativos y falsos positivos. Este enfoque permite adaptar el comportamiento del modelo a la criticidad de cada enfermedad, incorporando conocimiento del dominio en la fase de decisión.

Los umbrales definidos son los siguientes:

- **NORM (0.6):** se emplea un umbral más alto con el objetivo de reducir falsos positivos en la clase de normalidad, evitando clasificar pacientes enfermos como sanos.
- **STTC (0.4):** asociado a alteraciones del segmento ST/T, relacionadas con isquemia o infarto. Se utiliza un valor intermedio que equilibra sensibilidad y precisión.
- **CD (0.3):** asociado con alteraciones de conducción, con un umbral equilibrado debido a su detectabilidad relativamente estable.
- **MI (0.25):** indica infarto de miocardio, considerado la patología más crítica del sistema, por lo que se prioriza la sensibilidad reduciendo el umbral para minimizar falsos negativos.
- **HYP (0.35):** asociado con hipertrofia, con relevancia clínica intermedia, manteniendo un compromiso entre precisión y recall.

La aplicación de estos umbrales se realiza de forma homogénea en todos los modelos del sistema, garantizando un criterio de decisión consistente entre los distintos enfoques evaluados. De este modo, las diferencias de rendimiento entre modelos no dependen del proceso de binarización, sino de su capacidad de aprendizaje.

Desde un punto de vista clínico, esta estrategia permite controlar explícitamente el equilibrio entre sensibilidad y especificidad, priorizando la detección de patologías relevantes incluso a costa de incrementar ligeramente los falsos positivos, lo cual es aceptable en este tipo de sistemas de apoyo al diagnóstico.

5.5.3. Modelos de regresión probabilística

En este tercer escenario experimental se aborda el problema desde una perspectiva de regresión multisalida, en la que el objetivo no es realizar una decisión binaria sobre la presencia o ausencia de una patología, sino estimar un valor continuo entre 0 y 1 para cada una de

las 71 enfermedades consideradas. Este valor se interpreta como una medida de probabilidad o grado de asociación entre el registro ECG del paciente y cada patología.

Este enfoque resulta especialmente adecuado en problemas médicos multietiqueta, ya que en la práctica clínica un paciente puede presentar simultáneamente varias patologías con distinto grado de severidad, lo que convierte el problema en una tarea de estimación continua más que de clasificación estricta.

La estructura de los modelos implementados es la misma: son modelos de regresión multisalida que aprenden simultáneamente a predecir un vector de 71 dimensiones, donde cada componente representa una patología.

Dado que cada patología tiene una distribución y dificultad de detección distinta, las salidas del modelo se recortan al rango $[0, 1]$ mediante una operación de clipping, asegurando así una interpretación coherente como probabilidad.

Posteriormente, el rendimiento se evalúa mediante métricas de regresión y análisis específicos diseñados para variables desbalanceadas, como el error medio absoluto (MAE), el sesgo de predicción o la distribución de errores por clase.

Ridge Regression

La regresión Ridge es una extensión de la regresión lineal clásica que incorpora un término de regularización L2 con el objetivo de reducir el sobreajuste y mejorar la capacidad de generalización del modelo. Esta regularización penaliza la magnitud de los coeficientes aprendidos durante el entrenamiento, evitando que el modelo dependa en exceso de variables concretas o de posibles correlaciones entre características. De este modo, se obtiene un modelo más estable y robusto frente al ruido presente en los datos. (Murel y Kavlakoglu, s.f.-a)

Este tipo de modelo es especialmente relevante debido al gran número de variables que tiene cada paciente. Por este motivo, la regula-

rización es importante, ya que ayuda a que el modelo generalice mejor y no se ajuste en exceso a los datos de entrenamiento. En este caso, en lugar de realizar una clasificación directa, el modelo estima para cada paciente un vector de 71 valores entre 0 y 1, donde cada uno representa la probabilidad de que ese paciente tenga una determinada enfermedad, interpretando la salida como una estimación de riesgo en lugar de una decisión fija.

Dado que Ridge está pensado para problemas con una sola salida, se ha utilizado `MultiOutputRegressor`, que permite entrenar un modelo independiente por cada una de las 71 enfermedades. De esta forma, el sistema final está formado por 71 modelos Ridge en paralelo, cada uno especializado en una patología.

En la implementación realizada se ha utilizado un valor de regularización $\alpha = 10$. Este valor determina cuánto se penalizan los coeficientes del modelo: valores bajos permiten mayor flexibilidad pero aumentan el riesgo de sobreajuste, mientras que valores altos hacen el modelo más simple y estable, aunque con menor capacidad de ajuste. (Scikit-learn developers, s.f.-f) En pruebas iniciales se probaron valores más bajos, pero estos provocaban cierto sobreajuste, por lo que se decidió aumentar el valor. Este ajuste mejora la estabilidad del modelo y su comportamiento en datos no vistos.

Finalmente, las predicciones se recortan al rango [0, 1], ya que la regresión lineal puede generar valores fuera de ese intervalo. Este paso permite mantener una interpretación coherente como probabilidades.

En conjunto, Ridge se utiliza como un modelo base sencillo e interpretable dentro del enfoque de regresión probabilística, aunque su capacidad para capturar relaciones complejas es limitada.

ElasticNet

Elastic Net es una extensión de la regresión lineal que combina dos tipos de regularización: L1 (Lasso) y L2 (Ridge). Esto significa que, además de reducir el tamaño de los coeficientes para evitar el

sobreajuste, también puede llegar a eliminar algunos de ellos, lo que ayuda a seleccionar automáticamente las variables más relevantes. Por este motivo, es un modelo especialmente útil cuando existen muchas características correlacionadas o poco informativas, como ocurre en este problema. (Interactive Chaos, s.f.)

En este trabajo, este modelo se aplica al mismo escenario de regresión probabilística que Ridge, donde el objetivo es estimar para cada paciente un vector de 71 valores entre 0 y 1, interpretados como la probabilidad de presencia de cada patología. De esta forma, el modelo no realiza una clasificación directa, sino una estimación continua del riesgo asociado a cada enfermedad.

Al igual que en Ridge, se utiliza MultiOutputRegressor, ya que Elastic Net solo permite predecir una variable objetivo a la vez. Esto implica que se entrenan 71 modelos independientes en paralelo, uno por cada patología.

En la implementación se han utilizado los parámetros $\alpha = 0{,}001$ y l1_ratio = 0.5. El parámetro α controla la intensidad de la regularización, mientras que `l1_ratio` determina el equilibrio entre L1 y L2: valores cercanos a 0 se comportan como Ridge, y valores cercanos a 1 como Lasso. (Scikit-learn developers, s.f.-a) En este caso se ha elegido un valor intermedio para combinar estabilidad y capacidad de selección de variables, manteniendo un equilibrio entre simplicidad y ajuste del modelo.

Finalmente, al igual que en el resto de modelos de regresión, las predicciones se limitan al intervalo [0, 1] para mantener su interpretación como probabilidades.

En conjunto, Elastic Net se utiliza como una alternativa más flexible a Ridge, ya que permite no solo regularizar el modelo, sino también reducir la influencia de variables menos relevantes, lo que puede mejorar el rendimiento en datos de alta dimensionalidad.

Random Forest Regressor

Este modelo es una adaptación de los bosques aleatorios al ámbito de regresión. Al igual que en el caso de clasificación, se basa en la combinación de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y de las variables de entrada. La predicción final se obtiene promediando las predicciones generadas por todos los árboles del bosque, lo que permite reducir la varianza del modelo y obtener resultados más estables y robustos. (GeeksforGeeks, 2026)

A diferencia de los modelos lineales anteriores, Random Forest es capaz de capturar relaciones no lineales y patrones complejos entre las características extraídas del ECG y las patologías asociadas. Esta capacidad es reseñable en un problema como este, donde la relación entre las características fisiológicas y las enfermedades cardíacas no siempre sigue un comportamiento lineal. (C, 2025)

En este escenario de regresión probabilística, el modelo estima para cada paciente un vector de 71 valores continuos entre 0 y 1, uno por cada patología. Estas salidas no deben interpretarse como probabilidades clínicas directas, sino como una medida de similitud o grado de asociación entre el perfil del paciente y los patrones observados en el conjunto de entrenamiento. En otras palabras, el modelo estima qué tan frecuente es cada patología en pacientes con características similares.

Para su implementación se ha utilizado un bosque de 300 árboles (`n_estimators = 300`), buscando un equilibrio entre rendimiento y coste computacional. Inicialmente se probó una configuración con un número mayor de árboles, llegando hasta 800, con el objetivo de mejorar la estabilidad del modelo. Sin embargo, esta configuración no pudo evaluarse completamente debido al elevado tiempo de entrenamiento requerido, por lo que se optó finalmente por un número intermedio de árboles que permitiera completar el análisis de forma viable.

Además, se ha limitado la profundidad máxima de los árboles a 10 niveles (`max_depth = 10`) para evitar sobreajuste y controlar la com-

plejidad del modelo. También se ha fijado un mínimo de 10 muestras por hoja (`min_samples_leaf = 10`), lo que ayuda a suavizar las predicciones y reducir la sensibilidad al ruido presente en los datos. Por otro lado, se utiliza `max_features = "sqrt"`, lo que implica que cada árbol considera únicamente una selección aleatoria de variables en cada división, reduciendo la correlación entre árboles y mejorando la capacidad de generalización del modelo.

Además, en la fase de diseño se planteó el uso del criterio de división `squared_error`, con la intención de mejorar la precisión en la reducción del error cuadrático. Sin embargo, esta configuración incrementaba de forma notable el tiempo de ejecución, por lo que se descartó en favor de la configuración por defecto, que ofrecía un compromiso más adecuado entre rendimiento y coste computacional.

Finalmente, las predicciones se recortan al intervalo $[0, 1]$ para mantener una interpretación coherente como medida de asociación o probabilidad relativa de cada patología.

En conjunto, este modelo se utiliza como referencia dentro de los enfoques no lineales en el escenario de regresión probabilística, permitiendo evaluar el comportamiento del sistema frente a modelos más simples.

HistGradientBoosting Regressor

El HistGradientBoosting Regressor es una versión actualizada y optimizada del algoritmo de *Gradient Boosting*, que construye el modelo mediante una combinación de múltiples árboles débiles. En cada iteración, el modelo intenta corregir los errores cometidos por el conjunto de árboles anteriores, lo que permite mejorar progresivamente la precisión de las predicciones. A diferencia del Random Forest, donde los árboles se entrenan de forma independiente, en este modelo existe una dependencia secuencial entre ellos. (Picard, 2024)

Este enfoque es especialmente potente en problemas con relaciones no lineales y estructuras complejas en los datos, como ocurre en

el análisis de señales ECG. Además, la versión basada en histogramas implementa optimizaciones que reducen el coste computacional al agrupar los valores continuos en bins, lo que acelera el entrenamiento sin una pérdida significativa de rendimiento. (Riaux, 2024)

En este trabajo, el modelo se aplica al escenario de regresión probabilística, donde se predice un vector de 71 valores entre 0 y 1, uno por cada patología. Estas salidas representan el grado de asociación entre las características del paciente y la presencia de cada enfermedad, más que una probabilidad clínica estricta.

Para su implementación se ha utilizado una configuración controlada con `max_depth = 6`, con el objetivo de limitar la complejidad de los árboles base y reducir el riesgo de sobreajuste. El parámetro `learning_rate = 0.05` regula la contribución de cada árbol al modelo final, permitiendo un aprendizaje más estable y progresivo. Asimismo, se ha fijado `max_iter = 300`, que determina el número de iteraciones del proceso de boosting. (Scikit-learn developers, s.f.-b)

Dado el desbalance presente en las distintas patologías, se ha establecido `min_samples_leaf = 20`, lo que ayuda a evitar divisiones demasiado específicas en clases poco representadas. Además, se ha incluido una regularización L2 con `l2_regularization = 0.1`, con el objetivo de mejorar la estabilidad del modelo y reducir el sobreajuste. Al igual que en los modelos anteriores, se utiliza `MultiOutputRegressor` para poder entrenar de forma independiente un modelo por cada una de las 71 patologías.

Finalmente, las predicciones se recortan al intervalo $[0, 1]$ para mantener una interpretación coherente como medida de asociación entre el paciente y cada patología.

En conjunto, este modelo representa una alternativa más avanzada dentro de los enfoques de regresión no lineal, permitiendo comparar el rendimiento del boosting frente a métodos como Random Forest o regresión lineal regularizada.

Red Neuronal

En este caso se emplea una red neuronal multicapa con el objetivo de obtener una salida continua por cada una de las 71 patologías, interpretada como un valor de asociación entre el paciente y cada enfermedad. A diferencia del enfoque de clasificación, no se aplica posteriormente un umbral fijo para la toma de decisiones, sino que se trabajan directamente las salidas continuas del modelo.

La arquitectura utilizada es equivalente a la empleada en el modelo de clasificación, con tres capas totalmente conectadas de tamaño decreciente (256, 128 y 64 neuronas), junto con ReLU en las capas intermedias y técnicas de *Batch Normalization* y *Dropout* para estabilizar el entrenamiento y reducir el sobreajuste.

El entrenamiento se realiza mediante `BCEWithLogitsLoss`, trabajando directamente con los logits generados por la red, lo que permite una formulación numéricamente estable para problemas multisalida. El optimizador utilizado es Adam, complementado con un esquema de reducción adaptativa del learning rate mediante `ReduceLROnPlateau`.

Para mejorar la estabilidad del entrenamiento, se calcula una ponderación basada en el desbalance entre clases a partir de la proporción de ejemplos positivos y negativos por etiqueta. Sin embargo, este peso no se aplica directamente en la función de pérdida.

Finalmente, las salidas del modelo se obtienen aplicando la función sigmoide sobre los logits, obteniéndose valores en el rango [0,1] que se interpretan como una medida continua de probabilidad o asociación.

5.5.4. Configuración 1

En este escenario se trabajó con la configuración más sencilla, utilizando la señal en formato reducido con una única derivación y un enfoque de clasificación binaria por superclases. A nivel de implementación, esta configuración se utilizó principalmente como punto de partida para validar el correcto funcionamiento de los modelos antes de abordar el problema más complejo.

Como entradas del sistema se utilizaron tanto el megavector de la señal en bruto como las características extraídas a partir del mismo. Esto permitió comparar el comportamiento de los modelos con diferentes representaciones de los datos.

Durante el desarrollo experimental los modelos utilizados fueron Regresión Logística, Random Forest y una Red Neuronal Multicapa. Esto permitió establecer una primera referencia de rendimiento y detectar limitaciones asociadas a la simplicidad del espacio de características.

En cuanto a la decisión final de clasificación, se utilizaron las tres opciones de umbralización explicadas en 5.5.2, con el objetivo de analizar su impacto en el equilibrio entre sensibilidad y precisión.

5.5.5. Configuración 2

Esta segunda opción introduce un incremento en la complejidad del problema mediante el uso de las 12 derivaciones completas del ECG, pero manteniendo el enfoque de clasificación por superclases.

Desde el punto de vista de la implementación, el principal cambio respecto a la Configuración 1 fue el aumento masivo del volumen de información disponible. El uso de las señales completas de las 12 derivaciones multiplicó considerablemente el tamaño de los datos de entrada, haciendo inviable el entrenamiento de los modelos utilizando directamente la señal bruta debido al elevado coste computacional asociado. Por este motivo, se optó por trabajar exclusivamente con las características extraídas de cada latido, reduciendo drásticamente la dimensionalidad del problema sin perder la información más relevante desde el punto de vista clínico.

Esta configuración permitió evaluar el impacto que tiene la incorporación de información procedente de todas las derivaciones del ECG sobre la capacidad predictiva de los modelos, manteniendo constante la estructura del problema de clasificación. Para ello se emplearon los mismos algoritmos utilizados en la configuración anterior así como la misma estrategia de ajuste de umbrales.

5.5.6. Configuración 3

La tercera configuración representa el escenario más exigente de todo el trabajo. En este caso se trabaja directamente con las 71 patologías originales, manteniendo además la naturaleza multietiqueta del problema.

Al igual que en la configuración anterior, únicamente se emplean las características extraídas de las 12 derivaciones del ECG, ya que el volumen de información asociado a la señal bruta completa resulta excesivamente elevado para los recursos computacionales disponibles.

La principal diferencia respecto a las configuraciones anteriores está en la salida del sistema. Mientras que en los escenarios de clasificación cada modelo debía determinar la presencia o ausencia de una serie de patologías mediante decisiones binarias, en esta configuración los modelos generan directamente valores continuos comprendidos entre 0 y 1 para cada una de las 71 patologías. Estos valores pueden interpretarse como una medida del grado de asociación entre el paciente analizado y cada una de las patologías consideradas.

Debido a este cambio de enfoque, fue necesario emplear modelos de regresión capaces de generar salidas continuas. Inicialmente se implementaron Ridge Regression, Random Forest Regressor y una Red Neuronal Multicapa, que sirvieron como referencia para evaluar la viabilidad del enfoque probabilístico. Tras analizar sus resultados, se incorporaron Elastic Net e HistGradientBoosting Regressor con el objetivo de explorar técnicas de regularización adicionales y modelos capaces de capturar relaciones no lineales más complejas entre las características extraídas y las patologías. De esta forma, fue posible comparar aproximaciones lineales, basadas en árboles y basadas en redes neuronales dentro de un mismo escenario experimental.

Capítulo 6

Análisis de Resultados

Una vez realizado el desarrollo experimental del sistema, es necesario analizar los resultados obtenidos. El objetivo principal es evaluar el rendimiento de los distintos modelos implementados según las distintas configuraciones planteadas, así como estudiar su comportamiento en función del tipo de enfoque utilizado.

El análisis empieza con una visión global de los resultados hasta un nivel detallado por modelo y patología, lo que permite comparar de forma estructurada las distintas alternativas propuestas.

En primer lugar, se estudia el impacto de las configuraciones experimentales definidas previamente, así como el efecto del ajuste de umbrales en la toma de decisiones del sistema. Seguidamente, se incluye un análisis por patología y una discusión de los resultados desde un punto de vista clínico, junto con las principales limitaciones del sistema desarrollado.

6.1. Resultados globales

Con el objetivo de evaluar el rendimiento de los modelos desarrollados bajo las distintas configuraciones propuestas, se presentan a continuación los resultados obtenidos en la fase experimental. El análisis se centra en comparar el comportamiento de los diferentes enfoques de modelado, tanto en clasificación como en regresión probabilística, así

como en estudiar el impacto de la representación de los datos y de la complejidad del problema sobre el rendimiento final.

Los resultados se organizan de forma progresiva, desde los escenarios más simples hasta los más complejos, lo que permite interpretar de manera más clara la influencia de cada decisión de diseño en la capacidad predictiva del sistema. Se incluyen comparaciones globales entre modelos y métricas representativas que permiten obtener una visión general del comportamiento de cada enfoque.

6.1.1. Configuración 1: Derivación I y clasificación binaria

La primera configuración corresponde con al escenario más sencillo del sistema, en el que se utiliza únicamente la derivación I del ECG junto con un enfoque de clasificación binaria por superclases. Permite validar el funcionamiento general de los modelos y establece una referencia inicial de rendimiento antes de aumentar la complejidad del problema.

En esta fase se emplearon dos representaciones distintas de los datos de entrada: la señal en bruto mediante un megavector y un conjunto de características extraídas a partir de los latidos. Esto permite analizar el impacto de la representación de la información en el rendimiento de los modelos.

Los mejores resultados de esta configuración se obtuvieron al utilizar una estrategia de umbralización personalizada por superclase, lo que permitió ajustar de forma independiente el punto de decisión de cada categoría y mejorar el equilibrio entre sensibilidad y precisión.

En la Tabla 6.1 se observan los resultados obtenidos del conjunto de test de estos tres modelos con cada una de las entradas. El promedio utilizado es tipo macro.

Modelo	Entrada	Hamming Acc.	Precision	Recall	F1-score	AUC global
Regresión Logística	Megavector	0.7215	0.4064	0.7789	0.5074	0.7656
Regresión Logística	Features	0.7426	0.4153	0.8417	0.5330	0.8091
Random Forest	Megavector	0.7684	0.4665	0.6931	0.4956	0.8011
Random Forest	Features	0.7714	0.4767	0.7011	0.5013	0.8153
MLP	Megavector	0.8127	0.7677	0.3327	0.4051	0.8137
MLP	Features	0.8121	0.8060	0.3020	0.3610	0.8296

Tabla 6.1: Resultados globales en test de la configuración 1

Aunque la Hamming Accuracy proporciona una visión global del número de etiquetas correctamente clasificadas, en este problema resulta especialmente relevante analizar métricas como Precision, Recall y F1-score macro, ya que permiten evaluar el comportamiento del modelo de forma equilibrada entre superclases independientemente de su frecuencia de aparición.

Era esperable que la Hamming Accuracy presentase valores superiores al F1-macro debido al desbalance entre superclases. La accuracy se ve influida principalmente por las clases mayoritarias, mientras que el F1-macro otorga el mismo peso a todas las clases, penalizando más los errores en las minoritarias.

En términos generales, se observa que la Red Neuronal Multicapa (MLP) es el modelo que alcanza los mejores valores de accuracy, independientemente del tipo de entrada utilizado. Este comportamiento refleja una mayor capacidad del modelo para capturar relaciones no lineales complejas en los datos, lo que se traduce en un mayor número de predicciones correctas a nivel global.

