Universitat Oberta de Catalunya

uoc.edu

Soporte al diagnóstico de patologías cardíacas estructurales mediante ECG, inteligencia artificial explicable (XAI) y Cuantificación de Incertidumbre (UQ)

Raúl Checa Marín
Grado de Ingeniería Informática
Inteligencia Artificial

Consultora: Dra. María Moreno de Castro

Profesor responsable de la asignatura: Dr. Friman Sánchez Castaño

Fecha de entrega: 01/2026

Fecha de la defensa: 20/01/2026

FICHA DEL TRABAJO FINAL

Título del trabajo:	**Soporte al diagnóstico de patologías cardíacas estructurales mediante ECG, inteligencia artificial explicable (XAI) y Cuantificación de Incertidumbre (UQ)**
Nombre del autor:	**Raúl Checa Marín**
Nombre de la consultora:	**Dra. María Moreno de Castro**
Nombre del PRA:	**Dr. Friman Sánchez Castaño**
Fecha de entrega:	**01/2026**
Titulación:	**Ingeniería Informática**
Área del Trabajo Final:	**Inteligencia Artificial**
Idioma del trabajo:	**Castellano**
Palabras clave	**Clasificación multiclase, Clases desbalanceadas, Medicina personalizada.**

Resumen del Trabajo (máximo 250 palabras): Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.

Este trabajo presenta el desarrollo de un Sistema de Soporte a la Decisión Clínica para clasificar patologías cardíacas estructurales a partir de datos tabulares demográficos y series temporales de electrocardiogramas (ECG). Abordando el desafío del severo desbalance de clases sin recurrir a datos sintéticos, se evaluaron múltiples arquitecturas, seleccionando CatBoost con ponderación de clases como el modelo óptimo tras una optimización bayesiana de los hiperparámetros del modelo clasificador.

Priorizando la seguridad del paciente y los requisitos de la IA confiable, el sistema trasciende la predicción puntual tradicional integrando Predicción Conforme. Esta técnica permite cuantificar la incertidumbre mediante conjuntos de predicción con garantías estadísticas formales, generando conjuntos de predicción con garantías estadísticas formales de cobertura, que contienen la clase verdadera con una probabilidad controlada.

Asimismo, se incorporó una capa de Explicabilidad (XAI) mediante SHAP y contrafactuales, validando que las decisiones del algoritmo se fundamentan en marcadores fisiológicos coherentes. Los resultados confirman la viabilidad de esta arquitectura como herramienta eficaz, auditable y alineada con los principios de la IA confiable auditable para poder ser aplicada en medicina personalizada, triaje clínico y priorización de listas de espera.

Abstract (in English, 250 words or less):

This thesis presents the development of a Clinical Decision Support System designed to classify structural cardiac pathologies using demographic tabular data and electrocardiogram (ECG) time series. Addressing the critical challenge of severe class imbalance without relying on synthetic data generation, multiple architectures were evaluated. CatBoost, employing a class weighting strategy, was selected as the optimal model following Bayesian optimization hyperparameter optimization.

Prioritizing patient safety and adhering to Trustworthy AI principles, the system advances beyond traditional point prediction by integrating Conformal Prediction. This technique enables uncertainty quantification through prediction sets offering formal statistical guarantees, providing a confidence measure that contains the true class.

Furthermore, an Explainability (XAI) layer was incorporated using SHAP and counterfactual analysis, validating that the algorithm's decisions rely on physiologically consistent markers. The results confirm the architecture's viability as an effective and auditable tool aligned with the principles of Trustworthy AI, with potential application in personalized medicine, clinical triage, and waiting list prioritization.