Sin embargo, este mejor rendimiento en accuracy no se refleja de forma equivalente en el equilibrio entre clases. En particular, los valores de F1 macro obtenidos por la red neuronal son más bajos que los de otros modelos, lo que indica una tendencia a favorecer las clases mayoritarias en detrimento de aquellas menos representadas. Este comportamiento es especialmente relevante en el contexto del problema, ya que implica una menor sensibilidad en la detección de determinadas

patologías.

Por otro lado, Random Forest presenta un comportamiento más equilibrado entre las distintas métricas. Aunque su accuracy es inferior al de la red neuronal, mantiene valores más consistentes en términos de F1 macro, lo que sugiere una mejor capacidad para manejar el desbalance entre clases. Este equilibrio lo convierte en un modelo más robusto desde el punto de vista clínico, al no centrarse únicamente en el rendimiento global.

En el caso de la regresión logística, se observa el rendimiento más limitado de los tres modelos, aunque se aprecia una mejora consistente cuando se utilizan características extraídas en lugar del megavector de la señal. Este resultado indica que una representación más estructurada de la señal ECG facilita la separación entre clases, aunque no es suficiente para alcanzar el rendimiento de los modelos no lineales.

En cuanto a la capacidad discriminativa, medida mediante el AUC global, todos los modelos alcanzan valores superiores a 0.75, lo que indica una capacidad razonable para diferenciar entre las distintas superclases. Los mejores resultados corresponden a la red neuronal multicapa utilizando características extraídas, seguida de Random Forest con características y la regresión logística con características. Estos valores muestran que los modelos son capaces de ordenar correctamente las probabilidades asociadas a las distintas clases incluso antes de aplicar los umbrales de decisión finales.

Respecto al tipo de entrada, se observa una tendencia general a una ligera mejora en los resultados cuando se emplean características extraídas. Este efecto es más notable en la red neuronal multicapa, mientras que en Random Forest y regresión logística es más moderado. Esto sugiere que la reducción de dimensionalidad y la extracción de información relevante contribuyen a mejorar la capacidad predictiva del sistema, especialmente en modelos con mayor capacidad de aprendizaje.

A continuación, en la Tabla 6.2 se muestra la diferencia entre el conjunto de train y test en Hamming Accuracy para poder comprobar

si el modelo presenta sobreajuste.

Modelo	Entrada	Accuracy Train	Accuracy Test	Diferencia
Logistic Regression	Megavector	0.7769	0.7215	0.0554
Logistic Regression	Features	0.7479	0.7426	0.0053
Random Forest	Megavector	0.8187	0.7684	0.0503
Random Forest	Features	0.8090	0.7714	0.0376
MLP	Megavector	0.8558	0.8127	0.0431
MLP	Features	0.8234	0.8121	0.0113

Tabla 6.2: Comparación del rendimiento entre train y test en la configuración 1

Desde el punto de vista de la generalización, el análisis de la diferencia entre el rendimiento en entrenamiento y test muestra que no existe un sobreajuste significativo en ninguno de los modelos evaluados. La regresión logística presenta prácticamente el mismo rendimiento en ambos conjuntos, mientras que Random Forest muestra una ligera variación en función del tipo de entrada, aunque dentro de márgenes aceptables. La red neuronal multicapa destaca por presentar diferencias muy reducidas entre entrenamiento y test, lo que indica una buena capacidad de generalización, favorecida por el uso de técnicas como dropout y batch normalization.

En conjunto, los resultados de esta configuración muestran que no existe un único modelo claramente superior en todas las métricas analizadas. Mientras que la red neuronal multicapa obtiene los mejores resultados en términos de Hamming Accuracy y AUC global, la regresión logística y Random Forest presentan un comportamiento más equilibrado entre superclases, reflejado en valores superiores de F1-score macro. El uso de características extraídas mejora de forma consistente el rendimiento de todos los modelos respecto al uso directo del megavector de la señal ECG, confirmando la utilidad de la etapa de extracción de características para este problema de clasificación.

6.1.2. Configuración 2: 12 derivaciones y clasificación binaria

La segunda configuración introduce un incremento en la complejidad del problema mediante el uso de las 12 derivaciones completas del ECG. Este cambio permite disponer de una mejor representación de la señal cardíaca, lo que se traduce en una mejora general del rendimiento de los modelos en comparación con la configuración anterior.

Como se ha comentado con anterioridad, debido al elevado coste computacional asociado al uso de un megavector que contenga las doce derivaciones, se ha optado por utilizar únicamente como entrada del sistema las características extraídas de la señal ECG.

En este contexto, se mantiene la misma estrategia experimental que en la configuración anterior, utilizando los mismos modelos de clasificación y el mismo esquema de evaluación, lo que permite una comparación directa del efecto producido por el incremento de información disponible.

La Tabla 6.3 recoge los resultados globales obtenidos sobre el conjunto de test para los distintos modelos evaluados en la Configuración 2. Además de la Hamming accuracy, se incluyen métricas de precisión, recall y F1-score calculadas tanto mediante promedio macro como micro. A diferencia de la Configuración 1, donde el análisis se centró principalmente en comparar el comportamiento general de los modelos y se complementará posteriormente con un estudio detallado por superclases, en esta configuración se incorporan también las métricas micro para obtener una visión más completa del rendimiento global del sistema. Mientras que las métricas macro evalúan el comportamiento medio sobre cada superclase otorgando la misma importancia a todas ellas, las métricas micro agregan conjuntamente todas las predicciones realizadas y reflejan mejor el rendimiento global del modelo teniendo en cuenta la distribución real de las clases. El uso combinado de ambas perspectivas permite analizar simultáneamente la capacidad de detección equilibrada entre patologías y el comportamiento global del

sistema en condiciones más próximas a un escenario real de utilización.

Modelo	Hamming Acc.	Prec. Macro	Rec. Macro	F1 Macro	Prec. Micro	Rec. Micro	F1 Micro
Logistic Regression	0.8324	0.5409	0.8660	0.6513	0.5501	0.8699	0.6740
Random Forest	0.8320	0.5693	0.7768	0.6122	0.5310	0.7458	0.6203
MLP	0.8496	0.6345	0.8059	0.7091	0.6633	0.8296	0.7372

Tabla 6.3: Resultados globales en test de la configuración 2

La incorporación de las doce derivaciones completas del ECG produce una mejora significativa del rendimiento respecto a la Configuración 1 en los tres modelos evaluados. Este resultado confirma que la información adicional contenida en las distintas derivaciones aporta características relevantes para la identificación de las superclases consideradas, permitiendo una representación más completa de la actividad eléctrica cardíaca.

Analizando los resultados globales, la red neuronal multicapa obtiene la mayor Hamming accuracy, superando claramente a la regresión logística y a Random Forest. Esto indica que es el modelo que mejor clasifica las etiquetas individuales, lo que sugiere una mejor capacidad para explotar la información adicional disponible en las doce derivaciones.

En términos de precisión, la red neuronal multicapa presenta también los mejores resultados tanto en la métrica macro como en la micro . Estos valores reflejan que sus predicciones son considerablemente más fiables que las del resto de modelos, generando un menor número de falsos positivos. Sin embargo, esta mejora en precisión viene acompañada de una reducción del recall, que alcanza únicamente 0.5342 en la métrica macro y 0.5977 en la micro. Esto indica que el modelo adopta una estrategia más conservadora, priorizando la confianza de las predicciones frente a la detección exhaustiva de todas las patologías presentes.

Por el contrario, la regresión logística presenta el comportamiento opuesto. Sus valores de recall son los más elevados de los tres modelos.

Esto demuestra una gran capacidad para detectar casos positivos, aunque a costa de una reducción significativa de la precisión, que apenas supera el 0.53 en ambas métricas. En consecuencia, el modelo consigue identificar una mayor proporción de patologías presentes, pero genera también un número mayor de falsos positivos.

Random Forest se sitúa en una posición intermedia entre ambos enfoques. Aunque no alcanza los mejores resultados en ninguna de las métricas analizadas, presenta un compromiso razonable entre precisión y recall, obteniendo valores más equilibrados que los observados en la regresión logística y la red neuronal. No obstante, este equilibrio no resulta suficiente para superar a ninguno de los otros modelos en términos de F1-score.

La comparación de los valores de F1-score permite analizar conjuntamente precisión y recall. En términos macro, la regresión logística obtiene el mejor resultado, ligeramente por encima de la red neuronal multicapa y de Random Forest. Esto indica que, considerando todas las superclases con la misma importancia, la regresión logística mantiene el mejor equilibrio global entre sensibilidad y precisión. Sin embargo, cuando se considera la distribución real de las clases mediante el F1 micro, la red neuronal multicapa pasa a ocupar la primera posición, seguida de la regresión logística y Random Forest. Este resultado sugiere que la red neuronal aprovecha mejor la información disponible en los casos más representativos del conjunto de datos y ofrece el mejor rendimiento global del sistema.

La Tabla 6.4 recoge la Hamming accuracy obtenida en entrenamiento y test para cada modelo, junto con la diferencia entre ambas. Esta comparación permite evaluar la capacidad de generalización de los modelos y analizar la posible presencia de sobreajuste.

Modelo	Acc. Train	Acc. Test	Diferencia
Logistic Regression	0.8432	0.8324	0.0108
Random Forest	0.8652	0.8320	0.0332
MLP	0.9360	0.8496	0.0864

Tabla 6.4: Comparación del rendimiento entre train y test en la Configuración 2

En cuanto a la generalización de los modelos, los tres presentan una diferencia muy pequeña entre la hamming accuracy de train y test, lo que indica una ausencia de problemas graves de sobreajuste. La regresión logística muestra el menor gap, confirmando su elevada capacidad de generalización y la simplicidad de su estructura. La red neuronal multicapa presenta un gap de 0.0259, manteniendo un buen equilibrio entre rendimiento y capacidad de generalización debido a las técnicas de regularización empleadas durante el entrenamiento. Por su parte, Random Forest es el modelo que presenta la mayor diferencia entre entrenamiento y test, aunque este valor sigue siendo suficientemente bajo como para considerar que el modelo generaliza adecuadamente sobre datos no vistos.

En conjunto, los resultados obtenidos muestran que el incremento de información proporcionado por las doce derivaciones beneficia especialmente a la red neuronal multicapa, que consigue el mejor rendimiento global en términos de Hamming accuracy y F1 micro. No obstante, la regresión logística continúa destacando por su elevada capacidad de detección, obteniendo los mejores valores de recall y F1 macro. Estos resultados ponen de manifiesto que la elección del modelo óptimo depende en gran medida del objetivo clínico: maximizar la detección de patologías o reducir el número de falsas alarmas manteniendo un elevado rendimiento global.

6.1.3. Configuración 3: 12 derivaciones y clasificación probabilística

La tercera configuración corresponde con el escenario más exigente. En este caso, se trabaja con las 71 patologías originales y las 12 derivaciones completas del ECG.

Debido a la elevada dimensionalidad de los datos, los modelos se entrenan exclusivamente a partir de las características extraídas de las doce derivaciones del ECG. Este escenario permite evaluar el comportamiento de los distintos enfoques en condiciones más próximas a un entorno clínico real, donde la complejidad diagnóstica es considerablemente mayor y las patologías presentan una distribución altamente desbalanceada.

En esta configuración se analizan los modelos planteados inicialmente (Ridge Regression, Random Forest Regressor y Red Neuronal Multicapa) y los modelos incorporados durante el desarrollo experimental (ElasticNet y HistGradientBoosting Regressor), con el objetivo de identificar las arquitecturas más adecuadas para abordar un problema de clasificación multietiqueta de alta complejidad.

Dado que el objetivo es evaluar la calidad de las predicciones probabilísticas y no únicamente las etiquetas finales, no se emplea la métrica de accuracy, ya que esta no resulta representativa en escenarios multietiqueta altamente desbalanceados. En su lugar, se utilizan métricas más informativas como el error absoluto medio (MAE), el error cuadrático medio (MSE), la log-loss, el coeficiente de correlación de Spearman y el AUC global.

Las métricas incluidas en la Tabla 6.5 se han seleccionado para ofrecer una evaluación completa del rendimiento de los modelos en un contexto de clasificación multietiqueta con salida probabilística. El MAE y el MSE permiten cuantificar el error entre las probabilidades predichas y los valores reales, proporcionando una medida global de precisión, siendo el MSE más sensible a errores grandes.

Por su parte, la log-loss evalúa la calidad de las probabilidades pre-

dichas, penalizando especialmente las predicciones incorrectas con alta confianza, lo que la hace adecuada para problemas probabilísticos.

Finalmente, el AUC global mide la capacidad del modelo para discriminar entre clases de forma independiente del umbral, siendo especialmente relevante en un escenario altamente desbalanceado, donde la accuracy no resulta representativa.

Modelo	**MAE**	**MSE**	**Log-loss**	**AUC global**
Ridge Regression	0.0253	0.0097	0.0392	0.8808
ElasticNet Regression	0.0242	0.0098	0.0373	0.8887
Random Forest Regressor	0.0229	0.0098	0.0358	0.9002
HistGradientBoosting	0.0188	0.0085	0.0330	0.8509
MLP Regressor	0.0140	0.0112	0.0558	0.8835

Tabla 6.5: Resultados globales en test de la configuración 3

En primer lugar, el HistGradientBoosting obtiene el mejor rendimiento en términos de error global, con el menor MAE y el menor MSE, lo que indica una buena capacidad para aproximar las probabilidades reales. No obstante, este bajo error debe interpretarse con cuidado, ya que en un problema altamente desbalanceado como el planteado, donde la clase mayoritaria corresponde a la condición normal, es posible que el modelo esté ajustando mejor las probabilidades en dicha clase dominante, lo que reduce el error global sin implicar necesariamente un mejor rendimiento en la detección de patologías. Esto puede explicar en parte la menor AUC obtenida por este modelo, que refleja una menor capacidad discriminativa.

El Random Forest Regressor presenta el mejor AUC global, lo que lo convierte en el modelo con mayor capacidad para distinguir entre clases positivas y negativas. Además, mantiene valores de error bajos, lo que indica un equilibrio entre precisión en las probabilidades y capacidad de discriminación.

Los modelos lineales regularizados (Ridge y ElasticNet) muestran un comportamiento muy similar entre sí. Ambos presentan un rendimiento intermedio y errores ligeramente superiores a los modelos

basados en árboles, lo que refleja su menor capacidad para capturar relaciones no lineales en los datos.

Finalmente, la MLP Regressor destaca por obtener el menor MAE, lo que indica un ajuste muy preciso a nivel de predicción continua. Sin embargo, este resultado no se traduce en un mejor comportamiento global, ya que presenta el MSE más alto y una log-loss elevada, lo que sugiere cierta inestabilidad en las probabilidades estimadas. Su AUC se sitúa en un nivel intermedio, por debajo de Random Forest pero por encima de los modelos lineales.

En conjunto, se observa que los modelos basados en árboles ofrecen el mejor compromiso global entre capacidad discriminativa y estabilidad en la predicción, destacando especialmente Random Forest como el más equilibrado en este escenario de alta dimensionalidad y fuerte desbalanceo.

6.2. Análisis por modelos y enfermedades

Con el objetivo de complementar las métricas globales presentadas anteriormente, en este apartado se analiza el rendimiento de los distintos modelos a nivel de superclase. Para ello se estudian las métricas de precisión, recall y F1-score obtenidas para cada categoría, así como las matrices de confusión y las curvas ROC asociadas.

Dado que el problema presenta un importante desbalance entre clases y que las distintas patologías tienen una relevancia clínica diferente, no se emplea un único umbral de decisión global. En su lugar, se utilizan umbrales específicos para cada superclase, definidos según el criterio descrito en el apartado 5.5.2. Este enfoque permite adaptar el equilibrio entre sensibilidad y precisión a las características de cada categoría, priorizando la detección de aquellas patologías donde los falsos negativos pueden tener consecuencias clínicas más relevantes.

A continuación se analiza el comportamiento de cada modelo utilizando tanto la representación basada en el megavector de la señal ECG como el conjunto de características extraídas.

6.2.1. Configuración 1

Regresión Logística

La regresión logística constituye el modelo de referencia más sencillo evaluado en esta configuración. Debido a su naturaleza lineal, permite analizar hasta qué punto las distintas superclases pueden separarse mediante fronteras de decisión simples, así como evaluar el impacto de la representación de entrada sobre su capacidad discriminativa.

Megavector

Los resultados agrupados por superclase se muestran en la Tabla 6.6, donde se reportan las métricas de precisión, recall y F1-score para cada categoría en el conjunto de test tras la aplicación de los umbrales definidos

Clase	Precision	Recall	F1-score
NORM	0.52	0.60	0.56
STTC	0.67	0.85	0.75
CD	0.37	0.82	0.51
MI	0.17	0.84	0.28
HYP	0.31	0.78	0.44

Tabla 6.6: Resultados de regresión logística por clase (Megavector)

En conjunto, los resultados muestran un comportamiento heterogéneo entre clases, reflejando tanto la complejidad del problema como el efecto directo del ajuste de umbrales sobre el equilibrio entre precisión y exhaustividad.

Se observa un patrón general en el que el modelo tiende a favorecer el recall en la mayoría de las superclases, lo que indica una estrategia orientada a minimizar falsos negativos. Este comportamiento es coherente con el diseño del sistema, donde se prioriza la detección de patologías frente a la exactitud estricta de la predicción.

Este efecto se refleja directamente en el F1-score, que actúa como métrica de compromiso entre precisión y recall. En aquellas clases donde existe un desequilibrio claro entre ambas, el F1 se mantiene en

valores intermedios, penalizando precisamente esta falta de equilibrio. Por tanto, el F1 permite sintetizar en una única métrica el impacto real del ajuste de umbrales sobre el rendimiento del modelo.

La clase STTC destaca como la superclase con mejor equilibrio global entre precisión y recall y con un F1-score relativamente alto, lo que sugiere que el modelo es capaz de identificar de forma relativamente estable los patrones asociados a estas patologías Este comportamiento se ve reforzado por su alta capacidad discriminativa, como se refleja posteriormente en las curvas ROC (ver Figura 6.2).

En el caso de NORM, se observa un comportamiento relativamente equilibrado. El umbral utilizado para esta superclase es más restrictivo que en el resto debido a la importancia clínica de evitar que un paciente con alguna patología sea clasificado como sano. Como consecuencia, el modelo adopta una estrategia más conservadora al asignar la etiqueta NORM, mejorando la precisión a costa de una reducción moderada del recall. A pesar de ello, mantiene una capacidad razonable para identificar registros normales, obteniendo un equilibrio aceptable entre ambas métricas.

La clase CD muestra un comportamiento caracterizado por una sensibilidad elevada, detectando la mayoría de los casos positivos, y una precisión más reducida, lo que se traduce en un F1-score intermedio. Este patrón sugiere la presencia de solapamiento en la representación del ECG entre alteraciones de conducción y otras clases relacionadas.

Por otro lado, MI presenta uno de los comportamientos más relevantes desde el punto de vista clínico, con un recall elevado y una precisión sensiblemente menor. Este resultado es coherente con un enfoque conservador del modelo, orientado a priorizar la detección de posibles infartos, aunque ello implique un mayor número de falsos positivos.

Finalmente, En la clase HYP se observa un comportamiento marcado por un recall elevado acompañado de una precisión significativamente menor. Esto indica que el modelo es capaz de detectar la mayoría de los casos positivos, aunque genera un número elevado de falsos positivos, lo que sugiere solapamiento de patrones electrocardiográficos en

la representación en crudo del ECG y una mayor sensibilidad inducida por el ajuste del umbral.

Las matrices de confusión de la Figura 6.1 permiten analizar en detalle el comportamiento del modelo más allá de las métricas globales, descomponiendo los aciertos y errores en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Este análisis es especialmente relevante en un problema clínico, donde el tipo de error cometido tiene una importancia diferente según la patología.

En este caso, el ajuste de umbrales por clase influye directamente en la distribución de errores, modificando el equilibrio entre sensibilidad y precisión en cada superclase.

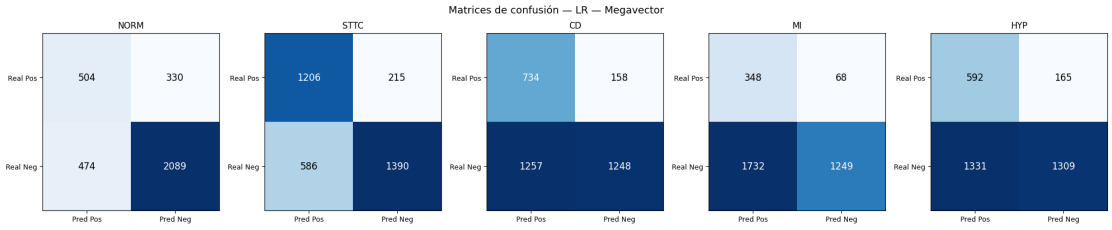

Figura 6.1: Matrices de confusión de regresión logística de la configuración 1 (Megavector)

En la clase NORM, el modelo presenta 504 verdaderos positivos, 2089 verdaderos negativos, 474 falsos positivos y 330 falsos negativos. A nivel global, se observa que la clase normal no es la más "limpia" del sistema, ya que acumula un número relativamente elevado de falsos positivos. Esto indica que una parte importante de registros que pertenecen a otras patologías están siendo clasificados como normales, lo cual es especialmente relevante desde el punto de vista clínico, ya que este tipo de error es el más crítico en esta clase. Sin embargo, el número de falsos positivos es menor respecto a configuraciones menos restrictivas, aunque aumenta ligeramente el número de falsos negativos. Desde el punto de vista clínico, este comportamiento es adecuado, ya que prioriza evitar falsos diagnósticos de normalidad.