Índice

Figuras

Tablas

1. Introducción

1.1. Contexto y justificación del trabajo

Las enfermedades isquémicas del corazón constituyen la principal causa de mortalidad a nivel mundial [1] [2] y representan un desafío clínico de gran relevancia por su impacto en la salud pública. La detección temprana de estas patologías estructurales del corazón es determinante para reducir la morbilidad y mejorar la supervivencia de las personas pacientes. En este contexto, el electrocardiograma (ECG) es una herramienta diagnóstica fundamental, económica y no invasiva, capaz de reflejar alteraciones eléctricas asociadas a diferentes trastornos cardíacos [3] [4].

No obstante, la interpretación del ECG requiere una especialización y puede estar sujeta a la persona encargada de la observación, especialmente en casos complejos o con señales ruidosas. Esta dependencia del criterio experto puede retrasar el diagnóstico y aumentar la carga asistencial del personal sanitario.

En los últimos años, la Inteligencia Artificial (IA) ha mostrado un notable potencial para asistir en el diagnóstico médico mediante la detección automatizada de patrones en datos clínicos [5] [6] [7], ofreciendo una segunda opinión objetiva y reproducible. En el ámbito cardiovascular, se han desarrollado modelos capaces de identificar alteraciones sutiles en el ECG que pueden pasar desapercibidas en una evaluación humana, mejorando la sensibilidad diagnóstica [8] [9].

No obstante, para que estos sistemas sean adoptados en entornos clínicos, deben cumplir requisitos de fiabilidad, explicabilidad y gestión de la incertidumbre, en línea con el Reglamento Europeo de Inteligencia Artificial (AI Act, UE 2024/1689) [10]. Dicho reglamento clasifica las herramientas de soporte al diagnóstico como sistemas de alto riesgo, lo que exige que sean transparentes, auditables y estén siempre bajo supervisión humana.

En este marco, la Inteligencia Artificial Explicable (XAI) y la Cuantificación de la Incertidumbre (UQ) se consolidan como pilares esenciales de una IA fiable. La primera permite interpretar cómo un modelo llega a sus predicciones [11], favoreciendo su validación clínica y aceptación por parte del personal sanitario. La segunda cuantifica el grado de confianza en las decisiones del modelo [12], elemento clave para reducir errores y garantizar la seguridad de las personas pacientes.

El presente TFG se enmarca en esta intersección entre tecnología y salud, con el propósito de desarrollar una herramienta de soporte al diagnóstico de patologías cardíacas estructurales mediante ECG, integrando técnicas de XAI y UQ. El objetivo último es reducir los tiempos de diagnóstico, aumentar la precisión clínica y disminuir la incertidumbre de las personas pacientes durante el proceso asistencial.

1.2. Objetivos del trabajo

El propósito general de este trabajo es diseñar y evaluar un sistema de soporte al diagnóstico clínico basado en Inteligencia Artificial, capaz de clasificar diversas patologías cardíacas estructurales (como valvulopatías o insuficiencias valvulares) [13] [14] [15] [16] a partir de datos multimodales procedentes del dataset EchoNext [17], que incluye señales de ECG de 12 derivaciones [18] (ver capítulo 2.3), y variables demográficas de las personas pacientes.

1.2.1. Objetivo general

Desarrollar un modelo de aprendizaje automático explicable y fiable que, a partir de las señales de ECG y los datos clínicos asociados para dar soporte a la persona a cargo del diagnóstico de la presencia y tipo de patología estructural cardíaca representada en la variable objetivo, con las siguientes categorías:

- Estenosis en la válvula aórtica.
- Insuficiencia valvular.
- Ambas patologías.
- Ninguna patología.

1.2.2. Objetivos específicos

- Implementar modelos de clasificación multiclase.
- Incorporar técnicas de Inteligencia Artificial Explicable (XAI) para ofrecer interpretabilidad local y global de las predicciones.
- Aplicar métodos de Cuantificación de Incertidumbre (UQ) para estimar la fiabilidad de las decisiones.
- Garantizar la transparencia y trazabilidad del desarrollo siguiendo las directrices de TRIPOD+AI [19] y las exigencias del AI Act europeo [10].