En la clase STTC, el modelo muestra 1206 verdaderos positivos, 1390 verdaderos negativos, 586 falsos positivos y 215 falsos negativos. En este caso, STTC destaca como la segunda clase con mayor número de falsos negativos, lo que indica que una parte relevante de casos con alteraciones del segmento ST/T no está siendo detectada. Esto sugiere que, aunque la clase presenta una capacidad discriminativa razonable, existen patrones que el modelo no consigue capturar completamente, probablemente debido a la similitud con otras alteraciones de repolarización o a variabilidad en la morfología de la señal.

En la clase CD, se observan 734 verdaderos positivos, 1248 verdaderos negativos, 1257 falsos positivos y 158 falsos negativos. El comportamiento de esta clase está dominado por un número elevado de falsos positivos, lo que indica una tendencia del modelo a sobre-asignar esta etiqueta. Este fenómeno suele aparecer en clases con características intermedias o menos definidas, donde el modelo encuentra similitudes parciales con múltiples categorías. Como consecuencia, la precisión se ve penalizada, aunque la capacidad de detección (recall) se mantiene elevada.

En la clase MI, el modelo presenta 348 verdaderos positivos, 1249 verdaderos negativos, 1732 falsos positivos y 68 falsos negativos. Este es uno de los comportamientos más relevantes desde el punto de vista clínico, ya que el número de falsos negativos es relativamente bajo en comparación con el resto de clases. Esto indica que el modelo está priorizando la detección de posibles infartos, incluso a costa de generar un número muy elevado de falsos positivos. Este comportamiento es coherente con un enfoque conservador, donde es preferible sobre-detectar eventos críticos antes que omitirlos.

Finalmente, en la clase HYP se observan 592 verdaderos positivos, 1309 verdaderos negativos, 1331 falsos positivos y 165 falsos negativos. El comportamiento es intermedio, con una sensibilidad relativamente alta pero una precisión limitada. Esto refleja que el modelo detecta una proporción importante de casos positivos, pero aún presenta dificultades para distinguir correctamente esta clase frente a otras alteraciones

con patrones parcialmente similares.

En conjunto, las matrices de confusión muestran un patrón consistente en el que el modelo tiende a priorizar la sensibilidad frente a la precisión, especialmente en las clases patológicas. Este comportamiento es intencionado dentro del diseño del sistema, ya que en un contexto clínico resulta más crítico minimizar falsos negativos que reducir falsos positivos. No obstante, este enfoque incrementa la carga de falsos positivos en varias clases, especialmente en CD y MI, lo que evidencia el compromiso inherente entre sensibilidad y especificidad en este tipo de problemas.

Se observa que la clase NORM actúa como principal punto de conflicto del sistema, debido a la dificultad de separar completamente la normalidad del resto de alteraciones cuando se trabaja con señal ECG en crudo, sin una representación más estructurada de la información.

Las curvas ROC mostradas en la Figura 6.2 permiten evaluar la capacidad discriminativa del modelo independientemente del umbral de decisión fijado, analizando la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para cada superclase. Este análisis complementa las métricas basadas en umbral fijo, ya que permite valorar la calidad intrínseca de las probabilidades generadas por el modelo.

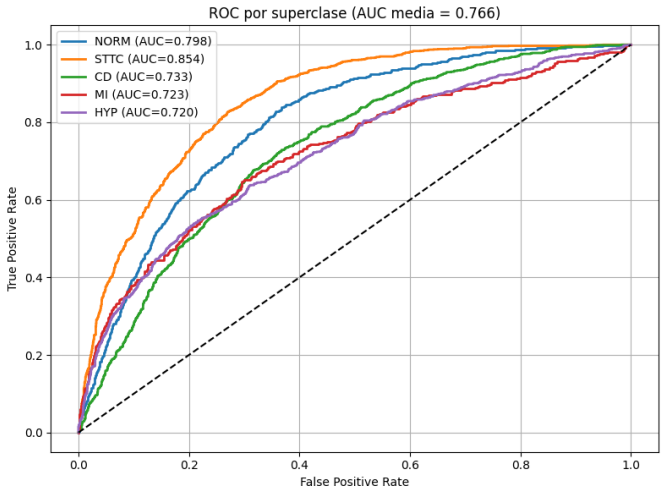

Figura 6.2: Curvas ROC de regresión logística de la configuración 1 (Megavector)

En términos generales, los resultados muestran un comportamiento razonablemente estable en la mayoría de las superclases, con valores de AUC superiores a 0.7 en todos los casos. La superclase STTC obtiene el mejor rendimiento, lo que indica una buena capacidad del modelo para diferenciar este tipo de alteraciones del resto de clases.

Por otro lado, CD presenta un AUC de 0.7330, mientras que MI y HYP obtienen valores similares, lo que sugiere una mayor dificultad del modelo para separar estas patologías en el espacio de probabilidad. NORM alcanza un AUC de 0.7983, reflejando una capacidad discriminativa intermedia, aunque condicionada por el solapamiento con determinadas clases patológicas.

Por todo esto, se confirma una capacidad discriminativa moderada del modelo sobre la representación en megavector, coherente con los resultados observados en las métricas por clase y en las matrices de confusión. Este comportamiento sugiere que, aunque el modelo es capaz de capturar patrones relevantes en la señal, existe todavía un grado significativo de solapamiento entre clases que limita su separación completa.

Características

Los resultados de la Tabla 6.7 muestran el rendimiento de la regresión logística utilizando como entrada las características extraídas de la señal ECG. En comparación con la representación basada en megavector, se observa un incremento generalizado del recall en prácticamente todas las superclases, lo que indica que esta representación facilita la detección de los casos positivos, aunque con un posible compromiso en la precisión en determinadas clases.

Clase	Precision	Recall	F1-score
NORM	0.54	0.69	0.61
STTC	0.67	0.88	0.76
CD	0.36	0.90	0.51
MI	0.17	0.94	0.28
HYP	0.34	0.82	0.48

Tabla 6.7: Resultados de regresión logística por clase (Características)

En la clase NORM se obtiene una precisión de 0.54, un recall de 0.69 y un F1-score de 0.61. Este comportamiento refleja una mejora en la capacidad del modelo para identificar correctamente registros normales respecto a la configuración anterior, aunque con un aumento moderado de falsos positivos, lo que sugiere una mayor sensibilidad del modelo a patrones cercanos a la normalidad.

La clase STTC presenta un rendimiento sólido. Este resultado indica que el modelo es capaz de detectar de forma consistente las alteraciones del segmento ST/T cuando se utilizan características estructuradas, lo que sugiere que este tipo de representación mejora la separabilidad de este tipo de patrones frente a la señal en crudo.

En la clase CD se observa un recall elevado acompañado de una precisión más limitada, lo que se traduce en un F1-score moderado. Este comportamiento indica que el modelo tiende a priorizar la detección de la mayoría de los casos positivos, aunque a costa de un incremento en los falsos positivos, lo que sugiere solapamiento entre esta clase y otras alteraciones con características similares.

En la clase MI destaca un recall muy alto frente a una precisión baja, lo que da lugar a un F1-score reducido. Este resultado refleja un comportamiento claramente conservador desde el punto de vista clínico, donde el modelo prioriza la detección de posibles infartos, minimizando los falsos negativos incluso a costa de una elevada tasa de falsos positivos. Este patrón es coherente con la relevancia clínica de esta patología.

Finalmente, la clase HYP presenta un comportamiento intermedio. Aunque el modelo consigue detectar la mayoría de los casos positivos, la precisión limitada indica una dificultad significativa para diferenciar esta clase de otras patologías con patrones electrocardiográficos similares.

En conjunto, estos resultados confirman que el uso de características extraídas mejora de forma consistente la sensibilidad del modelo respecto a la representación en megavector, especialmente en clases patológicas. Sin embargo, este incremento en recall viene acompañado de una reducción general en la precisión, lo que evidencia el compromiso habitual entre detección y tasa de falsas alarmas en este tipo de problemas desbalanceados.

Las matrices de confusión correspondientes a esta configuración, mostradas en la Figura 6.3, permiten analizar con mayor detalle el comportamiento del modelo en términos de aciertos y errores por superclase. En conjunto, se observa un patrón consistente con los resultados de las métricas anteriores, donde el modelo tiende a priorizar la detección de casos positivos a costa de un incremento en los falsos positivos en varias clases.

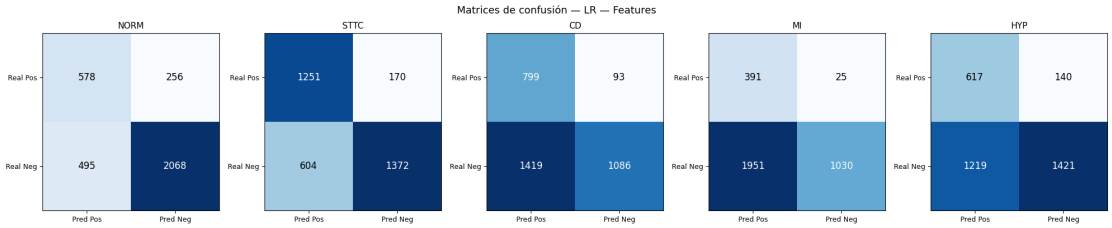

Figura 6.3: Matrices de confusión de regresión logística de la configuración 1 (Características)

En la clase NORM se obtienen 578 verdaderos positivos, 2068 verdaderos negativos, 495 falsos positivos y 256 falsos negativos. En este caso, se observa una mejora en la capacidad de detección de la clase normal en comparación con la representación en megavector, reflejada en la reducción de falsos negativos. Sin embargo, el número de falsos positivos se mantiene relativamente elevado, lo que indica que aún existe confusión entre registros normales y patológicos leves. Este comportamiento es coherente con la dificultad intrínseca de delimitar la clase NORM en presencia de alteraciones sutiles en la señal ECG.

En la clase STTC se observan 1251 verdaderos positivos, 1372 verdaderos negativos, 604 falsos positivos y 170 falsos negativos. Destaca especialmente la reducción de falsos negativos respecto a la configuración anterior, lo que indica una mejora en la capacidad del modelo para detectar alteraciones del segmento ST/T. Este resultado es consistente con el aumento del recall observado en las métricas por clase, sugiriendo que la representación mediante características facilita la identificación de este tipo de patrones.

En la clase CD, el modelo presenta 799 verdaderos positivos, 1086 verdaderos negativos, 1419 falsos positivos y 93 falsos negativos. Se observa una clara tendencia a la sobreasignación de esta clase, reflejada en el elevado número de falsos positivos. Esto sugiere que el modelo encuentra similitudes frecuentes entre CD y otras clases, lo que provoca una pérdida de precisión, aunque mantiene una alta sensibilidad.

En la clase MI se observa un comportamiento especialmente relevante desde el punto de vista clínico, con 391 verdaderos positivos y

únicamente 25 falsos negativos. Este resultado indica una capacidad muy elevada para detectar casos de infarto, priorizando claramente la sensibilidad del modelo. No obstante, esta mejora viene acompañada de un incremento significativo en falsos positivos, lo que refleja un comportamiento conservador en la toma de decisiones, coherente con la importancia crítica de esta patología.

Finalmente, la clase HYP presenta 617 verdaderos positivos, 1421 verdaderos negativos, 1219 falsos positivos y 140 falsos negativos. El comportamiento es intermedio, con una sensibilidad elevada pero una precisión limitada, lo que sugiere que el modelo sigue teniendo dificultades para diferenciar esta clase de otras alteraciones con patrones electrocardiográficos similares.

En la Figura 6.4 se observa el conjunto de curvas ROC para cada superclase en la configuración basada en características. De forma global, las curvas muestran un comportamiento consistente, situándose claramente por encima de la diagonal de no discriminación, lo que indica que el modelo es capaz de diferenciar de manera significativa entre clases positivas y negativas en todos los casos.

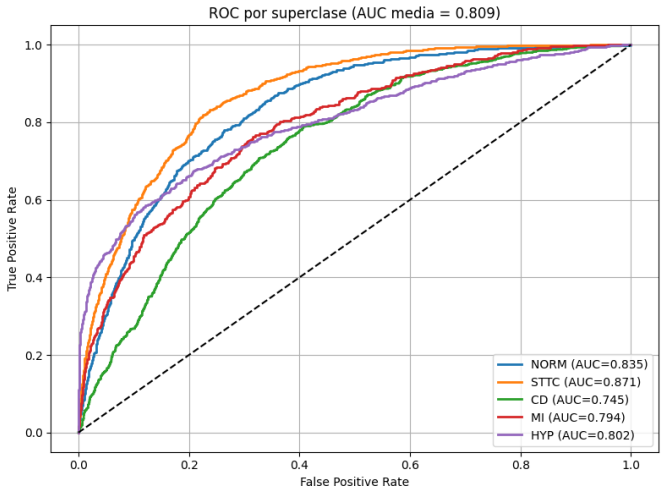

Figura 6.4: Curvas ROC de regresión logística de la configuración 1 (Características)

Este comportamiento se cuantifica mediante los valores de AUC por superclase, donde se observa un rendimiento especialmente alto en STTC y NORM, lo que indica una buena capacidad de separación entre estas clases y el resto del espacio de decisión. En el caso de STTC, este resultado es coherente con su buen comportamiento en las métricas de clasificación, mientras que en NORM refleja la capacidad del modelo para distinguir registros normales incluso en presencia de solapamiento con otras patologías.

Las clases MI y HYP presentan valores de AUC intermedios, lo que indica una capacidad discriminativa razonable, aunque con mayor dificultad para establecer fronteras claras entre clases debido a la similitud de patrones electrocardiográficos con otras alteraciones. A pesar de ello, el modelo mantiene una capacidad de ordenación de probabilidades adecuada, lo que justifica los buenos valores de recall observados previamente en estas clases.

Finalmente, la clase CD presenta el valor de AUC más bajo, lo que sugiere que es la superclase con mayor dificultad relativa de separación en este escenario. Este resultado es coherente con el comportamiento observado en las matrices de confusión, donde se aprecia una mayor tasa de falsos positivos, indicando solapamiento con otras categorías del conjunto.

En conjunto, los resultados confirman que el uso de características extraídas mejora de forma consistente la capacidad de detección del modelo, especialmente en clases patológicas, con una reducción clara de falsos negativos respecto a la representación en megavector. Este avance se acompaña de una disminución en la precisión y un aumento de falsos positivos, lo que refleja el compromiso habitual entre sensibilidad y especificidad en problemas clínicos desbalanceados.

Las matrices de confusión y las métricas globales muestran además una capacidad discriminativa general adecuada, coherente con la complejidad del ECG en crudo y el desbalance de clases. En conjunto, el comportamiento observado es estable y clínicamente razonable,

reforzando las conclusiones obtenidas a nivel por clase.

Random Forest

En este apartado se estudia el rendimiento del modelo Random Forest empleando tanto el megavector de la señal ECG como las características extraídas. El objetivo es analizar su capacidad para discriminar entre las distintas superclases y comparar el efecto de ambas representaciones sobre las métricas de clasificación, las matrices de confusión y las curvas ROC.

Megavector

Los resultados de la Tabla 6.8 muestran el rendimiento del modelo Random Forest utilizando como entrada el megavector de la señal ECG. A diferencia de la regresión logística, este modelo presenta un comportamiento más desigual entre precisión y recall según la superclase, lo que indica una estrategia de decisión más conservadora en ciertas clases y más restrictiva en otras.

Clase	Precision	Recall	F1-score
NORM	0.60	0.51	0.55
STTC	0.80	0.48	0.60
CD	0.38	0.80	0.52
MI	0.19	0.88	0.31
HYP	0.36	0.79	0.50

Tabla 6.8: Resultados de random forest por clase (Megavector)

En la clase NORM se obtiene una precisión de 0.60, un recall de 0.51 y un F1-score de 0.55. Este resultado refleja una mayor capacidad del modelo para evitar falsos positivos en comparación con la regresión logística, aunque a costa de una reducción en la sensibilidad. En este caso, el modelo es más estricto a la hora de asignar la etiqueta NORM, lo que reduce la probabilidad de clasificar como normal un caso que pueda presentar alteraciones.

La clase STTC presenta un comportamiento caracterizado por una precisión elevada pero un recall relativamente bajo, lo que da lugar a un F1-score de 0.60. Este patrón indica que el modelo es muy conservador a la hora de asignar esta clase, es decir, cuando predice STTC suele hacerlo con alta fiabilidad, pero deja sin detectar una parte importante de los casos positivos. Esto sugiere que el modelo encuentra patrones claros para esta clase, pero no logra generalizar bien todos los casos posibles dentro de la misma.

En la clase CD se observa un equilibrio más estable entre precisión y recall, lo que indica una tendencia del modelo a priorizar la detección de la mayoría de los casos positivos, aunque con un número considerable de falsos positivos. Este comportamiento es coherente con clases intermedias en complejidad, donde el modelo tiende a sobre-asignar la etiqueta para no perder sensibilidad.

La clase MI presenta un recall muy elevado frente a una precisión muy baja, lo que refleja un comportamiento claramente orientado a la detección de infartos, minimizando falsos negativos. Desde el punto de vista clínico, este comportamiento es coherente con la importancia crítica de esta patología, aunque reduce la fiabilidad de las predicciones positivas.

Finalmente, la clase HYP muestra un rendimiento intermedio. Esto indica que el modelo consigue detectar una proporción elevada de casos positivos, aunque con dificultades para discriminar correctamente esta clase frente a otras patologías con patrones electrocardiográficos similares.

La Figura 6.5 muestra las matrices de confusión por clase de random forest y entrada megavector. En conjunto, se observa un patrón más marcado de especialización del modelo, donde determinadas clases presentan una mejora clara en la reducción de falsos positivos, mientras que otras muestran una pérdida significativa de sensibilidad.

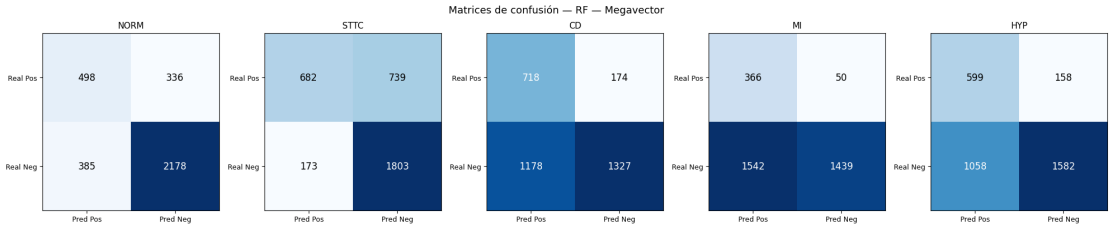

Figura 6.5: Matrices de confusión de random forest de la configuración 1 (Megavector)

En la clase NORM se obtienen 425 verdaderos positivos, 2283 verdaderos negativos, 280 falsos positivos y 409 falsos negativos. Este resultado muestra un comportamiento más conservador respecto a la asignación de la clase normal, reduciendo de forma notable los falsos positivos en comparación con la regresión logística. Sin embargo, esta mejora en precisión viene acompañada de un aumento de falsos negativos, lo que indica que el modelo tiende a clasificar como patológicos algunos registros que realmente son normales. En este caso, el modelo prioriza evitar falsos diagnósticos de normalidad, pero pierde sensibilidad en esta clase.

En la clase STTC se observan 682 verdaderos positivos, 1803 verdaderos negativos, 173 falsos positivos y 739 falsos negativos. Aquí se aprecia uno de los puntos débiles del modelo, ya que el número de falsos negativos es elevado, lo que implica que una proporción importante de alteraciones del segmento ST/T no está siendo detectada. Este comportamiento es coherente con el bajo recall observado en las métricas por clase y sugiere que el modelo no consigue capturar completamente la variabilidad de esta patología en el espacio del megavector.

En la clase CD se obtienen 718 verdaderos positivos, 1327 verdaderos negativos, 1178 falsos positivos y 174 falsos negativos. En este caso, el modelo presenta una tendencia clara a la sobrepredicción de esta clase, reflejada en el elevado número de falsos positivos. Esto indica que el modelo identifica patrones compatibles con CD en registros que pertenecen a otras superclases, lo que penaliza la precisión pero mantiene una sensibilidad relativamente alta.

En la clase MI se observa un comportamiento muy orientado a la detección de la patología, con 366 verdaderos positivos y únicamente 50 falsos negativos. Este resultado indica una alta sensibilidad en la detección de infartos, lo que es coherente con un comportamiento conservador desde el punto de vista clínico. Sin embargo, este rendimiento se acompaña de un número elevado de falsos positivos, lo que refleja una baja precisión en esta clase.

Finalmente, la clase HYP presenta 599 verdaderos positivos, 1582 verdaderos negativos, 1058 falsos positivos y 158 falsos negativos. El comportamiento es intermedio, con una sensibilidad relativamente alta pero con una proporción significativa de falsos positivos, lo que indica dificultades del modelo para diferenciar esta clase de otras alteraciones con patrones electrocardiográficos similares.