1.3. Enfoque y metodología usada

El trabajo adopta el marco metodológico CRISP-DM (*Cross Industry Standard Process for Data Mining*) [20], ampliamente utilizado en proyectos de ciencia de datos por su estructura iterativa, su orientación práctica y su capacidad para garantizar la trazabilidad del proceso analítico. Este modelo organiza el desarrollo en seis fases principales: comprensión del dominio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

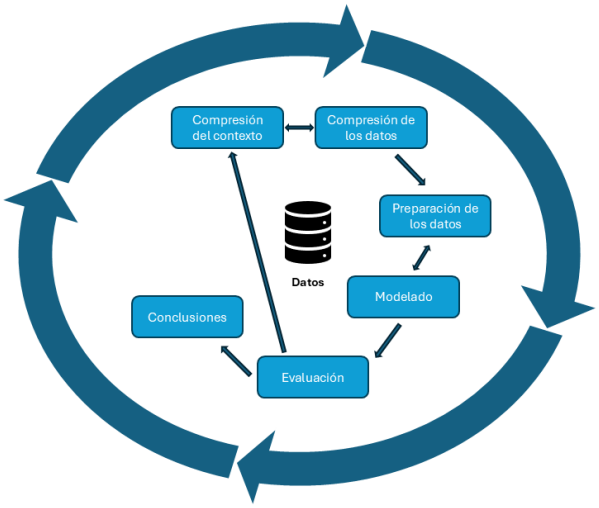

Figura 1 - Esquema del modelo CRISP-DM.[1]

La Figura 1 muestra las fases del modelo CRISP-DM, que servirán como guía para el desarrollo del presente trabajo.

En este proyecto, cada una de estas etapas se adapta al contexto de diagnóstico médico asistido por IA, estructurando el flujo de trabajo de la siguiente forma:

- **Comprensión del contexto:**

  Se realizará una revisión bibliográfica y técnica del estado del arte para comprender los fundamentos clínicos del ECG y las posibilidades de aplicación de la IA en la detección de patologías cardíacas estructurales. Esto permitirá definir los objetivos específicos y establecer los criterios de validación del modelo.

- **Comprensión de los datos**:

  Se examinará el *dataset* EchoNext [17], que contiene señales de ECG de 12 derivaciones y variables demográficas asociadas. Se identificarán la estructura de los ficheros, el tipo de variables y la distribución de clases, verificando su idoneidad para el problema de clasificación planteado.

[1] Diseño propio creado con Powerpoint.

- **Preparación de los datos**:

  Se llevará a cabo la limpieza de los datos tabulares junto con el preprocesado de las señales de los ECG, ficheros de onda representados como series temporales. De ellos se extraerán características relevantes mediante la biblioteca Neurokit2 [21] [22], generando un conjunto adecuado y discriminativo de descriptores para el modelado.

  Posteriormente, el conjunto de datos del *dataset* se dividirá en cuatro particiones: entrenamiento (**TRAIN**), calibración (**CAL**), validación (**VAL**), y prueba (**TEST**), y se eliminará la partición **NO_SPLIT**, garantizando un desarrollo riguroso y reproducible.

- **Modelado**:

  Se iniciará el proceso con un modelo de referencia (*baseline*) basado en un Árbol de Decisión, que servirá como punto de partida para la comparación posterior.

  A continuación, se entrenarán modelos de ensamble pertenecientes a dos familias principales:

  - ***Bagging***: *Random Forest* [23].
  - ***Gradient Boosting***: XGBoost [24] [25], LightGBM [26] [27], y CatBoost [28] [29].

  Se ha demostrado la eficacia de XGBoost en predicción cardíaca [30], mientras que otras investigaciones definen rangos similares para arquitecturas PSO-XGBoost [31]. Asimismo, Bentéjac, C., Csörgő, A. & Martínez-Muñoz (2021) establecen comparativas de referencia para algoritmos *gradient boost* en datos médicos tabulares [32].