En conjunto, las matrices de confusión muestran que Random Forest tiende a mejorar la especificidad en comparación con modelos lineales en ciertas clases, especialmente NORM, pero introduce una mayor inestabilidad en la detección de clases como STTC, donde el incremento de falsos negativos es especialmente relevante. Este comportamiento refleja una mayor rigidez del modelo en el espacio de alta dimensionalidad del megavector, lo que afecta de forma desigual a las distintas superclases.

En la Figura 6.6, se observa que las curvas ROC se sitúan claramente por encima de la diagonal de aleatoriedad, lo que indica que el modelo es capaz de diferenciar de manera consistente entre las distintas clases frente al resto, incluso en un problema multietiqueta desbalanceado.

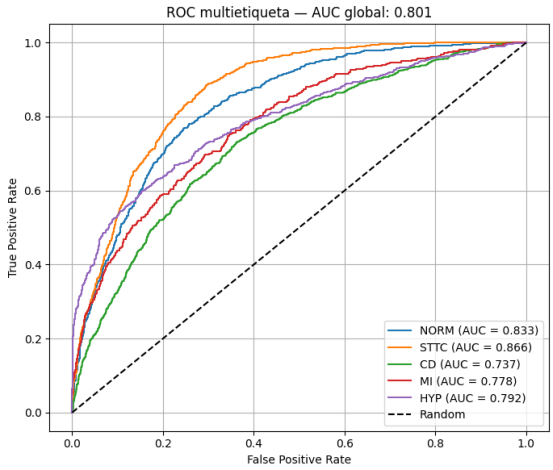

Figura 6.6: Curvas ROC de random forest de la configuración 1 (Megavector)

Este comportamiento se cuantifica mediante los valores de AUC por superclase, donde se observa un rendimiento especialmente elevado en STTC y NORM. En el caso de STTC, este valor confirma una buena capacidad del modelo para ordenar correctamente las probabilidades de pertenencia a esta clase, aunque este resultado no se refleja completamente en el recall obtenido previamente, lo que sugiere que el umbral utilizado limita parte de su sensibilidad. En NORM, el AUC elevado indica una buena capacidad de separación global entre registros normales y patológicos, coherente con la reducción de falsos positivos observada en las matrices de confusión.

Las clases MI y HYP presentan valores de AUC intermedios, lo que indica una capacidad discriminativa razonable, aunque inferior a las clases más separables. En el caso de MI, este resultado es especialmente relevante desde el punto de vista clínico, ya que confirma que, a pesar del elevado número de falsos positivos observado, el modelo es capaz de asignar puntuaciones de probabilidad adecuadas para distinguir esta clase de forma global. En HYP, el comportamiento sugiere una separabilidad moderada, afectada probablemente por la similitud de patrones con otras alteraciones estructurales del ECG.

Finalmente, la clase CD presenta el valor de AUC más bajo, lo que indica que es la superclase con mayor dificultad relativa de separación

en este escenario. Este resultado es consistente con el comportamiento observado en las matrices de confusión, donde se aprecia una tendencia a la sobreasignación de la clase, reflejando solapamiento en el espacio de representación del megavector.

En conjunto, Random Forest con entrada megavector presenta un rendimiento global sólido, aunque desigual entre clases. Destaca por una mayor sensibilidad en patologías críticas como MI y CD, mientras que en STTC el recall disminuye de forma notable.

Las matrices de confusión reflejan este comportamiento, con mejoras en especificidad frente a modelos lineales en NORM, pero mayor inestabilidad en la detección de algunas clases. En general, el modelo muestra una buena capacidad discriminativa, aunque con variaciones significativas según la superclase analizada.

Características

Los resultados de la Tabla 6.9 muestran el rendimiento del modelo Random Forest utilizando como entrada las características extraídas de la señal ECG en la configuración 1. En comparación con la representación basada en megavector, se observa un comportamiento más equilibrado en algunas clases, aunque persiste una clara tendencia a priorizar la sensibilidad en las patologías más relevantes clínicamente.

Clase	Precision	Recall	F1-score
NORM	0.58	0.56	0.57
STTC	0.85	0.46	0.59
CD	0.40	0.83	0.54
MI	0.19	0.90	0.31
HYP	0.37	0.76	0.49

Tabla 6.9: Resultados de random forest por clase (Características)

En la clase NORM se obtiene una precisión de 0.58, un recall de 0.56 y un F1-score de 0.57. Este resultado refleja un comportamiento más equilibrado respecto a la versión con megavector, con una ligera mejora

en precisión pero una reducción en la capacidad de detección de casos normales. Esto sugiere que la representación basada en características no mejora de forma clara la separación de la clase NORM, aunque sí estabiliza parcialmente el comportamiento del modelo.

La clase STTC presenta una precisión elevada junto con un recall bajo, lo que da lugar a un F1-score de 0.59. Este patrón indica que el modelo es muy conservador a la hora de asignar esta clase, de forma que cuando predice STTC lo hace con alta fiabilidad, pero deja sin detectar una proporción importante de casos positivos. Esto sugiere que, aunque las características extraídas permiten identificar patrones muy representativos, el modelo no consigue generalizar completamente la variabilidad de esta clase.

En la clase CD se observa un recall elevado frente a una precisión de 0.40, lo que indica una tendencia clara a sobreasignar esta clase. Este comportamiento refleja que el modelo prioriza la detección de la mayoría de los casos positivos, aunque a costa de incrementar el número de falsos positivos, lo que penaliza la precisión.

La clase MI presenta un recall muy alto y una precisión baja, manteniendo un F1-score reducido. Este comportamiento confirma el enfoque conservador del modelo en esta patología, donde se prioriza claramente la detección de infartos frente a la reducción de falsas alarmas.

Finalmente, la clase HYP muestra un comportamiento intermedio, con un recall de 0.76 y una precisión de 0.37. Aunque el modelo es capaz de detectar una proporción significativa de casos positivos, la baja precisión indica dificultades para diferenciar esta clase de otras patologías con patrones electrocardiográficos similares.

Las matrices de la Figura 6.7 muestran un patrón consistente con los resultados de las métricas previas, donde el modelo mantiene una tendencia clara a priorizar la sensibilidad en las clases patológicas, aunque con variaciones importantes en el equilibrio entre falsos positivos y falsos negativos según la clase.

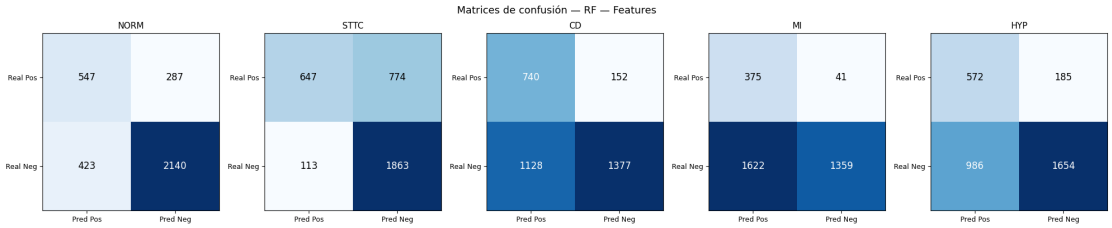

Figura 6.7: Matrices de confusión de random forest de la configuración 1 (Características)

En la clase NORM se obtienen 470 verdaderos positivos, 2224 verdaderos negativos, 339 falsos positivos y 364 falsos negativos. Se observa una reducción de falsos positivos respecto a la configuración con megavector, lo que indica una mejora en la capacidad del modelo para evitar clasificaciones erróneas de normalidad. El número de falsos negativos sigue siendo elevado, lo que refleja que una parte relevante de registros normales es clasificada como patológica. Sin embargo, este comportamiento es coherente con una estrategia conservadora en la asignación de la clase NORM.

En la clase STTC se observan 647 verdaderos positivos, 1863 verdaderos negativos, 113 falsos positivos y 774 falsos negativos. Destaca especialmente el elevado número de falsos negativos, lo que indica una pérdida significativa de sensibilidad en esta clase. A pesar de la baja tasa de falsos positivos, el modelo no es capaz de detectar una proporción importante de alteraciones del segmento ST/T, lo que sugiere una limitación en la generalización de los patrones asociados a esta clase dentro del espacio de características.

En la clase CD se obtienen 740 verdaderos positivos, 1377 verdaderos negativos, 1128 falsos positivos y 152 falsos negativos. El comportamiento muestra una tendencia clara a la sobreasignación de esta clase, reflejada en el elevado número de falsos positivos. Esto indica que el modelo identifica patrones compatibles con CD en registros que pertenecen a otras superclases, lo que reduce la precisión aunque mantiene una sensibilidad relativamente alta.

En la clase MI se observa un comportamiento especialmente rele-

vante, con 375 verdaderos positivos y únicamente 41 falsos negativos. Este resultado confirma una alta capacidad del modelo para detectar casos de infarto, priorizando claramente la sensibilidad. No obstante, este comportamiento viene acompañado de un número elevado de falsos positivos, lo que indica una estrategia conservadora orientada a minimizar errores críticos de no detección.

Finalmente, la clase HYP presenta 572 verdaderos positivos, 1654 verdaderos negativos, 986 falsos positivos y 185 falsos negativos. El modelo muestra un comportamiento intermedio, con una sensibilidad aceptable pero con una precisión limitada, lo que sugiere dificultades para diferenciar esta clase frente a otras patologías con patrones electrocardiográficos similares.

Las curvas ROC de la Figura 6.8 se sitúan claramente por encima de la diagonal de no discriminación, lo que indica que el modelo es capaz de diferenciar de manera consistente entre clases positivas y negativas en todos los casos, incluso en un escenario multiclase desbalanceado.

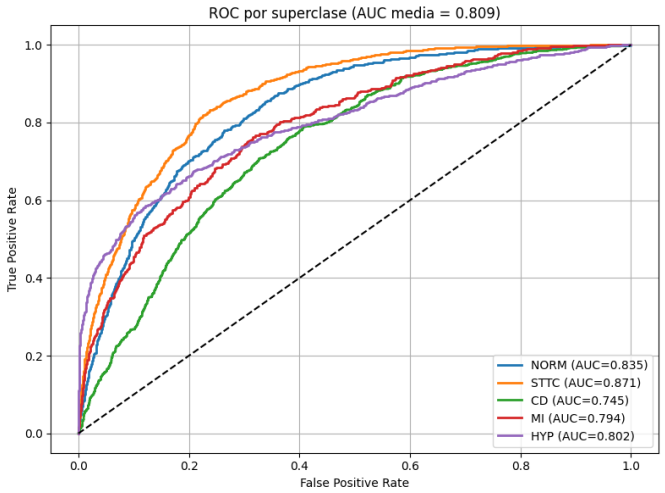

Figura 6.8: Curvas ROC de random forest de la configuración 1 (Características)

Este comportamiento se cuantifica mediante los valores de AUC por superclase, donde se observa un rendimiento especialmente elevado en STTC y NORM. En el caso de STTC, este valor confirma una buena

capacidad del modelo para ordenar correctamente las probabilidades asociadas a esta clase, aunque este rendimiento no se traduce completamente en un recall elevado, lo que sugiere que el umbral de decisión penaliza parte de la sensibilidad. En NORM, el AUC elevado refleja una buena capacidad de separación global entre registros normales y patológicos, coherente con la reducción parcial de falsos positivos observada en las matrices de confusión.

Las clases MI y HYP presentan valores de AUC intermedios, lo que indica una capacidad discriminativa razonable, aunque inferior a las clases más separables. En el caso de MI, este resultado es especialmente relevante, ya que confirma que el modelo es capaz de asignar puntuaciones de probabilidad adecuadas para diferenciar esta clase del resto, a pesar de la presencia de falsos positivos derivados del enfoque conservador del modelo. En HYP, el comportamiento sugiere una separabilidad moderada, afectada por la similitud de patrones electrocardiográficos con otras alteraciones estructurales.

Finalmente, la clase CD presenta el valor de AUC más bajo, lo que indica que sigue siendo la superclase con mayor dificultad relativa de separación en este escenario. Este resultado es coherente con el comportamiento observado en las matrices de confusión, donde se aprecia una tendencia a la sobreasignación de la clase, reflejando solapamiento en el espacio de representación incluso con características extraídas.

En conjunto, Random Forest con características muestra un comportamiento ligeramente más equilibrado que con megavector en términos de precisión y recall, aunque sin mejoras globales claras. Mantiene una fuerte orientación hacia la sensibilidad en clases patológicas como MI y CD, coherente con un enfoque clínicamente conservador.

Las matrices de confusión indican una mejora parcial de la especificidad, especialmente en NORM, aunque persiste la tendencia a priorizar la detección de patologías. En conjunto, este modelo presenta la mejor capacidad discriminativa dentro de la configuración analizada, lo que respalda el uso de características extraídas en este problema.

Red Neuronal Multicapa

En este apartado se analiza el rendimiento de la red neuronal multicapa utilizando tanto el megavector de la señal ECG como las características extraídas. Debido a su capacidad para modelar relaciones no lineales complejas, este tipo de arquitectura puede capturar patrones difíciles de representar mediante modelos más simples. El objetivo es evaluar hasta qué punto dicha capacidad se traduce en una mejora de la detección de las distintas superclases, así como estudiar el equilibrio alcanzado entre sensibilidad y precisión en cada una de ellas.

Megavector

La Tabla 6.10 muestra los resultados obtenidos por la red neuronal multicapa utilizando como entrada el megavector de la señal ECG en la configuración 1. Este modelo se evalúa en un escenario de clasificación multietiqueta desbalanceado, donde el objetivo es analizar su capacidad para discriminar entre distintas superclases a partir de la señal en bruto.

Clase	Precision	Recall	F1-score
NORM	0.76	0.15	0.26
STTC	0.71	0.85	0.77
CD	0.63	0.23	0.33
MI	0.87	0.10	0.18
HYP	0.87	0.34	0.49

Tabla 6.10: Resultados de red neuronal por clase (Megavector)

En conjunto, los resultados muestran un comportamiento desigual entre clases, con una clara especialización del modelo en la detección de STTC, mientras que el rendimiento en el resto de superclases se ve limitado principalmente por valores bajos de recall.

En la clase NORM se obtiene una precisión de 0.76 junto con un recall de 0.15, lo que indica que el modelo es capaz de identificar correctamente los casos normales cuando los predice, aunque presenta dificultades para detectarlos de forma consistente. Este comportamiento

sugiere una tendencia del modelo a ser conservador en la asignación de esta clase, probablemente debido a la similitud entre registros normales y ciertos patrones patológicos.

La clase STTC presenta el mejor rendimiento global, lo que indica una buena capacidad de detección de alteraciones del segmento ST/T. Este resultado sugiere que esta clase presenta patrones suficientemente diferenciables en la representación del megavector, lo que facilita su identificación por parte del modelo.

En el caso de CD y MI se observan precisiones elevadas, pero recalls bajos. Esto indica que el modelo solo asigna estas etiquetas cuando existe una alta confianza, lo que reduce la aparición de falsos positivos, pero limita de forma importante la sensibilidad del sistema en estas patologías. Este comportamiento es especialmente relevante en MI, donde la baja sensibilidad implica una menor capacidad de detección de eventos críticos.

Finalmente, la clase HYP presenta un comportamiento intermedio, lo que indica una capacidad de detección moderada, aunque todavía limitada en comparación con otras clases más fácilmente identificables.

En la Figura 6.9, que muestra las matrices de confusión de la red neuronal con entrada megavector, se observa un patrón claramente desigual, donde el modelo muestra una buena capacidad de discriminación en STTC, mientras que presenta dificultades importantes en la detección del resto de patologías.

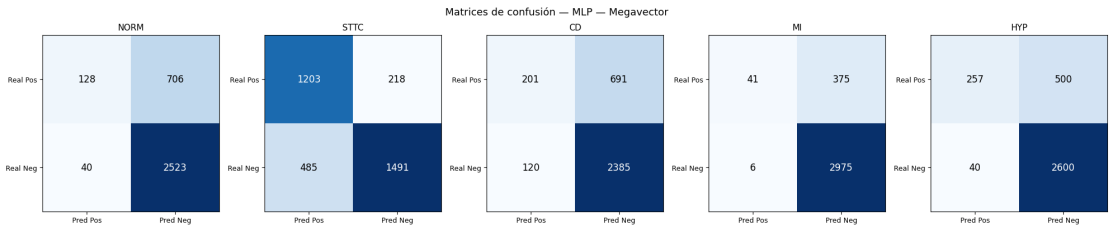

Figura 6.9: Matrices de confusión de red neuronal de la configuración 1 (Megavector)

En la clase NORM se obtienen 128 verdaderos positivos, 2523 ver-

daderos negativos, 40 falsos positivos y 706 falsos negativos. Este resultado refleja una capacidad muy limitada del modelo para identificar correctamente los registros normales, lo que se traduce en un recall bajo. El elevado número de falsos negativos indica que una gran parte de los casos normales está siendo clasificada como patológica, lo que sugiere que el modelo es especialmente restrictivo a la hora de asignar esta clase.

En la clase STTC se observan 1203 verdaderos positivos, 1491 verdaderos negativos, 485 falsos positivos y 218 falsos negativos. En este caso, el modelo muestra su mejor comportamiento, con una capacidad de detección relativamente sólida y un equilibrio razonable entre sensibilidad y precisión. Aunque existen falsos positivos relevantes, la tasa de falsos negativos es moderada, lo que explica el buen recall observado previamente en las métricas por clase.

En la clase CD se obtienen 201 verdaderos positivos, 2385 verdaderos negativos, 120 falsos positivos y 691 falsos negativos. Este resultado evidencia una baja capacidad de detección de la clase, con un número elevado de falsos negativos. Esto indica que el modelo no consigue capturar adecuadamente los patrones asociados a CD dentro del espacio de representación del megavector, lo que explica su bajo recall.

En la clase MI se observa un comportamiento especialmente crítico desde el punto de vista clínico, con 41 verdaderos positivos y 375 falsos negativos. Este resultado refleja una sensibilidad extremadamente baja, lo que implica que el modelo no es capaz de detectar la mayoría de los casos de infarto. Aunque el número de falsos positivos es reducido, este comportamiento es problemático en un contexto clínico, ya que prioriza en exceso la precisión frente a la detección de casos positivos.

Finalmente, la clase HYP presenta 257 verdaderos positivos, 2600 verdaderos negativos, 40 falsos positivos y 500 falsos negativos. El modelo muestra una baja sensibilidad en esta clase, con un número elevado de falsos negativos, lo que indica dificultades claras para distinguirla de otras patologías con patrones electrocardiográficos similares.

En conjunto, las curvas ROC (Figura 6.10) se sitúan por encima de la diagonal de aleatoriedad en todas las superclases, lo que indica que el modelo es capaz de establecer una ordenación coherente de las probabilidades asignadas a cada clase.

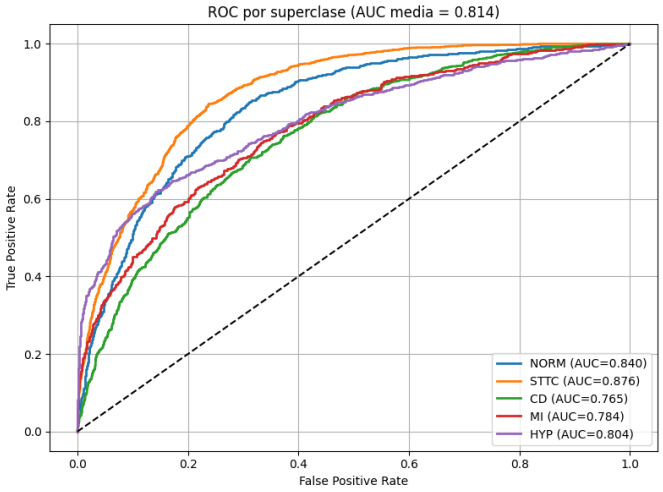

Figura 6.10: Curvas ROC de red neuronal de la configuración 1 (Megavector)

Este comportamiento se cuantifica mediante los valores de AUC por superclase, donde se observa un rendimiento especialmente elevado en STTC y NORM. En el caso de STTC, este valor confirma una buena capacidad del modelo para distinguir esta clase del resto, coherente con su buen rendimiento en términos de recall y F1-score. En NORM, el AUC elevado indica que, a pesar del bajo recall observado en las matrices de confusión, el modelo sí es capaz de asignar puntuaciones de probabilidad adecuadas, lo que sugiere que el problema no reside tanto en la separabilidad, sino en el punto de decisión utilizado.

Las clases MI y HYP presentan valores de AUC intermedios, lo que indica una capacidad discriminativa razonable. En el caso de MI, este resultado es especialmente relevante desde el punto de vista clínico, ya que demuestra que el modelo sí es capaz de diferenciar parcialmente esta clase del resto, aunque la elección del umbral y la distribución de errores penalicen fuertemente su sensibilidad. En HYP, el comportamiento refleja una separabilidad moderada, consistente con la dificul-

tad del modelo para capturar completamente sus patrones en la señal ECG.

Finalmente, la clase CD presenta el valor de AUC más bajo, lo que indica que es la superclase con mayor dificultad relativa de separación en este escenario. Este resultado es coherente con el comportamiento observado en las matrices de confusión, donde se aprecia una baja capacidad de detección y un elevado número de falsos negativos.

En conjunto, la red neuronal multicapa con megavector depende fuertemente de la separabilidad de las clases, destacando en STTC pero con baja sensibilidad en el resto de superclases.

Las matrices de confusión confirman esta especialización, evidenciando que el megavector no es suficientemente discriminativo para varias patologías.

En conjunto, el modelo muestra buena capacidad global, aunque con un rendimiento claramente desequilibrado entre clases, condicionado tanto por la representación como por la toma de decisiones.

Características

La Tabla 6.11 muestra los resultados obtenidos por la red neuronal multicapa utilizando como entrada las características extraídas de la señal ECG en la configuración 1. En este caso, se observa un comportamiento similar al obtenido con la representación basada en megavector, aunque con ciertas variaciones en el equilibrio entre precisión y recall según la superclase.

Clase	Precision	Recall	F1-score
NORM	0.80	0.05	0.09
STTC	0.72	0.86	0.78
CD	0.71	0.16	0.26
MI	0.88	0.09	0.17
HYP	0.92	0.35	0.51

Tabla 6.11: Resultados de red neuronal por clase (Características)

En conjunto, los resultados muestran un patrón altamente conser-

vador en la mayoría de las clases, con valores de precisión elevados pero recalls significativamente bajos, lo que indica que el modelo solo asigna ciertas etiquetas cuando existe una alta confianza en la predicción.

En la clase NORM se obtiene una precisión de 0.80 junto con un recall de 0.05, lo que refleja una capacidad muy limitada para identificar correctamente registros normales. Aunque las predicciones positivas de esta clase son altamente fiables, el modelo apenas las asigna, lo que implica una sensibilidad extremadamente baja.

La clase STTC presenta el mejor rendimiento global, lo que indica una buena capacidad de detección de alteraciones del segmento ST/T. Este resultado sugiere que esta clase presenta patrones suficientemente diferenciables dentro del espacio de características extraídas, lo que facilita su identificación por parte del modelo.

En el caso de CD y MI se observa un comportamiento especialmente restrictivo, con recalls muy bajos, a pesar de presentar precisiones elevadas. Esto indica que el modelo solo asigna estas etiquetas en casos muy concretos, reduciendo los falsos positivos pero a costa de una pérdida importante de sensibilidad. Este comportamiento es especialmente crítico en MI, donde la baja capacidad de detección implica una limitación importante desde el punto de vista clínico.

Finalmente, la clase HYP presenta un comportamiento intermedio, lo que refleja una mejora respecto a las clases más problemáticas, aunque todavía con una sensibilidad insuficiente para una detección completa.

En la Figura 6.11 se observa un patrón claramente conservador, donde el modelo tiende a reducir los falsos positivos, pero a costa de incrementar de forma significativa los falsos negativos en varias clases.

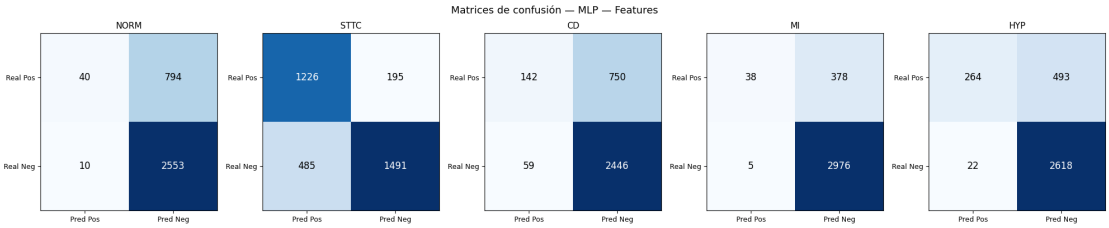

Figura 6.11: Matrices de confusión de red neuronal de la configuración 1 (Características)

En la clase NORM se obtienen 40 verdaderos positivos, 2553 verdaderos negativos, 10 falsos positivos y 794 falsos negativos. Este resultado refleja una capacidad muy limitada del modelo para identificar correctamente los registros normales, con una tendencia marcada a clasificar la mayoría de los casos como patológicos. Aunque el número de falsos positivos es prácticamente nulo, el elevado número de falsos negativos indica una sensibilidad extremadamente baja en esta clase.

En la clase STTC se observan 1226 verdaderos positivos, 1491 verdaderos negativos, 485 falsos positivos y 195 falsos negativos. En este caso, el modelo presenta su mejor comportamiento, con una capacidad de detección elevada y un equilibrio razonable entre sensibilidad y precisión. Este resultado es coherente con las métricas por clase, donde STTC destaca como la superclase mejor identificada por el modelo.

En la clase CD se obtienen 142 verdaderos positivos, 2446 verdaderos negativos, 59 falsos positivos y 750 falsos negativos. Este comportamiento evidencia una sensibilidad muy baja, ya que la mayoría de los casos positivos no son detectados. A pesar de la baja tasa de falsos positivos, el modelo no consigue capturar adecuadamente los patrones asociados a esta patología.

En la clase MI se observa un comportamiento especialmente crítico desde el punto de vista clínico, con 38 verdaderos positivos y 378 falsos negativos. Este resultado indica una capacidad prácticamente nula de detección de infartos en comparación con el número total de casos reales, lo que refleja una limitación importante del modelo en esta clase, a pesar de la baja tasa de falsos positivos.

Finalmente, la clase HYP presenta 264 verdaderos positivos, 2618 verdaderos negativos, 22 falsos positivos y 493 falsos negativos. Aunque el modelo reduce significativamente los falsos positivos en esta clase, la sensibilidad sigue siendo limitada, lo que indica dificultades para capturar completamente sus patrones característicos.

Las curvas ROC de la Figura 6.12 permiten evaluar la capacidad discriminativa de la red neuronal multicapa con entrada basada en características de forma independiente al umbral de decisión utilizado en la clasificación final. Este análisis resulta especialmente relevante en este caso, ya que las métricas de clasificación y las matrices de confusión mostraron una sensibilidad reducida en varias superclases.

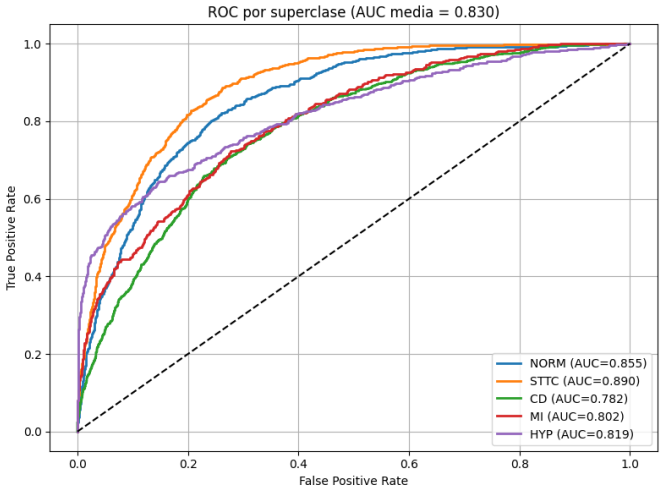

Figura 6.12: Curvas ROC de red neuronal de la configuración 1 (Características)

Los valores de AUC obtenidos indican una buena capacidad de separación entre clases en todas las categorías analizadas. La superclase STTC presenta el mejor resultado, lo que confirma la elevada capacidad del modelo para distinguir este tipo de alteraciones del resto de registros. Este comportamiento es coherente con los buenos valores de recall y F1-score observados anteriormente.

La clase NORM alcanza un AUC de 0.8547, uno de los valores más elevados del conjunto. Este resultado indica que el modelo es capaz

de diferenciar adecuadamente entre registros normales y patológicos a nivel probabilístico. Sin embargo, esta capacidad discriminativa no se traduce en una sensibilidad elevada, como se observó en las matrices de confusión, lo que sugiere que las limitaciones del modelo están más relacionadas con el proceso de decisión final que con la separación de clases en sí misma.

Las clases MI y HYP presentan valores de AUC de 0.8021 y 0.8194 respectivamente, lo que indica una capacidad discriminativa razonable. En el caso de MI, este resultado es especialmente relevante, ya que demuestra que el modelo es capaz de identificar patrones asociados al infarto a nivel probabilístico, aunque posteriormente la clasificación final resulte excesivamente conservadora y produzca un número elevado de falsos negativos.

Por su parte, la clase CD obtiene el valor de AUC más reducido, lo que confirma que se trata de una de las superclases más difíciles de separar en este escenario. Este comportamiento es consistente con las dificultades observadas previamente en las métricas por clase y en las matrices de confusión.

En conjunto, la red neuronal multicapa con características extraídas prioriza la precisión frente a la sensibilidad, generando predicciones fiables pero con menor capacidad para detectar todos los casos positivos. Las matrices de confusión confirman este patrón, especialmente en clases como MI y CD.

En conjunto, el modelo muestra buena capacidad discriminativa, aunque con un recall limitado que condiciona su rendimiento final.

De forma global, los resultados obtenidos muestran que todos los modelos son capaces de discriminar adecuadamente entre las distintas superclases, alcanzando valores de AUC superiores a 0.8 en la mayoría de los casos. De esta forma, se observa que el uso de características extraídas produce una mejora consistente respecto al uso directo del

megavector, especialmente en términos de capacidad discriminativa.

La regresión logística presenta el rendimiento más limitado, aunque mantiene un comportamiento estable y mejora notablemente cuando se utilizan características extraídas. Por su parte, la red neuronal multicapa obtiene los mayores valores de Hamming Accuracy y AUC global, pero muestra una tendencia a favorecer las clases mayoritarias, reduciendo significativamente el recall en varias patologías relevantes.

En cambio, Random Forest ofrece el mejor equilibrio entre precisión y recall en la mayoría de las superclases, manteniendo además una elevada sensibilidad en patologías críticas como MI. Desde una perspectiva clínica, este comportamiento resulta especialmente interesante, ya que reduce el riesgo de falsos negativos sin penalizar excesivamente el rendimiento global.

Por tanto, aunque la red neuronal multicapa obtiene las mejores métricas globales, Random Forest con características extraídas puede considerarse el modelo más adecuado dentro de esta configuración, al combinar una buena capacidad discriminativa con un comportamiento más equilibrado entre superclases y una mayor sensibilidad en las patologías de mayor interés clínico.

6.2.2. Configuración 2

La segunda configuración se basa exclusivamente en el uso de características extraídas a partir de la señal electrocardiográfica, sustituyendo la representación en bruto empleada en la configuración anterior. Este cambio permite trabajar con una representación más compacta y estructurada de la señal, en la que se sintetiza información relevante del ECG en variables diseñadas para capturar patrones clínicamente significativos.

Al reducir la dimensionalidad del problema y centrarse en descriptores más informativos, esta configuración busca mejorar la capacidad de generalización de los modelos y facilitar la separación entre superclases. En este contexto, se analiza si la transformación de la señal en un

conjunto de características derivadas permite mejorar el rendimiento de clasificación frente a representaciones más directas.

El análisis se realiza de forma análoga a la configuración 1, utilizando métricas por clase (precisión, recall y F1-score), junto con las matrices de confusión y las curvas ROC asociadas a cada modelo. Esto permite evaluar de manera comparativa el impacto del tipo de representación sobre la capacidad discriminativa de los algoritmos.

Dado que el problema mantiene un importante desbalance entre clases y diferentes implicaciones clínicas según la patología, se conservan los umbrales de decisión específicos por superclase definidos en el apartado 5.5.2, con el objetivo de priorizar la detección de aquellas enfermedades donde los falsos negativos tienen mayor relevancia clínica.

A continuación, se presenta el análisis detallado del comportamiento de los distintos modelos bajo esta representación de entrada.

Regresión Logística

La regresión logística constituye el modelo lineal de referencia dentro de esta configuración. Su utilización permite evaluar hasta qué punto las distintas superclases pueden ser separadas mediante combinaciones lineales de la información procedente de las doce derivaciones del ECG. Este enfoque sirve como punto de partida para analizar el valor añadido de incorporar una representación multiderivación completa frente a escenarios más limitados.

Los resultados obtenidos por la regresión logística en la Configuración 2, basada en la utilización de las doce derivaciones del ECG, se recogen en la Tabla 6.12. En esta configuración se observa una mejora generalizada del rendimiento respecto a la Configuración 1, especialmente en términos de recall, lo que indica una mayor capacidad del modelo para detectar correctamente las distintas superclases.

Clase	Precision	Recall	F1-score
NORM	0.63	0.80	0.71
STTC	0.76	0.92	0.83
CD	0.49	0.86	0.63
MI	0.30	0.91	0.45
HYP	0.52	0.84	0.64

Tabla 6.12: Resultados de regresión logística por clase (Configuración 2)

En conjunto, se observa un incremento significativo en la sensibilidad del modelo en todas las clases, lo que indica que la incorporación de la información procedente de las doce derivaciones mejora la capacidad de detección de patrones patológicos. Este comportamiento es especialmente relevante en un contexto clínico, donde la reducción de falsos negativos tiene un impacto directo en la seguridad del diagnóstico.

La clase STTC presenta el mejor rendimiento global, lo que confirma que se trata de la superclase mejor representada por el modelo en esta configuración. Este resultado sugiere que los patrones asociados a alteraciones del segmento ST/T son claramente identificables cuando se dispone de información multiderivación completa.

En el caso de NORM, se observa una mejora notable en comparación con la configuración anterior. Esto indica una mayor capacidad del modelo para identificar correctamente registros normales, aunque la precisión se mantiene moderada, lo que sugiere cierta confusión residual con clases patológicas.

La clase CD muestra una mejora significativa en recall, lo que indica una mayor sensibilidad en la detección de alteraciones de conducción. Sin embargo, la precisión moderada refleja la persistencia de cierto solapamiento con otras clases, lo que afecta al equilibrio global de la métrica F1.

Por otro lado, MI presenta una mejora especialmente relevante desde el punto de vista clínico. Este resultado indica que el modelo es capaz de detectar la gran mayoría de los casos de infarto, reduciendo de forma importante el riesgo de falsos negativos, aunque a costa de una

precisión más reducida.

Finalmente, la clase HYP alcanza un recall de 0.84, mostrando también una mejora respecto a la configuración anterior. Sin embargo, su precisión moderada indica que aún existe cierta dificultad para separar completamente esta clase del resto de patologías.

Las matrices de confusión de la Figura 6.13 permiten analizar en detalle el comportamiento de la regresión logística en la Configuración 2, basada en las doce derivaciones del ECG. En conjunto, se observa una mejora clara en la capacidad de detección de las distintas superclases, especialmente en términos de reducción de falsos negativos, lo que refuerza el efecto positivo de la información multiderivación en el rendimiento del modelo.

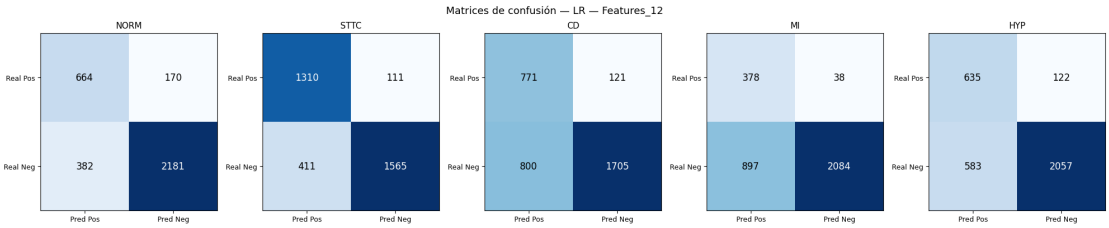

Figura 6.13: Matrices de confusión de regresión logística de la configuración 2

En la clase NORM, el modelo presenta 664 verdaderos positivos, 2181 verdaderos negativos, 382 falsos positivos y 170 falsos negativos. Estos resultados muestran una mejora significativa respecto a la configuración anterior, especialmente en la reducción de falsos negativos, lo que indica una mayor capacidad para identificar correctamente registros normales. No obstante, el número de falsos positivos se mantiene relativamente elevado, lo que sugiere cierta confusión con patrones patológicos leves.

En la clase STTC se obtienen 1310 verdaderos positivos, 1565 verdaderos negativos, 411 falsos positivos y 111 falsos negativos. Este resultado refleja un rendimiento muy sólido, con una alta capacidad de detección y una reducción notable de los falsos negativos. Es la clase

que presenta el comportamiento más equilibrado, consolidándose como una de las superclases mejor identificadas por el modelo.

En la clase CD se observan 771 verdaderos positivos, 1705 verdaderos negativos, 800 falsos positivos y 121 falsos negativos. Aunque la sensibilidad mejora de forma evidente respecto a la configuración anterior, el incremento de falsos positivos indica que la mayor capacidad de detección viene acompañada de una mayor tendencia a la sobreclasificación de esta patología.

En la clase MI destaca especialmente la reducción de falsos negativos, con solo 38 casos, lo que supone una mejora muy relevante desde el punto de vista clínico. El modelo es capaz de detectar prácticamente todos los casos de infarto, aunque esto se acompaña de un aumento de falsos positivos, lo que refleja un comportamiento claramente orientado a maximizar la sensibilidad.

Finalmente, la clase HYP presenta 635 verdaderos positivos, 2057 verdaderos negativos, 583 falsos positivos y 122 falsos negativos. Se observa también una mejora importante en la detección de esta patología, con una reducción significativa de falsos negativos, aunque persiste cierta confusión con otras clases que incrementa el número de falsos positivos.

Las curvas ROC de la Figura 6.14 permiten evaluar la capacidad discriminativa de la regresión logística en la Configuración 2 de forma independiente al umbral de decisión utilizado en la clasificación final. Este análisis resulta especialmente relevante, ya que permite comprobar si las mejoras observadas en las métricas de clasificación se deben a una mejor separación de las clases o únicamente a un ajuste del umbral.

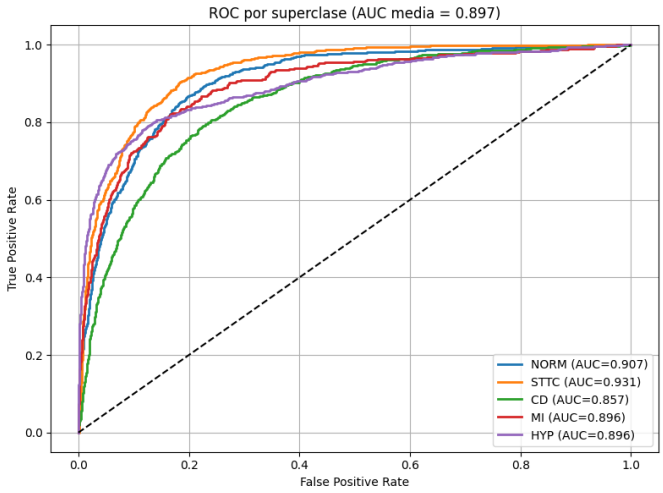

Figura 6.14: Curvas ROC de regresión logística de la configuración 2

En conjunto, las curvas ROC muestran un comportamiento claramente superior al de la configuración anterior, situándose todas las clases en valores próximos a la esquina superior izquierda del plano ROC, lo que indica una alta capacidad de discriminación entre superclases.

La clase STTC obtiene el mejor resultado global, lo que confirma que el modelo es capaz de distinguir de forma muy robusta este tipo de alteraciones respecto al resto de clases. Este resultado es coherente con el elevado recall y F1-score observados previamente, consolidando a STTC como la superclase mejor modelada en esta configuración.

La clase NORM alcanza un AUC de 0.9072, lo que representa una mejora muy significativa respecto a la configuración anterior. Este resultado indica que el modelo es capaz de separar adecuadamente los registros normales de los patológicos a nivel de probabilidad, incluso aunque la clasificación final dependa del umbral aplicado.

Las clases MI y HYP presentan valores de AUC de 0.8962 en ambos casos, lo que refleja una capacidad discriminativa muy elevada. En el caso de MI, este resultado es especialmente relevante desde el punto de vista clínico, ya que confirma que el modelo aprende a identificar patrones asociados al infarto de forma consistente, aunque la decisión

final pueda verse afectada por la elección del umbral.

Finalmente, la clase CD obtiene un AUC de 0.8569, el menor de entre las superclases, aunque sigue reflejando una capacidad discriminativa alta dentro del modelo. Esto indica que, aunque sigue siendo la superclase más compleja, la incorporación de las doce derivaciones mejora de forma notable su separabilidad respecto a la configuración anterior.

En conjunto, el uso de las doce derivaciones mejora de forma consistente la capacidad de detección del modelo lineal, especialmente la sensibilidad en patologías críticas como MI, lo que resulta clave en un contexto clínico donde los falsos negativos son especialmente relevantes.

Las matrices de confusión confirman esta mejora, aunque acompañada de un ligero aumento de falsos positivos, lo que sugiere que la representación multiderivación captura mejor los patrones ECG a costa de cierta confusión entre clases.

En conjunto, se observa una mejora clara en la capacidad discriminativa del modelo en esta configuración, coherente con las métricas obtenidas por clase.

Random Forest

Random Forest introduce un enfoque basado en conjuntos de árboles de decisión, lo que le permite modelar relaciones no lineales entre las variables procedentes de las doce derivaciones del ECG. En este contexto, su análisis permite evaluar si la información espacial completa de la señal mejora la capacidad de discriminación entre superclases en comparación con modelos lineales.

Los resultados obtenidos por la regresión logística en la Configuración 2, empleando características extraídas junto con la información de las doce derivaciones del ECG, se recogen en la Tabla 6.13. En términos generales, se observa una mejora en el equilibrio global del modelo

respecto a la configuración anterior, especialmente en aquellas clases donde se había identificado mayor dificultad de separación.