  Cabe destacar que en este estudio se ha decidido excluir técnicas de aprendizaje profundo (*Deep Learning*). Si bien las redes neuronales representan el estado del arte en datos no estructurados (imágenes, texto), la literatura reciente sugiere que los modelos basados en árboles (*tree-based models*) continúan superando a las arquitecturas profundas en tareas con datos tabulares típicos [33] [34].

  Esta decisión se fundamenta en que la complejidad computacional y estructural del *Deep Learning* no garantiza necesariamente un mejor rendimiento predictivo en este dominio. Como señalan Mignan y Broccardo en *"One neuron versus deep learning in aftershock prediction"* [35], modelos más simples o tradicionales pueden rendir igual o mejor que arquitecturas profundas complejas si el volumen de datos o la naturaleza del problema no lo justifican. Asimismo, estudios comparativos indican que la superioridad de las redes neuronales sobre los árboles de decisión potenciados (*boosted trees*) suele manifestarse únicamente bajo regímenes de datos masivos o condiciones muy específicas que escapan al alcance de este conjunto de datos [36].

  La optimización de hiperparámetros se realizará mediante búsqueda sistemática, evaluando cada modelo mediante validación cruzada. El análisis se centrará tanto en la discriminación como en la calibración de las probabilidades predichas.

  - **Evaluación y fiabilidad**:

  Siguiendo las mejores prácticas para clasificación multiclase [37], la calidad del estudio se estimará observando indicadores tanto de discriminación [38] como de calibración:

  - **Métricas de discriminación:** Se utilizarán Precisión (*Precision*), Sensibilidad (*Recall)* y *F1-score* para evaluar la capacidad del modelo de distinguir entre clases. Adicionalmente, debido a su extendido uso en el ámbito clínico, se calculará el ROC AUC; no obstante, su interpretación se realizará con cautela dadas sus limitaciones informativas frente a las métricas de calibración directas [39].

  - **Métricas de calibración:** Dado que las salidas crudas de los modelos de IA son puntuaciones (*scores*) y no probabilidades reales, se evaluará la calidad de estas estimaciones mediante el *Log Loss* y el *Brier Score* [40] [41].

- **Fiabilidad e Incertidumbre (UQ):** Para dotar de fiabilidad a las predicciones y cuantificar su incertidumbre [12], se aplicará Predicción Conforme (*Conformal Prediction*) [42]. Esta técnica *post-hoc* (que no requiere reentrenar el modelo) permite transformar las puntuaciones del modelo en conjuntos de predicción con garantías estadísticas formales, siendo especialmente relevante en ciencias médicas [43]. La evaluación de esta incertidumbre se medirá a través de la cobertura (validez estadística del intervalo) y la cardinalidad (precisión o tamaño del conjunto de predicción).

- **Explicabilidad (XAI)** [11] [44] [37]:

  En la selección del modelo óptimo se tendrán en cuenta no solo métricas de rendimiento predictivo, sino también criterios de interpretabilidad y explicabilidad. Cabe destacar que la Inteligencia Artificial confiable puede extenderse al establecimiento de relaciones causa-efecto, ámbito propio de la epidemiología y del análisis causal, si bien este enfoque queda fuera del alcance del presente trabajo. Sobre dicho modelo, se aplicarán técnicas de Inteligencia Artificial Explicable tales como *Permutation Feature Importance* [45], SHAP [46] [47] [48], LIME [49] [50] y Contrafactuales [51], con el objetivo de identificar los factores diagnósticos más relevantes y generar explicaciones locales por paciente. Estas herramientas facilitarán su interpretación por parte del personal sanitario.

- **Obtención de conclusiones:**
  Finalmente, se realizará un análisis global de los resultados obtenidos durante las fases previas. Se evaluarán el rendimiento, la fiabilidad y la explicabilidad de los modelos seleccionados, contrastando los hallazgos con la literatura revisada. A partir de este estudio se elaborarán las conclusiones del TFG, incluyendo las limitaciones del enfoque, posibles mejoras y líneas futuras de trabajo, cerrando así el ciclo metodológico establecido por CRISP-DM.