Clase	Precision	Recall	F1-score
NORM	0.72	0.59	0.65
STTC	0.90	0.67	0.76
CD	0.47	0.87	0.61
MI	0.28	0.95	0.43
HYP	0.48	0.81	0.60

Tabla 6.13: Resultados de regresión logística por clase (Configuración 2)

En conjunto, se observa un cambio en el comportamiento del modelo respecto a la Configuración 1, con una tendencia más equilibrada entre precisión y recall en varias superclases. En particular, se aprecia una mejora en la precisión de clases como STTC y NORM, lo que indica una reducción en el número de falsos positivos respecto a la versión anterior del modelo.

La clase STTC presenta el mejor equilibrio global, lo que refleja una capacidad de clasificación muy sólida. Sin embargo, el recall de 0.67 indica que el modelo es más conservador en la asignación de esta etiqueta, priorizando la fiabilidad de las predicciones frente a la sensibilidad.

En el caso de NORM, se observa una mejora en precisión, aunque con una reducción del recall respecto a la configuración anterior. Este comportamiento indica que el modelo es más restrictivo a la hora de clasificar registros como normales, reduciendo falsos positivos a costa de un mayor número de falsos negativos.

La clase CD presenta un comportamiento caracterizado por una alta sensibilidad pero una precisión moderada, lo que sugiere que el modelo tiende a sobrepredecir esta clase en algunos casos, aunque consigue detectar la mayoría de los positivos reales.

Por otro lado, MI destaca por un recall muy elevado, lo que refleja una estrategia claramente orientada a maximizar la detección de infartos, reduciendo al mínimo los falsos negativos. Sin embargo, la baja precisión indica un elevado número de falsos positivos, lo que represen-

ta un compromiso explícito hacia la sensibilidad en esta clase crítica.

Finalmente, HYP muestra un comportamiento intermedio, lo que refleja una mejora en la capacidad de detección respecto a configuraciones anteriores, aunque todavía con cierta inestabilidad en la asignación de esta etiqueta.

Las matrices de confusión de la Figura 6.15 permiten analizar con mayor detalle el comportamiento de la regresión logística en la Configuración 2, basada en características extraídas junto con las doce derivaciones del ECG. En conjunto, se observa una mejora en la capacidad de detección de las distintas superclases, aunque acompañada de un incremento en el número de falsos positivos en determinadas categorías.

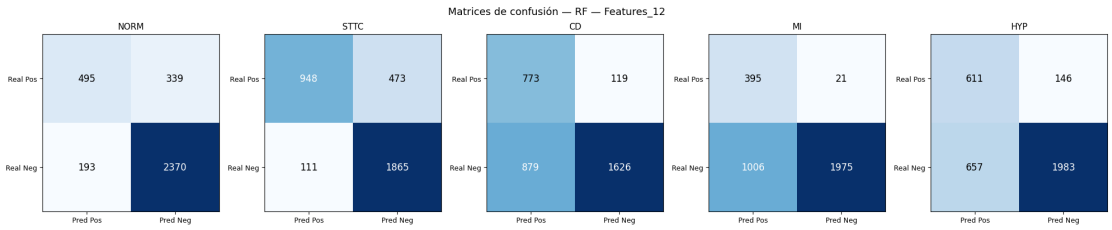

Figura 6.15: Matrices de confusión de random forest de la configuración 2

En la clase NORM se obtienen 495 verdaderos positivos, 2370 verdaderos negativos, 193 falsos positivos y 339 falsos negativos. En este caso, el modelo muestra una reducción significativa de falsos positivos respecto a configuraciones anteriores, lo que indica una mayor fiabilidad a la hora de identificar registros normales. Sin embargo, el número de falsos negativos se mantiene en un nivel moderado, lo que refleja cierta tendencia a no clasificar como normales algunos casos que sí lo son.

En la clase STTC se observan 948 verdaderos positivos, 1865 verdaderos negativos, 111 falsos positivos y 473 falsos negativos. Este resultado muestra un comportamiento más conservador del modelo, con una reducción clara de falsos positivos, aunque a costa de un incre-

mento en los falsos negativos, lo que sugiere una menor sensibilidad en comparación con la configuración anterior.

En la clase CD se obtienen 773 verdaderos positivos, 1626 verdaderos negativos, 879 falsos positivos y 119 falsos negativos. En este caso, se mantiene una alta capacidad de detección de la clase, aunque el aumento de falsos positivos indica una tendencia del modelo a sobreasignar esta categoría en presencia de patrones similares a otras alteraciones.

En la clase MI se observa un comportamiento especialmente relevante desde el punto de vista clínico, con 395 verdaderos positivos y únicamente 21 falsos negativos. Esto indica una capacidad muy elevada para detectar casos de infarto, minimizando el riesgo de no identificación de esta patología. No obstante, este comportamiento se acompaña de un elevado número de falsos positivos, lo que refleja una estrategia claramente orientada a maximizar la sensibilidad.

Finalmente, la clase HYP presenta 611 verdaderos positivos, 1983 verdaderos negativos, 657 falsos positivos y 146 falsos negativos. El modelo muestra una capacidad de detección razonable, aunque con cierta confusión con otras clases, lo que se traduce en un equilibrio intermedio entre precisión y sensibilidad.

Las curvas ROC de la Figura 6.16 permiten analizar la capacidad discriminativa del Random Forest, evaluando su rendimiento independientemente al umbral de decisión utilizado. Este análisis resulta especialmente relevante en modelos basados en árboles, donde la capacidad de modelar relaciones no lineales puede influir significativamente en la separación entre clases.

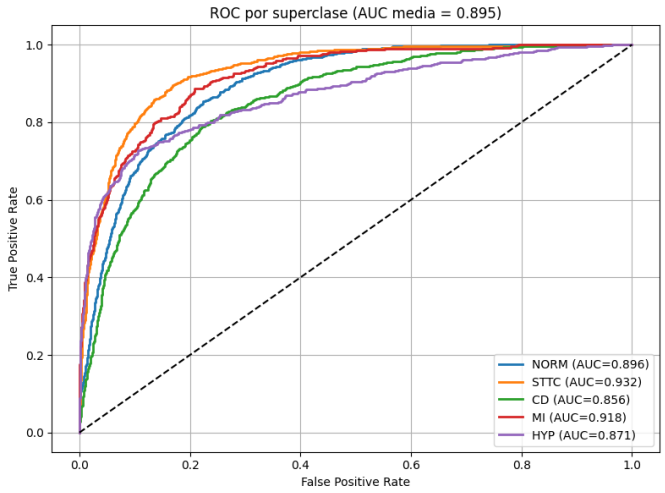

Figura 6.16: Curvas ROC de random forest de la configuración 2

En términos generales, las curvas ROC reflejan un comportamiento estable y consistente del modelo, con valores de AUC elevados en todas las superclases. Esto indica que Random Forest es capaz de establecer una ordenación fiable de las probabilidades asociadas a cada clase, incluso en presencia de solapamiento entre patrones electrocardiográficos.

El mejor rendimiento lo presenta la clase STTC, demostrando así que las combinaciones de caracterísiticas de esta patología se capturan correctamente por el conjunto de árboles del modelo. Esto refuerza su papel como una de las superclases mejor diferenciadas.

La clase MI tiene una AUC superior a 0.9, que confirma la capacidad del modelo para identificar patrones asociados al infarto. Esto es relevante ya que indica que la estructura basada en árboles es capaz de mantener una buena separación incluso en una clase clínicamente crítica y con patrones complejos.

La AUC de la clase NORM refleja una buena capacidad del modelo de diferenciar registros normales de aquellos con patologías. Aunque no es la clase con mejor rendimiento, mantiene una separación sólida en el espacio de decisión del modelo.

Por su parte, la superclase HYP es ligeramente más complicada de

diagnosticar, lo que hace que surja una mayor dificultad en la representación de alteraciones dentro del modelo.

Finalmente, la clase CD tiene la AUC con el valor más bajo, aunque sigue siendo elevada. Este resultado confirma que, a pesar de ser la clase más compleja en términos de separación, Random Forest es capaz de capturar parte de su estructura discriminativa.

En conjunto, la combinación de características extraídas y las doce derivaciones mejora la capacidad de detección del modelo, especialmente la sensibilidad, aunque introduce cierto desequilibrio en la precisión en algunas superclases.

Las matrices de confusión confirman una reducción de falsos negativos en clases críticas como MI, a costa de un aumento de falsos positivos, en línea con un enfoque clínico conservador.

En conjunto, Random Forest muestra un rendimiento estable y una buena capacidad para modelar relaciones no lineales entre clases.

Red Neuronal Multicapa

La red neuronal multicapa representa el modelo con mayor capacidad de aprendizaje dentro de esta configuración. Su arquitectura permite capturar relaciones no lineales complejas entre las distintas derivaciones del ECG, lo que la convierte en una herramienta especialmente adecuada para evaluar el potencial discriminativo de la señal multiderivación completa en la detección de patrones clínicamente relevantes.

Los resultados obtenidos por la red neuronal multicapa en la Configuración 2 se recogen en la Tabla 6.14. En esta configuración se observa un comportamiento global equilibrado del modelo en la clasificación de las distintas superclases, con un buen compromiso entre precisión y recall en la mayoría de ellas.

Clase	Precision	Recall	F1-score
NORM	0.64	0.82	0.72
STTC	0.78	0.92	0.84
CD	0.65	0.78	0.71
MI	0.49	0.71	0.58
HYP	0.62	0.80	0.70

Tabla 6.14: Resultados de red neuronal por clase (Configuración 2)

La clase STTC destaca como la superclase con mejor rendimiento global, mostrando un comportamiento especialmente estable entre precisión y recall. Esto indica que el modelo es capaz de capturar de forma clara los patrones asociados a este tipo de alteraciones electrocardiográficas, siendo la clase mejor diferenciada en este escenario.

En el caso de NORM, se observa un rendimiento equilibrado, con una buena capacidad de detección de registros normales y una precisión consistente. Esto sugiere que el modelo logra diferenciar de manera adecuada entre casos normales y patológicos, aunque con cierto grado de solapamiento inherente al problema.

La clase CD presenta un comportamiento sólido, con una capacidad de detección razonable y una precisión moderada. Esto indica que el modelo es capaz de identificar la mayoría de los casos asociados a esta alteración, aunque aún existe cierta confusión con otras superclases con patrones similares.

Por otro lado, MI se mantiene como una de las clases más complejas del problema, con una sensibilidad menor en comparación con el resto, lo que refleja la dificultad intrínseca de identificar este tipo de patrones electrocardiográficos. Aun así, el modelo consigue mantener una capacidad de detección aceptable.

Finalmente, HYP muestra un comportamiento intermedio, con un rendimiento relativamente equilibrado, aunque ligeramente inferior a las clases más separables. Esto sugiere que el modelo es capaz de capturar parte de sus características, pero aún existen solapamientos con otras patologías.

Las matrices de confusión de la Figura 6.17 permiten analizar en detalle el comportamiento de la red neuronal multicapa en la Configuración 2, descomponiendo el rendimiento del modelo en términos de aciertos y errores para cada superclase.

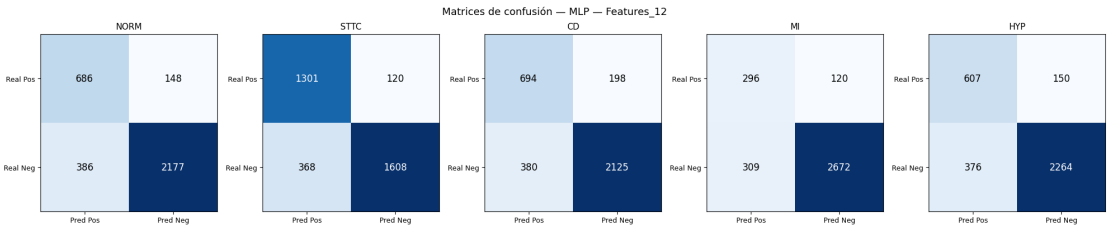

Figura 6.17: Matrices de confusión de red neuronal de la configuración 2

En la clase NORM, el modelo presenta 686 verdaderos positivos, 2177 verdaderos negativos, 386 falsos positivos y 148 falsos negativos. En conjunto, se observa un comportamiento estable, con una buena capacidad para identificar registros normales y un número relativamente controlado de falsos positivos. Esto indica que el modelo consigue diferenciar de forma adecuada entre patrones normales y patológicos, aunque todavía existe cierto solapamiento con algunas alteraciones leves.

En la clase STTC, el modelo muestra 1301 verdaderos positivos, 1608 verdaderos negativos, 368 falsos positivos y 120 falsos negativos. Este resultado refleja uno de los mejores comportamientos globales, con una alta capacidad de detección y una baja tasa de error por omisión, lo que confirma la buena separabilidad de este tipo de alteraciones.

En la clase CD se observan 694 verdaderos positivos, 2125 verdaderos negativos, 380 falsos positivos y 198 falsos negativos. El modelo presenta una capacidad de detección adecuada, aunque con cierta presencia de falsos positivos, lo que sugiere solapamiento con otras superclases con patrones electrocardiográficos similares.

En la clase MI, el modelo obtiene 296 verdaderos positivos, 2672 verdaderos negativos, 309 falsos positivos y 120 falsos negativos. Este resultado es especialmente relevante desde el punto de vista clínico, ya

que muestra una buena capacidad de detección de infartos, manteniendo una tasa de falsos negativos relativamente baja, aunque con cierto incremento de falsos positivos debido a la dificultad inherente de esta clase.

En la clase HYP se registran 607 verdaderos positivos, 2264 verdaderos negativos, 376 falsos positivos y 150 falsos negativos. El comportamiento es intermedio, con una capacidad de detección razonable y un equilibrio aceptable entre errores, aunque todavía existe cierto solapamiento con otras patologías.

Las curvas ROC de la Figura 6.18 permiten evaluar la capacidad discriminativa de la red neuronal multicapa en la Configuración 2, analizando su comportamiento de forma independiente al umbral de decisión utilizado en la clasificación final. Este análisis resulta especialmente relevante en este tipo de modelos, ya que permite valorar su capacidad para asignar probabilidades coherentes a cada superclase incluso en presencia de solapamiento entre patrones.

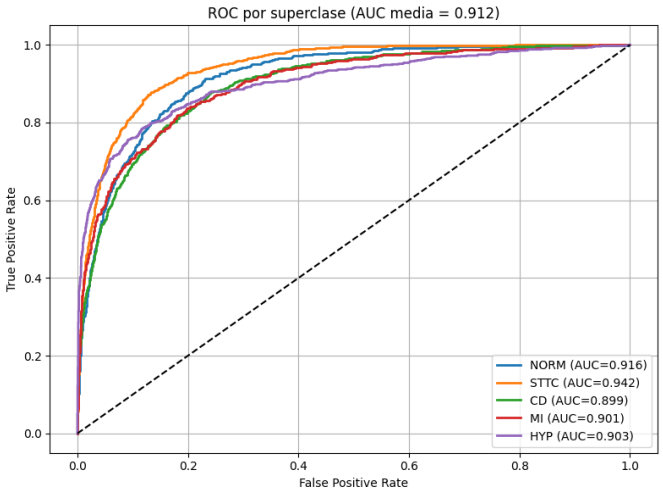

Figura 6.18: Curvas ROC de red neuronal de la configuración 2

En conjunto, las curvas ROC muestran un comportamiento muy sólido del modelo, con valores de AUC elevados en todas las superclases, lo que indica una alta capacidad de separación entre clases a nivel de

puntuaciones de probabilidad. Esto refleja que la red neuronal no solo clasifica correctamente en el punto de decisión final, sino que además mantiene una ordenación consistente de las predicciones.

La clase STTC obtiene el mejor rendimiento global, lo que confirma una excelente capacidad del modelo para diferenciar este tipo de alteraciones del resto de clases. Este resultado es coherente con el comportamiento observado en las métricas por clase y en las matrices de confusión, consolidándose como la superclase más claramente separable.

La clase NORM alcanza un AUC de 0.9164, lo que indica una muy buena capacidad del modelo para distinguir registros normales de patológicos. Este resultado refuerza la fiabilidad del modelo en la identificación de la clase de referencia dentro del problema.

Por su parte, la clase HYP presenta un AUC de 0.9025, mostrando una capacidad discriminativa alta, aunque ligeramente inferior a STTC, lo que sugiere una mayor complejidad en la representación de este tipo de alteraciones.

La clase MI obtiene un AUC de 0.9006, lo que refleja una muy buena capacidad del modelo para diferenciar patrones asociados a infarto del resto de clases, manteniendo una separación sólida incluso en una patología clínicamente crítica.

Finalmente, la clase CD presenta un AUC de 0.8986, siendo la más baja del conjunto, aunque sigue situándose en valores elevados. Esto indica que, pese a ser la clase más compleja en términos de separación, el modelo mantiene una capacidad discriminativa muy adecuada.

En conjunto, la red neuronal multicapa muestra un comportamiento estable y coherente entre superclases, con una buena capacidad de discriminación global.

Las matrices de confusión reflejan un modelo robusto, con una ligera tendencia a priorizar la detección de patologías frente a la reducción de falsos positivos, en línea con el objetivo clínico.

En conjunto, el modelo destaca por su capacidad para capturar

relaciones no lineales complejas y mantener una separación consistente entre clases.

De forma global, la Configuración 2 mejora el rendimiento de todos los modelos con respecto a la configuración anterior. En general, todos los clasificadores aumentan su capacidad de discriminación entre superclases, con valores de AUC que se mantienen consistentemente altos, lo que indica que la representación basada en características extraídas junto con las doce derivaciones del ECG es más informativa que la señal en bruto.

La regresión logística sigue siendo el modelo más débil, como era esperable, ya que su capacidad para separar clases no lineales es limitada. Aun así, su mejora en recall respecto a la configuración anterior es importante: el modelo deja de "fallar por falta de sensibilidad" y pasa a detectar más patologías, aunque a costa de más falsos positivos. Es decir, mejora clínicamente en detección, pero no en precisión.

Random Forest se comporta como un modelo intermedio muy sólido. No es el mejor en métricas globales puras, pero sí el más estable: mantiene recalls altos en clases críticas como MI y CD, y no se descompensa tanto como la red neuronal. Su patrón es claro: prioriza sensibilidad en patologías importantes, pero introduce ruido, especialmente en clases solapadas. Esto en clínica suele ser aceptable porque reduce riesgo de no detección.

La red neuronal multicapa es la que mejor rendimiento global alcanza, lo que indica que es el modelo que mejor aprovecha la representación del problema. Sin embargo, su comportamiento no es el más "limpio": tiende a favorecer clases mejor representadas y más fáciles, mientras que en clases más complejas o minoritarias el recall cae. Es decir, aprende muy bien el patrón global, pero no es tan consistente clínicamente en todas las patologías.

En conclusión, aunque la red neuronal gana en rendimiento global, Random Forest ofrece un comportamiento más robusto y controlado

en patologías críticas, lo que lo hace más fiable si el criterio principal es minimizar falsos negativos de forma consistente.

6.2.3. Configuración 3

La Configuración 3 constituye el escenario más completo del sistema, al emplear las doce derivaciones del ECG junto con características extraídas de la señal. En este caso, el problema se plantea como una estimación probabilística sobre 71 patologías, en lugar de una clasificación basada en superclases.

Esto permite estimar la probabilidad de cada patología y capturar relaciones más complejas entre patrones electrocardiográficos, aumentando la riqueza de la información respecto a las configuraciones anteriores.

Se evalúan cinco modelos: Ridge Regression, Elastic Net, Random Forest Regressor, HistGradient Boosting y MLP, lo que permite comparar enfoques lineales, de ensamblado y redes neuronales.

Dado el elevado número de clases, los resultados se presentan de forma resumida. En la tabla solo se incluyen patologías representativas seleccionadas por su relevancia clínica: alteraciones de conducción (CLBBB, CRBBB, LAFB), infarto e isquemia (IMI, ASMI), alteraciones generales del ECG (IVCD, NDT), y clases relevantes como NORM, LVH y PVC. Esto permite interpretar el comportamiento del modelo sin necesidad de mostrar las 71 clases.

El objetivo es evaluar la capacidad de los modelos para estimar probabilidades coherentes en un problema altamente desbalanceado y de alta complejidad clínica.

Ridge Regression

Los resultados de la Tabla 6.15 muestran diferencias claras de rendimiento según el tipo de patología, lo que indica que el comportamiento del modelo está fuertemente condicionado por la complejidad de los patrones electrocardiográficos.

Patología	AUC	MAE	Spearman
NORM	0.919	0.252	0.721
Alteraciones de conducción			
CLBBB	0.995	0.025	0.293
CRBBB	0.998	0.027	0.277
LAFB	0.982	0.074	0.458
Infarto e isquemia			
IMI	0.904	0.109	0.448
ASMI	0.918	0.119	0.462
Alteraciones generales ECG			
IVCD	0.805	0.069	0.187
NDT	0.844	0.155	0.343
Otras clases relevantes			
LVH	0.942	0.088	0.474
PVC	0.969	0.059	0.383

Tabla 6.15: Métricas de Ridge Regressor por patología agrupadas por tipo clínico

En las alteraciones de conducción, el modelo alcanza los mejores resultados globales, con valores de AUC muy elevados. Esto sugiere que estas patologías presentan patrones más estructurados y consistentes en la señal ECG, lo que facilita su discriminación. Sin embargo, el valor de Spearman es más moderado en algunos casos, lo que indica que, aunque la separación entre clases es buena, la calibración de las probabilidades no siempre es completamente proporcional.