Todo el proceso se implementará mediante cuadernos de Python documentados, garantizando la trazabilidad, reproducibilidad y transparencia del flujo de trabajo.

Con el fin de asegurar el rigor metodológico y la validez científica de los resultados, se seguirán las directrices de la declaración TRIPOD-AI [19], que amplía las guías de buenas prácticas para el desarrollo y reporte de modelos de predicción clínica basados en Inteligencia Artificial. En este trabajo, dichas directrices se aplicarán de la siguiente manera:

- **Transparencia de la fuente**: Detallando el origen, estructura y características del dataset empleado.
- **Trazabilidad**: Documentando los procesos de preprocesamiento, transformación y extracción de características de las señales de ECG.
- **Rigor técnico**: Justificando la división de los datos de entrenamiento, calibración y prueba, así como los criterios de evaluación y selección del modelo final.
- **Prueba de rendimiento**: Definiendo métricas de discriminación y calibración apropiadas para el ámbito clínico, e incorporando un análisis de los casos de error de categorización.
- **Explicabilidad**: Presentando los resultados de los modelos en un formato comprensible y visualmente interpretable para las personas destinatarias del trabajo (profesionales sanitarios y revisores académicos).

El enfoque sigue, por tanto, los principios de una IA fiable y responsable, en consonancia con las exigencias del Reglamento Europeo de IA (AI Act, 2024/1689), garantizando la transparencia, seguridad y auditabilidad del modelo propuesto.

1.4. Planificación del trabajo

El desarrollo del TFG sigue una planificación adaptada al calendario docente de la asignatura, organizada según las fases del modelo CRISP-DM. La Figura 2 muestra el diagrama de Gantt correspondiente.

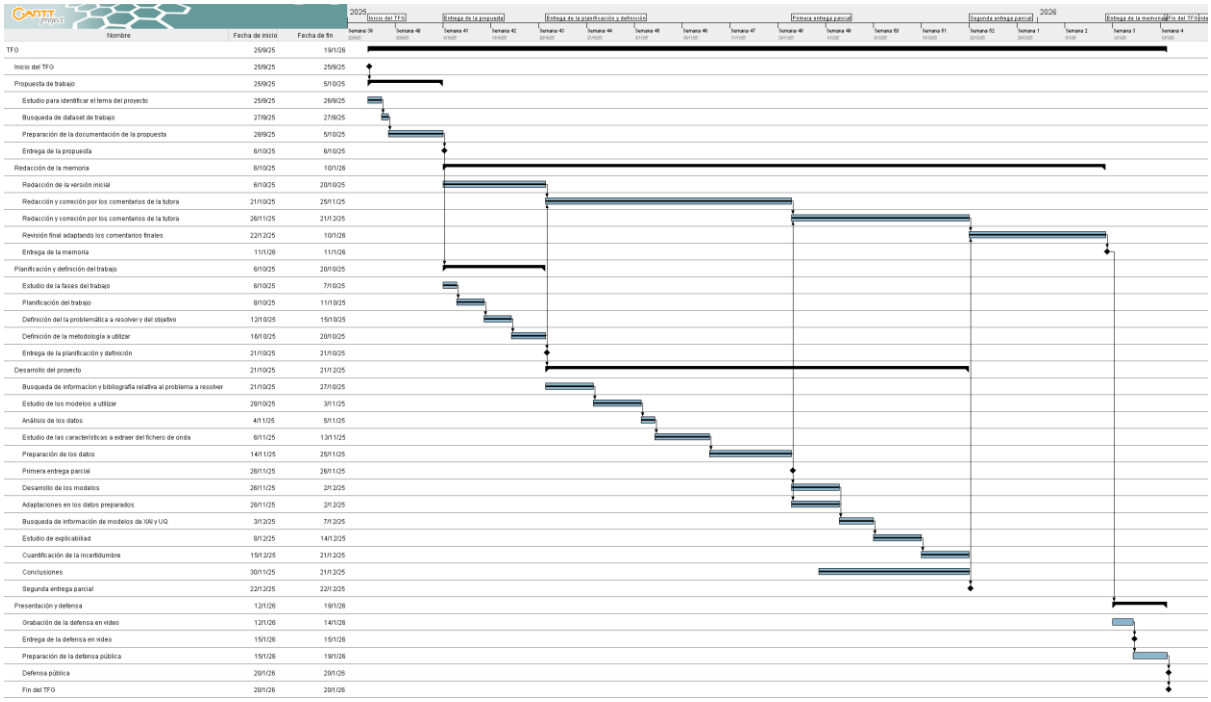

Figura 2 - Gantt de planificación del trabajo.[2]

Fases principales:

- **Propuesta del trabajo y revisión bibliográfica**:
  Definición del contexto, justificación del problema y alcance del estudio.
- **Análisis exploratorio y preparación de datos**:
  Estudio del dataset, limpieza, normalización y creación de conjuntos de entrenamiento, validación, calibración y prueba.
- **Desarrollo y evaluación de modelos**:
  Entrenamiento de los algoritmos, optimización y selección del modelo óptimo.
- **Explicabilidad e incertidumbre**:
  Aplicación de técnicas XAI y UQ al modelo final.
- **Redacción, conclusiones y defensa**:
  Documentación final de los resultados, elaboración de la memoria y preparación de la defensa pública.

Las fases de redacción y análisis se desarrollan en paralelo para mantener la coherencia entre el avance del proyecto y su documentación, asegurando que los resultados más recientes se reflejen fielmente en la memoria.

Riesgos y mitigación:

Durante el desarrollo del trabajo se han identificado una serie de riesgos potenciales que podrían afectar al cumplimiento del cronograma y a la calidad de los resultados. A continuación, se detallan los principales riesgos detectados y las medidas previstas para su mitigación:

- **Desconocimiento del tiempo de procesamiento de los datos**:
  El dataset EchoNext presenta un volumen elevado de información, lo que puede conllevar tiempos de cálculo significativos, especialmente durante las fases de extracción de características y ajuste de

[2] Diseño propio creado con GanttProject (https://www.ganttproject.biz/)

modelos. Si se requiere regenerar los modelos repetidamente para afinar su rendimiento, esto podría suponer un retraso en la planificación prevista.

Estrategias de mitigación:

- Evaluar la reducción del tamaño de la muestra o del número de características en fases iniciales para estimar la complejidad computacional.
- Utilizar Google Colab Pro u otros entornos con capacidad de procesamiento acelerado (GPU) para disminuir los tiempos de entrenamiento y optimización.
- Definir puntos de control intermedios que permitan validar resultados parciales antes de entrenamientos completos.

- **Inexperiencia con los algoritmos propuestos**

  El desarrollador no ha trabajado previamente con algunos de los algoritmos de ensamble seleccionados (*Gradient Boosting* y *Bagging*), lo que podría originar un retraso en la curva de aprendizaje y en la implementación técnica del proyecto.

  Estrategias de mitigación:

  - Dedicación de un periodo previo de estudio y experimentación con ejemplos prácticos para comprender el funcionamiento y los parámetros de los algoritmos.
  - En caso de dificultad o limitación temporal, contemplar una reducción en la complejidad del modelo final, priorizando la coherencia metodológica sobre la amplitud experimental.
  - Apoyarse en la documentación oficial y recursos de la comunidad científica (*papers*, *notebooks* públicos, foros especializados) para resolver incidencias técnicas.