En las patologías relacionadas con infarto e isquemia, el rendimiento es ligeramente inferior. Aunque el modelo mantiene una buena capacidad discriminativa, se observa una mayor variabilidad en las correlaciones, lo que refleja el mayor solapamiento clínico y morfológico entre estas clases y otras alteraciones.

Las clases agrupadas como alteraciones generales del ECG presentan el rendimiento más bajo dentro del conjunto, especialmente en términos de AUC y correlación. Esto es coherente con su mayor heterogeneidad interna, lo que dificulta la captura de un patrón común estable.

Por último, las clases restantes muestran un comportamiento intermedio, con buena capacidad de discriminación pero con diferencias en

la calidad de la predicción probabilística según la patología.

En cuanto a las métricas de error, el MAE se mantiene globalmente bajo, lo que indica que el modelo comete errores reducidos en la estimación de probabilidades. No obstante, este comportamiento no implica necesariamente una buena separación entre clases, sino únicamente una calibración razonable de las salidas.

Por otro lado, los valores de Spearman reflejan una coherencia moderada en el ranking de probabilidades, especialmente en patologías con mayor número de muestras, lo que sugiere que el modelo es capaz de ordenar correctamente parte de las clases, aunque con limitaciones en patologías más complejas o minoritarias.

En conjunto, se observa que el modelo Ridge es especialmente eficaz en patologías con patrones bien definidos, mientras que su rendimiento disminuye en aquellas con mayor variabilidad o solapamiento clínico, lo que es consistente con las limitaciones inherentes a los modelos lineales en problemas de alta complejidad.

Elastic Net

Los resultados recogidos en la Tabla 6.16 muestran que Elastic Net presenta un comportamiento muy similar al observado en Ridge Regression, manteniendo una elevada capacidad discriminativa en la mayoría de las patologías analizadas. La clase NORM alcanza valores elevados tanto de AUC como de correlación, lo que indica una buena capacidad para distinguir registros normales y una estimación probabilística consistente.

Patología	AUC	MAE	Spearman
NORM	0.920	0.254	0.723
Alteraciones de conducción			
CLBBB	0.997	0.026	0.296
CRBBB	0.998	0.027	0.278
LAFB	0.983	0.074	0.460
Alteraciones generales ECG			
IVCD	0.825	0.066	0.199
NDT	0.845	0.154	0.344
Infarto / Isquemia			
IMI	0.904	0.110	0.449
ASMI	0.919	0.119	0.461
Clases relevantes			
PVC	0.971	0.059	0.382
LVH	0.946	0.088	0.477

Tabla 6.16: Métricas de Elastic Net por patología agrupadas por tipo clínico

Las alteraciones de conducción continúan siendo el grupo mejor identificado por el modelo. Patologías como CLBBB, CRBBB y LAFB obtienen los valores de AUC más altos del conjunto, lo que sugiere que sus patrones electrocardiográficos resultan fácilmente distinguibles para el modelo. Además, los valores de MAE permanecen reducidos, reflejando una estimación probabilística estable.

Por el contrario, las alteraciones generales del ECG presentan los resultados más modestos. Estas patologías muestran los valores de AUC y Spearman más bajos entre las patologías analizadas, evidenciando una mayor dificultad para discriminar correctamente estas clases. Este comportamiento es coherente con su naturaleza más heterogénea y con el solapamiento existente con otras alteraciones electrocardiográficas.

Las patologías relacionadas con infarto e isquemia mantienen una buena capacidad discriminativa, aunque con errores de predicción superiores a los observados en las alteraciones de conducción. Esto sugiere que la identificación de estas patologías continúa siendo más compleja debido a la similitud de sus manifestaciones electrocardiográficas con otras enfermedades cardiovasculares.

Por último, las clases PVC y LVH presentan un comportamiento sólido en las tres métricas consideradas. Especialmente destacable es el caso de LVH, que muestra uno de los valores de correlación más elevados del conjunto, indicando una buena correspondencia entre las probabilidades estimadas por el modelo y la presencia real de la patología.

En conjunto, Elastic Net mantiene fortalezas y limitaciones muy similares a las observadas en Ridge Regression. La incorporación de regularización L1 no produce mejoras sustanciales en la capacidad discriminativa, aunque contribuye a conservar un comportamiento estable en la mayoría de las patologías analizadas.

Random Forest Regressor

El modelo Random Forest Regressor, cuyos resultados se observan en la Tabla 6.17 muestra un comportamiento global sólido, destacando por una capacidad de discriminación elevada en la mayoría de patologías, especialmente en aquellas con patrones electrocardiográficos bien definidos.

Patología	AUC	MAE	Spearman
NORM	0.935	0.242	0.754
Alteraciones en la conducción			
CLBBB	0.998	0.020	0.277
CRBBB	0.998	0.028	0.261
LAFB	0.984	0.075	0.443
Alteraciones generales de ECG			
IVCD	0.799	0.064	0.182
NDT	0.874	0.143	0.371
Infarto / Isquemia			
IMI	0.879	0.113	0.395
ASMI	0.923	0.118	0.459
Clases relevantes			
LVH	0.952	0.080	0.467
PVC	0.978	0.064	0.373

Tabla 6.17: Métricas de Random Forest Regressor por patología agrupadas por tipo clínico

Las mejores prestaciones se observan en las alteraciones de la conducción, donde el modelo alcanza valores de AUC muy cercanos a la perfección en clases como CLBBB y CRBBB. Este comportamiento sugiere que este tipo de patrones electrocardiográficos presentan una señal bien definida, fácilmente capturable por modelos basados en ensembles de árboles.

En el caso de las alteraciones generales del ECG, el rendimiento disminuye ligeramente, especialmente en métricas de correlación, lo que refleja una mayor heterogeneidad clínica dentro de estas clases. Aun así, el modelo mantiene una capacidad de discriminación aceptable, lo que indica que sigue capturando parte de la estructura subyacente de los datos.

En el grupo de infarto e isquemia, el modelo muestra un comportamiento intermedio, con valores de AUC y correlación algo más moderados. Esto puede explicarse por la mayor variabilidad morfológica de estas patologías en el ECG, lo que dificulta una predicción completamente estable.

Finalmente, en las clases consideradas relevantes como hipertrofia ventricular izquierda (LVH) y extrasístoles ventriculares (PVC), el modelo vuelve a mostrar un rendimiento sólido, con una buena capacidad de discriminación y correlación entre predicción y valor real.

En conjunto, Random Forest se posiciona como uno de los modelos más robustos de la configuración, destacando especialmente por su estabilidad en clases bien definidas, aunque con cierta variabilidad en patologías más heterogéneas.

HistGradientBoosting

Los resultados de la Tabla 6.18 muestran que, de momento, el HistGradienteBoosting presenta el mejor comportamiento global dentro de la Configuración 3, destacando tanto en capacidad de discriminación como en coherencia de las predicciones.

Patología	**AUC**	**MAE**	**Spearman**
NORM	0.981	0.146	0.835
Alteraciones en la conducción			
CLBBB	0.992	0.012	0.380
CRBBB	0.998	0.005	0.389
LAFB	0.997	0.035	0.491
Alteraciones generales de ECG			
IVCD	0.949	0.055	0.294
NDT	0.979	0.093	0.467
Infarto / Isquemia			
IMI	0.989	0.053	0.547
ASMI	0.988	0.055	0.529
Clases relevantes			
LVH	0.954	0.061	0.494
PVC	0.990	0.056	0.368

Tabla 6.18: Métricas de HistGradientBoosting por patología agrupadas por tipo clínico

En términos generales, se observa una mejora clara respecto a modelos basados en regresores lineales y Random Forest, especialmente en

aquellas patologías con patrones electrocardiográficos más definidos. Las alteraciones de la conducción vuelven a ser el grupo con mejor rendimiento, alcanzando valores de AUC prácticamente máximos en CLBBB y CRBBB, lo que confirma que estas clases son fácilmente separables a partir de las características utilizadas.

Una de las diferencias más relevantes respecto a modelos anteriores se encuentra en las métricas de correlación (Spearman), donde se aprecia un incremento consistente en prácticamente todas las categorías. Esto indica que el modelo no solo discrimina correctamente entre presencia y ausencia de patología, sino que además ordena mejor las probabilidades estimadas, lo que resulta especialmente relevante en un contexto probabilístico.

En el caso de las alteraciones generales del ECG y de las patologías de infarto e isquemia, el modelo mantiene un rendimiento elevado y más estable que en enfoques anteriores, reduciendo la variabilidad observada entre clases. Esto sugiere una mejor capacidad de generalización ante patrones más heterogéneos.

Finalmente, en las clases relevantes como LVH y PVC, el modelo vuelve a mostrar valores altos tanto en AUC como en correlación, consolidando su comportamiento robusto en el conjunto del problema.

En conjunto, HistGradientBoosting se posiciona como el modelo más equilibrado de la configuración, combinando alta capacidad discriminativa con una mejora notable en la calidad de la predicción probabilística.

Red Neuronal Multicapa

La Tabla 6.19 muestra los resultados obtenidos en la regresión usando la Red Neuronal Multicapa. Se observa que el modelo alcanza un rendimiento global consistente en el conjunto de test, con una capacidad de discriminación alta en la mayoría de las patologías analizadas.

Patología	AUC	MAE	Spearman
NORM	0.940	0.161	0.761
Alteraciones de conducción			
CLBBB	0.991	0.012	0.305
CRBBB	0.997	0.021	0.289
LAFB	0.985	0.053	0.445
Alteraciones generales ECG			
IVCD	0.790	0.033	0.177
NDT	0.883	0.096	0.380
Infarto / Isquemia			
IMI	0.913	0.073	0.453
ASMI	0.941	0.078	0.481
Clases relevantes			
PVC	0.969	0.045	0.366
LVH	0.933	0.056	0.448

Tabla 6.19: Métricas de la Red Neuronal Multicapa por patología agrupadas por tipo clínico

En términos generales, el modelo obtiene valores de AUC muy altos en prácticamente todas las clases, destacando especialmente en las alteraciones de conducción, donde se alcanzan valores cercanos a 1 en CLBBB y CRBBB, lo que indica una separación muy clara entre casos positivos y negativos. Este comportamiento sugiere que el modelo es especialmente eficaz en patologías con patrones eléctricos bien estructurados.

En el grupo de infarto e isquemia, el modelo mantiene un rendimiento consistente, con AUC elevados en IMI y ASMI, acompañado de valores de Spearman moderados, lo que indica una buena capacidad no solo de clasificación, sino también de preservación del orden relativo de las predicciones.

En contraste, las alteraciones generales del ECG, como IVCD y NDT, presentan un rendimiento inferior, con una disminución tanto en AUC como en correlación de Spearman. Esto refleja la mayor complejidad de estas clases, donde los patrones electrocardiográficos son menos definidos y más heterogéneos.

En conjunto, la red neuronal multicapa se posiciona como uno de los modelos más equilibrados, mostrando mejoras claras frente a enfoques más lineales, especialmente en términos de capacidad de modelado no lineal y consistencia en la correlación entre predicción y valor real.

De forma global, la Configuración 3 plantea el escenario más complejo de estudio, ya que plantea una predicción probabilística sobre las 71 patologías. A pesar de ello, todos los modelos mantienen una capacidad discriminativa elevada en la mayoría de las clases, con valores de AUC generalmente altos.

Los enfoques lineales presentan un rendimiento más limitado, con tendencia a generar predicciones conservadoras concentradas en valores bajos, lo que reduce la sensibilidad en varias patologías, especialmente en los casos positivos.

Sin embargo, los modelos no lineales presentan un comportamiento superior. Random Forest presenta un rendimiento estable y robusto, aunque tiene una capacidad de calibración probabilística más limitada. HistGradientBoosting mejora tanto la discriminación como la coherencia en las predicciones, siendo así uno de los modelos más competitivos.

La Red Neuronal Multicapa obtiene el mejor rendimiento global, destacando por una AUC muy elevada y una mejor capacidad para adaptarse a la complejidad del problema, manteniendo un comportamiento más equilibrado entre patologías.

En conjunto, las alteraciones de conducción son las mejor resueltas por todos los modelos, mientras que las alteraciones generales del ECG presentan la mayor dificultad. Las patologías de infarto e isquemia se sitúan en un nivel intermedio, con mejor desempeño en los modelos no lineales.

En conclusión, la Configuración 3 confirma que la Red Neuronal Multicapa es el modelo más adecuado en este escenario, al combinar la mejor capacidad discriminativa con el comportamiento más consistente.

6.2.4. Limitaciones del análisis

Este análisis presenta ciertas limitaciones que condicionan la interpretación de los resultados obtenidos. En primer lugar, el desbalance de clases afecta de forma directa al comportamiento de los modelos, haciendo que en muchos casos se priorice la reducción del error global frente a la detección de las patologías menos representadas.

Por otro lado, se observa que no todas las métricas reflejan de la misma forma el rendimiento clínico de los modelos. Medidas como el MAE pueden resultar engañosas, ya que un error medio bajo no garantiza una buena capacidad de detección de casos positivos, especialmente cuando los modelos tienden a hacer predicciones conservadoras.

Además, la complejidad creciente del problema a lo largo de las configuraciones, especialmente en el caso de la formulación con 71 patologías, introduce un mayor grado de solapamiento entre clases, lo que dificulta la separación clara entre algunas patologías y genera variabilidad en el comportamiento de los modelos.

Por último, la heterogeneidad propia de algunas clases clínicas reduce la capacidad de generalización del modelo, generando variaciones significativas en el rendimiento según la patología analizada.

Capítulo 7

Dashboard interactivo

El sistema desarrollado se complementa con un *dashboard* interactivo diseñado para facilitar la visualización de señales electrocardiográficas y la interpretación de las predicciones generadas por el modelo final seleccionado en la Configuración 3. En este caso, el modelo integrado corresponde a la Red Neuronal Multicapa (MLP), al ser el que presenta el mejor equilibrio entre capacidad discriminativa y coherencia en las predicciones probabilísticas.

La herramienta permite la exploración individual de registros ECG de forma estructurada, organizando la información en distintas secciones que integran representación de la señal, análisis clínico y resultados del modelo, con el objetivo de facilitar su interpretación conjunta.

7.1. Resumen clínico del paciente

En primer lugar, el dashboard genera un resumen clínico automático a partir de características extraídas de la señal electrocardiográfica. Este bloque permite una interpretación rápida del estado del paciente mediante variables relevantes desde el punto de vista fisiológico y clínico:

- **Frecuencia cardíaca (HR):** expresada en latidos por minuto (bpm), calculada a partir de los intervalos RR medios, permitiendo identificar situaciones de bradicardia o taquicardia.

- **Intervalo RR medio:** medido en segundos, representa la distancia temporal entre picos R consecutivos y refleja la regularidad del ritmo cardíaco.
- **Variabilidad cardíaca (HRV - SDNN):** desviación estándar de los intervalos RR, utilizada como indicador de la variabilidad del sistema nervioso autónomo.
- **Eje eléctrico cardíaco:** expresado en grados, describe la orientación media de la actividad eléctrica ventricular, siendo útil para detectar desviaciones asociadas a alteraciones de conducción o hipertrofias.
- **Duración del QRS:** medida en número de muestras, representa el tiempo de despolarización ventricular.
- **Segmento ST:** variable de amplitud asociada a posibles alteraciones isquémicas cuando presenta elevación o depresión.
- **Amplitud del pico R:** valor medio del pico R en cada latido, relacionado con la intensidad de la despolarización ventricular.
- **Número de latidos detectados:** conteo de complejos QRS identificados, útil para evaluar la calidad del registro.
- **Edad y sexo:** variables demográficas que aportan contexto clínico adicional.

Este resumen proporciona una visión general del paciente antes de analizar la señal completa o las predicciones del modelo.

7.2. Visualización del electrocardiograma

A continuación, el dashboard muestra el electrocardiograma completo en sus 12 derivaciones. Esta representación permite observar la

actividad eléctrica del corazón desde diferentes perspectivas, lo que resulta esencial para la detección de anomalías específicas que pueden no ser visibles en una única derivación.

La visualización está sincronizada entre derivaciones, lo que permite analizar la coherencia temporal de los eventos cardíacos y facilita la identificación de patrones característicos de distintas patologías.

7.3. Latidos segmentados por derivación

El sistema también incluye una representación de los latidos individuales segmentados por derivación. Esta vista permite aislar complejos QRS y analizar su morfología de forma independiente.

Esta segmentación resulta especialmente útil para identificar variaciones en la forma del latido entre diferentes regiones del corazón, lo que puede aportar información relevante en patologías como bloqueos de rama, hipertrofias o alteraciones de la repolarización.

7.4. Predicción del modelo

Finalmente, el dashboard incorpora la salida del modelo de aprendizaje automático, basado en la Red Neuronal Multicapa de la Configuración 3. El sistema proporciona un diagnóstico probabilístico para las distintas patologías consideradas en el estudio.

Esta salida permite interpretar el resultado no como una única clase, sino como un conjunto de probabilidades asociadas a cada patología, lo que resulta especialmente adecuado en un contexto clínico donde pueden coexistir múltiples alteraciones.

La presentación de estas predicciones facilita la comprensión del comportamiento del modelo y su relación con las características extraídas del ECG.

7.5. Generación de informe

Como complemento final, el dashboard permite la generación y descarga de un informe estructurado en formato PDF. Este informe incluye la información del paciente, el resumen clínico, las visualizaciones del ECG y las predicciones del modelo.

Este documento está diseñado para facilitar su uso como soporte de análisis, permitiendo conservar los resultados obtenidos de forma organizada y reutilizable fuera del entorno interactivo.

Capítulo 8

Conclusiones y Trabajos futuros

8.1. Conclusiones

1. La clasificación automática de patologías cardíacas mediante aprendizaje automático es viable y clínicamente razonable. Los modelos desarrollados han demostrado capacidad para discriminar entre superclases de patologías y para estimar probabilidades sobre 71 condiciones cardíacas, obteniendo valores de AUC consistentemente altos. Esto confirma que el aprendizaje automático constituye una herramienta válida para el apoyo al diagnóstico electrocardiográfico, especialmente como sistema de análisis preliminar del ECG. Su objetivo no es sustituir al profesional médico, sino servir como apoyo en la detección temprana y priorización de posibles patologías.

2. El uso de las doce derivaciones es imprescindible para un rendimiento adecuado. Los experimentos de la Configuración 1 demuestran que limitar el sistema a una única derivación reduce significativamente la capacidad discriminativa de todos los modelos. La incorporación de las doce derivaciones en la Configuración 2 y 3 produce una mejora generalizada y consistente, confirmando que la información espacial del electrocardiograma es esencial para capturar la variedad de patrones asociados a las distintas patologías cardíacas.

3. La extracción de características mejora la representación de la señal frente al uso del megavector en bruto. La combinación de características temporales, morfológicas y estadísticas proporciona una representación más compacta, informativa y manejable que permite a los modelos generalizar mejor. Frente al megavector de la señal en crudo, la representación basada en características reduce la dimensionalidad del problema sin pérdida de información relevante, lo que se traduce en mejoras tanto en rendimiento como en eficiencia computacional.

4. Los modelos no lineales superan consistentemente a los enfoques lineales. La regresión logística y los enfoques basados en regularización lineal (Ridge, Elastic Net) presentan el rendimiento más limitado en todos los escenarios, mostrando tendencia a generar predicciones conservadoras que reducen la sensibilidad en las patologías minoritarias. En contraste, Random Forest, HistGradientBoosting y la red neuronal multicapa son capaces de capturar relaciones complejas entre características, lo que se traduce en una mejora significativa en las métricas de clasificación y discriminación.

5. Random Forest ofrece el comportamiento más robusto y controlado desde una perspectiva clínica. Aunque no es el modelo con mayor rendimiento global, Random Forest destaca por mantener recalls altos de forma consistente en las patologías de mayor riesgo clínico, como el infarto de miocardio (MI) y las alteraciones de conducción (CD). Su patrón de decisión tiende a priorizar la sensibilidad frente a la precisión, reduciendo el riesgo de falsos negativos, lo cual resulta especialmente valioso en entornos clínicos donde no detectar una patología grave supone un riesgo mayor que una alarma falsa.

6. La red neuronal multicapa obtiene el mejor rendimiento global, pero con menor consistencia clínica. El MLP es el modelo que

mejor aprovecha la representación del problema y el que alcanza los valores más altos de AUC en prácticamente todas las configuraciones. Sin embargo, su comportamiento no es homogéneo entre clases: tiende a favorecer las patologías mejor representadas en el conjunto de datos, mientras que el recall cae en clases complejas o minoritarias. Por ello, su uso debe evaluarse cuidadosamente en función del contexto clínico y los criterios de evaluación más relevantes.

7. La elección del modelo debe guiarse por el criterio clínico prioritario, no únicamente por las métricas globales. Un modelo con mayor AUC no es necesariamente el más adecuado para una aplicación clínica real. Si el objetivo es minimizar los falsos negativos en patologías críticas, Random Forest resulta preferible. Si se busca el mejor rendimiento global y una buena calibración probabilística, HistGradientBoosting o el MLP son las opciones más adecuadas. Esta distinción es fundamental para el diseño de sistemas de apoyo al diagnóstico.

8. El desbalance de clases y la heterogeneidad de los patrones son las principales limitaciones del sistema. El desbalance entre patologías afecta directamente al comportamiento de los modelos, que tienden a priorizar la reducción del error global frente a la detección de las clases menos representadas. Además, la heterogeneidad de ciertas condiciones clínicas introduce solapamiento que dificulta la separación limpia entre clases y genera variabilidad en los resultados.

9. El sistema desarrollado sienta una base sólida para su extensión hacia herramientas clínicas reales. La integración de los modelos en el dashboard interactivo desarrollado permite visualizar las probabilidades de cada patología de forma intuitiva, acercando el sistema a un entorno de apoyo al diagnóstico real. Los resultados obtenidos, combinados con la modularidad del diseño, ofre-

cen una plataforma válida sobre la que construir mejoras futuras, incorporar nuevas fuentes de datos o integrar arquitecturas más avanzadas.

8.2. Trabajos futuros

A partir de los resultados obtenidos y las limitaciones identificadas a lo largo del trabajo, se plantean las siguientes líneas de mejora y extensión:

1. Incorporación de early stopping en la red neuronal multicapa. Durante el entrenamiento del MLP se ha observado que el modelo puede beneficiarse de una parada anticipada basada en la evolución de la pérdida en validación, evitando el sobreajuste en configuraciones de alta complejidad como la Configuración 3. La incorporación de early stopping permitiría obtener modelos más generalizables sin necesidad de aumentar la regularización manual, mejorando especialmente el comportamiento en las patologías minoritarias.

2. Exploración de redes neuronales convolucionales (CNN). Las CNN han demostrado ser especialmente eficaces en el procesamiento de señales temporales como el ECG, al ser capaces de aprender automáticamente patrones locales relevantes directamente desde la señal en bruto. Durante el desarrollo de este trabajo se inició una primera aproximación a su implementación, si bien no llegó a completarse ni a producir resultados evaluables dentro del alcance del proyecto. Como trabajo futuro, se propone retomar esta línea, completar el diseño de la arquitectura convolucional y comparar su rendimiento frente a los modelos desarrollados, con especial atención a las patologías con patrones morfológicos más complejos donde la extracción manual de características puede resultar insuficiente.

3. Optimización de umbrales de decisión por patología. A lo largo de las tres configuraciones, se ha utilizado un umbral de decisión personalizado para todas las clases, lo que no resultó del todo óptimo dado el elevado desbalance existente entre patologías. Una línea natural de mejora consistiría en optimizar el umbral de forma individual para cada clase, buscando maximizar métricas como el recall, sin penalizar en exceso la precisión. En las Configuraciones 1 y 2 esto permitiría ajustar el criterio de clasificación por superclase, mientras que en la Configuración 3 aportaría una mayor granularidad al poder adaptar el umbral a cada una de las 71 patologías según su gravedad clínica y su representación en el conjunto de datos.

4. Soporte para señales ECG de longitud variable. El sistema actual requiere que los registros ECG tengan una longitud fija, lo que limita su aplicabilidad en entornos clínicos reales donde la duración del registro puede variar. Adaptar el pipeline para procesar señales de longitud arbitraria ampliaría significativamente la utilidad del sistema y facilitaría su integración con equipos de adquisición heterogéneos.

5. Despliegue e integración clínica del sistema. Como línea final, se propone evolucionar el dashboard interactivo desarrollado hacia una herramienta clínica desplegable en entornos reales, con capacidad para recibir registros ECG directamente desde equipos de adquisición, generar informes automáticos con las probabilidades de cada patología y alertar al personal sanitario ante la detección de condiciones de alto riesgo. Este paso requeriría además una validación regulatoria y la colaboración con profesionales médicos para ajustar el sistema a los flujos de trabajo clínicos existentes.

Bibliografía

Adetiba, E., Iweanya, V. C., Popoola, S. I., Adetiba, J. N., & Menon, C. (2017). Automated detection of heart defects in athletes based on electrocardiography and artificial neural network [Accedido: 2025-11-07]. *Cogent Engineering*, *4*(1), 1411220. https://doi.org/10.1080/23311916.2017.1411220

Amit, H. (2024). Evaluation metrics for classification models [Accedido: 2026-04-24]. https://medium.com/biased-algorithms/evaluation-metrics-for-classification-models-b995f9980716

Attal, M. (2026). Entender la curva ROC AUC: métricas esenciales para la evaluación de modelos de Machine Learning [Accedido: 2026-04-24]. https://liora.io/es/curva-roc-auc-todo-sobre

Ayano, Y. M., Schwenker, F., Dufera, B. D., & Debelee, T. G. (2022). Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review [Accedido: 2026-04-17]. *Diagnostics (Basel)*, *13*(1), 111. https://doi.org/10.3390/diagnostics13010111

Bergquist, J. A., Zenger, B., Brundage, J., MacLeod, R. S., Shah, R., Ye, X., Lyones, A., Ranjan, R., Tasdizen, T., Bunch, T. J., & Steinberg, B. A. (2023). Comparison of Machine Learning Detection of Low Left Ventricular Ejection Fraction Using Individual ECG Leads [Accedido: 2026-04-17]. *Computing in Cardiology*, *50*, 1. https://doi.org/10.22489/CinC.2023.047

Buhl, N. (2023). F1 Score in Machine Learning [Accedido: 2026-04-24]. https://encord.com/blog/f1-score-in-machine-learning/

C, P. (2025). Mastering RandomForestRegressor in Scikit-learn: A Practical Guide (ML Quickies #3) [Accedido: 2026-05-16]. https://medium.com/@prathik.codes/mastering-randomforestregressor-in-scikit-learn-a-practical-guide-ba8615097100

Chauhan, C., Agrawal, M., & Shabherwal, P. (2021). A Multi-Lead Fusion Method for the Accurate Delineation of QRS Complex Location in 12 Lead ECG Signal [Accedido: 2025-12-28]. *arXiv preprint arXiv:2107.05469*. https://arxiv.org/abs/2107.05469

Chugh, A. (2020). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? [Accedido: 2026-04-24]. https://

medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Cigna. (s.f.). El corazón y su sistema eléctrico [Accedido: 2026-01-15]. https://www.cigna.com/es-us/knowledge-center/hw/el-corazn-y-su-sistema-elctrico-zm2272

Codificando Bits. (s.f.). Función de activación [Accedido: 2026-05-24]. https://codificandobits.com/blog/funcion-de-activacion/

Durán, J. (2019). Técnicas de regularización básicas para redes neuronales [Accedido: 2026-03-27]. https://medium.com/metadatos/t%C3%A9cnicas-de-regularizaci%C3%B3n-b%C3%A1sicas-para-redes-neuronales-b48f396924d4

Eje cardíaco en el ECG: cálculo e interpretación [Accedido: 2026-04-10]. (2026). https://www.cardioteca.com/imagen/8008-eje-cardiaco-ecg-calculo.html

European Society of Cardiology. (2024). Artificial intelligence in ECG diagnostics - where are we now? [Accedido: 2026-04-17]. https://www.escardio.org/communities/councils/cardiology-practice/education/cardiopractice/artificial-intelligence-in-ecg-diagnostics-where-are-we-now/

Evidently AI. (2025). Accuracy, precision, recall and related classification metrics [Accedido: 2026-04-24]. https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

GeeksforGeeks. (2026). Random Forest Regression in Python [Accedido: 2026-05-16]. https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network [Accedido: 2026-04-17]. *Nature Medicine*, *25*(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

IBM. (s.f.). ¿Qué es Random Forest? [Accedido: 2026-05-16]. https://www.ibm.com/es-es/think/topics/random-forest

Interactive Chaos. (s.f.). Elastic Net [Accedido: 2026-05-16]. https://interactivechaos.com/es/manual/tutorial-de-machine-learning/elastic-net

Kalmady, S. V., Salimi, A., Sun, W., Sepehrvand, N., Nademi, Y., Bainey, K., Ezekowitz, J., Hindle, A., McAlister, F., Greiner, R., Sandhu, R., & Kaul, P. (2024). Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level [Accedido: 2026-04-17]. *npj Digital Medicine*, *7*, 133. https://doi.org/10.1038/s41746-024-01133-?

Klabunde, R. E. (s.f.-a). Electrocardiogram Augmented Limb Leads (Unipolar) [Accedido: 2025-11-07]. https://cvphysiology.com/arrhythmias/a013b

Klabunde, R. E. (s.f.-b). Electrocardiogram Chest Leads (Unipolar) [Accedido: 2025-11-07]. https://cvphysiology.com/arrhythmias/a013c

Klabunde, R. E. (s.f.-c). Electrocardiogram Standard Limb Leads (Bipolar) [Accedido: 2025-11-07]. https://cvphysiology.com/arrhythmias/a013a

Klabunde, Richard E. (2023). Electrocardiogram Leads [Accedido: 2025-11-07]. https://cvphysiology.com/arrhythmias/a013

KoshurAI. (2024). Understanding Log Loss: A Comprehensive Guide with Code Examples [Accedido: 2026-04-24]. https://koshurai.medium.com/understanding-log-loss-a-comprehensive-guide-with-code-examples-c79cf5411426

Lee, F. (2025). ¿Qué es la regresión logística? [Accedido: 2026-05-16]. https://www.ibm.com/es-es/think/topics/logistic-regression

Lee, Fangfang. (2021). ¿Qué son las redes neuronales? [Accedido: 2026-05-16]. https://www.ibm.com/es-es/think/topics/neural-networks

Lera, L., Leyton, B., & Lizana, P. A. (2025). La regresión logística y su aplicación en la investigación biomédica [Accedido: 2026-05-16]. *International Journal of Morphology*, *43*(5). https://doi.org/10.4067/s0717-95022025000501545

McCaffrey, J. (2023). Regression Using a scikit MLPRegressor Neural Network [Accedido: 2026-05-16]. https://visualstudiomagazine.com/articles/2023/05/01/regression-scikit.aspx

MedlinePlus. (s.f.). Electrocardiograma [Accedido: 2025-11-06]. https://medlineplus.gov/spanish/pruebas-de-laboratorio/electrocardiograma/

Murel, J., & Kavlakoglu, E. (s.f.-a). ¿Qué es la regresión Ridge? [Accedido: 2026-05-16]. https://www.ibm.com/es-es/think/topics/ridge-regression

Murel, J., & Kavlakoglu, E. (s.f.-b). ¿Qué es una matriz de confusión? [Accedido: 2026-04-24]. https://www.ibm.com/es-es/think/topics/confusion-matrix

My-EKG. (s.f.-a). Derivaciones cardíacas [Accedido: 2025-11-07]. https://www.my-ekg.com/generalidades-ekg/derivaciones-cardiacas.php

My-EKG. (s.f.-b). Dilatación de aurícula izquierda [Accedido: 2025-11-07]. https://www.my-ekg.com/hipertrofia-dilatacion/dilatacion-auricula-izquierda.php

PhysioNet. (2005). MIT-BIH Arrhythmia Database [Dataset disponible en PhysioNet. Accedido: 2026-04-17]. https://physionet.org/content/mitdb/1.0.0/

Picard. (2024). Hist Gradient Boosting Regressor (Scikit-learn) [Accedido: 2026-05-16]. https://datasciencepythonblog.net/hist-gradient-boosting-regressor-scikit-learn/

Prakash, A. J., Belkacem, A. N., Elfadel, I. M., Jelinek, H. F., & Atef, M. (2025). Advances in machine and deep learning for ECG beat classification: a systematic review [Accedido: 2026-04-17]. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1649923

Prutkin, J. M., Goldberger, A. L., & Botkin, N. F. (2025). ECG tutorial: Basic principles of ECG analysis [Topic 2115, Version 36.0. Accedido: 2025-11-21].

PyTorch contributors. (s.f.). torch.nn.BCEWithLogitsLoss [Accedido: 2026-05-16]. https://docs.pytorch.org/docs/2.12/generated/torch.nn.BCEWithLogitsLoss.html

Q2B Studio. (2026). Redes neuronales interpretables para analizar electrocardiogramas de 12 derivaciones: diagnóstico de enfermedades cardíacas [Accedido: 2026-03-27]. https://www.q2bstudio.com/nuestro-blog/677225/redes-neuronales-interpretables-para-analizar-electrocardiogramas-12-derivaciones-diagnostico-enfermedades-cardiacas

Qlik. (s.f.). Scoring multiclass classification [Accedido: 2026-04-24]. https://help.qlik.com/es-ES/cloud-services/Subsystems/Hub/Content/Sense_Hub/AutoML/scoring-multiclass-classification.htm

Repsol. (2026). Machine Learning: innovación y aplicaciones [Accedido: 2026-04-24]. https://www.repsol.com/es/energia-avanzar/innovacion/machine-learning/index.cshtml

Riaux, A. (2024). Mastering HistGBM — Sklearn Implementation [Accedido: 2026-05-16]. https://medium.com/@adrien.riaux/mastering-histgbm-sklearn-implementation-d3a19fbe729d

Salvavidas. (2025). Diferencias de género en enfermedades cardíacas [Accedido: 2026-05-24]. https://salvavidas.com/blog/diferencias-de-genero-en-enfermedades-cardiacas/

Scikit-learn developers. (s.f.-a). ElasticNet [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

Scikit-learn developers. (s.f.-b). HistGradientBoostingRegressor [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html

Scikit-learn developers. (s.f.-c). log_loss [Accedido: 2026-04-24]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

Scikit-learn developers. (s.f.-d). LogisticRegression [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit-learn developers. (s.f.-e). OneVsRestClassifier [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

Scikit-learn developers. (s.f.-f). Ridge [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Scikit-learn developers. (s.f.-g). sklearn.ensemble.RandomForestClassifier [Accedido: 2026-05-16]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Sociedad Argentina de Cardiología (SADEC). (s.f.). El sistema eléctrico del corazón y qué es una arritmia [Accedido: 2026-01-15]. https://www.sociedadsadec.org.ar/el-sistema-electrico-del-corazon-que-es-una-arritmia/

Stanford Children's Health. (s.f.-a). Anatomy and Function of the Electrical System [Accedido: 2026-01-15]. https://www.stanfordchildrens.org/es/topic/default?id=anatomy-and-function-of-the-electrical-system-90-P04865

Stanford Children's Health. (s.f.-b). Anatomy and Function of the Heart's Electrical System [Accedido: 2026-01-15]. https://www.stanfordchildrens.org/es/topic/default?id=anatomy-and-function-of-the-hearts-electrical-system-85-P03337

Talebi, S. (2025). Fourier vs. Wavelet Transform: What's the Difference? [Accedido: 2025-12-28]. https://builtin.com/data-science/wavelet-transform

University of Utah - CEG. (s.f.). Machine Learning Research Area [Accedido: 2026-04-17]. https://ceg.sci.utah.edu/research-areas/machine-learning/

Vimal, C., & Sathish, B. (2010). Random Forest Classifier Based ECG Arrhythmia Classification [Accedido: 2026-04-17]. *International Journal of E-Health and Medical Communications (IGI Global)*. https://www.igi-global.com/article/random-forest-classifier-based-ecg/42992

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45

Wei, D. (2024). Essential Math for Machine Learning: Spearman's Rank Correlation [Accedido: 2026-04-24]. https://medium.com/@weidagang/essential-math-for-machine-learning-spearmans-rank-correlation-80e023045c96

Wysokinski, W. E., Meverden, R. A., Lopez-Jimenez, F., Harmon, D. M., Medina Inojosa, B. J., Baez Suarez, A., Liu, K., Medina Inojosa, J. R., Casanegra, A. I., McBane, R. D., & Houghton, D. E. (2024). Electrocardiogram Signal Analysis With a Machine Learning Model Predicts the Presence of Pulmonary Embolism With Accuracy Dependent on Embolism Burden [Accedido: 2026-04-17]. *Mayo Clinic Proceedings: Digital Health*, *2*(3), 453-462. https://doi.org/10.1016/j.mcpdig.2024.03.009

Zou, C., Muller, A., Wolfgang, U., Ruckert, D., Muller, P., Becker, M., Steger, A., & Martens, E. (2022). Heartbeat Classification by Random Forest With a Novel Context Feature: A Segment Label [Accedido: 2026-05-16]. *IEEE Journal of Translational Engineering in Health and Medicine*, *10*, 1900508. https://doi.org/10.1109/JTEHM.2022.3202749

Anexo I

El presente Trabajo Fin de Grado se alinea con varios de los Objetivos de Desarrollo Sostenible (ODS) establecidos por la Agenda 2030 de Naciones Unidas, principalmente debido a su aplicación en el ámbito de la salud cardiovascular mediante técnicas de aprendizaje automático. A continuación, se detalla la contribución del proyecto a cada uno de los ODS identificados como relevantes.

ODS 3: Salud y Bienestar

Este es el objetivo con el que el proyecto guarda una relación más directa. Las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad a nivel mundial, y su detección temprana es determinante para reducir complicaciones y mejorar el pronóstico de los pacientes. El sistema desarrollado en este trabajo, capaz de analizar señales de electrocardiograma y estimar la presencia de distintas patologías cardíacas, contribuye directamente a la meta 3.4 de los ODS, orientada a reducir la mortalidad prematura por enfermedades no transmisibles mediante la prevención y el tratamiento. Asimismo, al tratarse de una herramienta de apoyo al diagnóstico y no de sustitución del criterio médico, el proyecto refuerza la meta 3.8, relativa al acceso a servicios de salud de calidad, al facilitar al personal clínico una herramienta adicional para la interpretación de registros ECG.

ODS 9: Industria, Innovación e Infraestructura

El trabajo aplica técnicas avanzadas de aprendizaje automático y procesamiento de señales a un problema biomédico real, lo que representa una contribución a la meta 9.5, centrada en el fomento de la investigación científica y la mejora de las capacidades tecnológicas en sectores industriales y de salud. El desarrollo de un sistema automático de análisis de ECG, junto con una herramienta de visualización interactiva, supone una aportación a la infraestructura tecnológica disponible para el apoyo al diagnóstico clínico.

ODS 10: Reducción de las Desigualdades

El electrocardiograma es una prueba no invasiva, de bajo coste y de fácil acceso, lo que la convierte en una de las herramientas diagnósticas más extendidas, incluso en entornos con recursos limitados. Al automatizar parte de su interpretación, este trabajo contribuye potencialmente a reducir la dependencia de personal médico especializado para un primer análisis del ECG, lo que puede resultar especialmente valioso en sistemas sanitarios con escasez de cardiólogos o en zonas con menor acceso a atención especializada, favoreciendo así una mayor equidad en el acceso a herramientas de apoyo diagnóstico.

Anexo II

Este anexo recoge la información completa relativa a la base de datos del sistema. Se incluye la correspondencia entre las siglas utilizadas para identificar las patologías y su significado clínico, agrupadas por superclases, con el objetivo de facilitar la interpretación de los resultados.

De este modo, el anexo complementa el contenido del documento principal, proporcionando una referencia completa y estructurada de las patologías analizadas en la Configuración 3.

Tabla 1: Descripción de las patologías agrupadas por superclases

Código	**Descripción**
Cambios del segmento ST-T (STTC)	
NDT	Anomalías de la onda T no diagnósticas
NST_	Cambios del segmento ST no específicos
DIG	Efecto digitálico
LNGQT	Intervalo QT prolongado
ISC_	Isquemia no específica
ISCAL	Isquemia en derivaciones anterolaterales
ISCIN	Isquemia en derivaciones inferiores
ISCIL	Isquemia en derivaciones inferolaterales
ISCAS	Isquemia en derivaciones anteroseptales
ISCAN	Isquemia en derivaciones anteriores
ANEUR	Cambios compatibles con aneurisma ventricular
EL	Alteración electrolítica o farmacológica
STD_	Depresión del segmento ST no específica

Código	Descripción
STE_	Elevación del segmento ST no específica
TAB_	Alteración de la onda T
LOWT	Ondas T de baja amplitud
NT_	Cambios inespecíficos de la onda T
INVT	Inversión de la onda T
Alteraciones de conducción (CD)	
LAFB	Bloqueo fascicular anterior izquierdo
IRBBB	Bloqueo incompleto de rama derecha
1AVB	Bloqueo AV de primer grado
IVCD	Trastorno inespecífico de conducción intraventricular
CRBBB	Bloqueo completo de rama derecha
CLBBB	Bloqueo completo de rama izquierda
WPW	Síndrome de Wolff-Parkinson-White
ILBBB	Bloqueo incompleto de rama izquierda
LPFB	Bloqueo fascicular posterior izquierdo
3AVB	Bloqueo AV de tercer grado
2AVB	Bloqueo AV de segundo grado
Infarto de miocardio (MI)	
IMI	Infarto de miocardio inferior
ASMI	Infarto anteroseptal
ILMI	Infarto inferolateral
AMI	Infarto anterior
ALMI	Infarto anterolateral
INJAS	Lesión subendocárdica anteroseptal
LMI	Infarto lateral
IPLMI	Infarto inferoposterolateral
IPMI	Infarto inferoposterior
INJIN	Lesión subendocárdica inferior
INJLA	Lesión subendocárdica lateral
PMI	Infarto posterior
INJIL	Lesión subendocárdica inferolateral

Código	Descripción
Hipertrofias y sobrecargas (HYP)	
LVH	Hipertrofia ventricular izquierda
LAO/LAE	Sobrecarga auricular izquierda
RVH	Hipertrofia ventricular derecha
RAO/RAE	Sobrecarga auricular derecha
SEHYP	Hipertrofia septal
Normal	
NORM	Electrocardiograma normal

Repositorio del proyecto

El código fuente desarrollado durante este Trabajo Fin de Grado se encuentra disponible en el siguiente repositorio de GitHub:

https://github.com/AArreguiB/analisis_ecg.git