1.5. Marco ético y responsabilidad social

El desarrollo de sistemas de IA en salud plantea desafíos éticos que trascienden lo puramente técnico. Este trabajo se alinea con las exigencias del AI Act europeo [10], priorizando cuatro pilares bioéticos fundamentales [52]:

- Beneficencia: El sistema busca mejorar los resultados clínicos mediante detección temprana, sin sustituir el criterio médico.
- No maleficencia: La cuantificación de incertidumbre permite abstenerse de predicciones en casos ambiguos, minimizando riesgos.
- Autonomía: La persona especialista en sanidad mantiene control total sobre las decisiones finales, usando el sistema como herramienta de apoyo.
- Justicia: La auditoría de sesgos demográficos (ver apartado 1.9) busca garantizar equidad en el acceso a diagnóstico preciso. Adicionalmente, se reconocen las limitaciones inherentes a un trabajo puramente académico.

1.6. Diversidad y sesgos demográficos en IA médica

Los algoritmos de aprendizaje automático son tan justos o tan parciales como los datos con los que se entrenan. Cuando dichos datos reflejan desequilibrios históricos, como la infrarepresentación de determinados grupos étnicos en ensayos clínicos [53], el modelo puede aprender patrones que perpetúan estas injusticias. En medicina, diversos estudios han documentado sesgos significativos según características demográficas en poblaciones minoritarias [54] [55].

En el contexto clínico, un modelo sesgado puede agravar las desigualdades existentes en el acceso al diagnóstico y tratamiento, con consecuencias directas sobre la salud de las personas pacientes. Por ello, el presente trabajo adopta un enfoque de equidad consciente [56], auditando sistemáticamente el comportamiento del modelo por subgrupos demográficos y aplicando técnicas de mitigación cuando se detecten disparidades injustificadas. Este enfoque es coherente con el marco regulatorio europeo [10], que exige transparencia y evaluación de riesgos de discriminación en sistemas de IA de alto riesgo desplegados en el ámbito sanitario.

1.7. Sostenibilidad y eficiencia computacional

El desarrollo de sistemas de Inteligencia Artificial conlleva una responsabilidad creciente respecto a su consumo energético. Este trabajo se adhiere desde su diseño a los principios de la "Green AI" [57], priorizando metodologías que maximicen la precisión diagnóstica con el mínimo coste computacional posible.

Esta perspectiva de sostenibilidad fundamenta la elección de la metodología de este estudio:

- **Eficiencia algorítmica**: Para el análisis de datos estructurados, este trabajo se basa en arquitecturas de *Bagging* y *Gradient Boosting*. La literatura científica destaca estos métodos de *Ensemble* por su alta eficiencia energética y bajo coste computacional en comparación con otras familias de algoritmos [58]. Su capacidad para procesar grandes volúmenes de datos con tiempos de entrenamiento reducidos y un uso moderado de memoria permite iterar y optimizar los modelos con una huella de carbono mínima.
- **Eficiencia del dato**: El uso de pruebas de bajo coste energético como fuente única de información refuerza el carácter sostenible de la propuesta [59]. El ECG al ser una prueba no invasiva, rápida y de muy bajo consumo energético, representa una herramienta diagnóstica eficiente que maximiza la utilidad clínica con un impacto ambiental reducido.

1.8. Breve sumario de productos obtenidos

Para el estudio y desarrollo del presente trabajo, se ha trabajado en *notebooks* Jupyter que se están distribuidos de forma pública en un repositorio de GitHub[3].

Cuaderno relativo al capítulo 3 (Análisis exploratorio de datos): TFG-EDA.ipynb.

En este cuaderno se realiza la carga y revisión estructural del dataset, junto con el análisis descriptivo de las variables demográficas, electrocardiográficas y de los predictores tabulares disponibles. Se estudian las relaciones entre variables y se verifica la coherencia entre las distintas particiones del conjunto de datos proporcionadas por los autores.

Cuaderno relativo al capítulo 4 (Preparación de datos): TFG-PreparacionDatos.ipynb.

En este cuaderno se desarrollan los procesos de limpieza y enriquecimiento del dataset, incluyendo la depuración del fichero de metadatos y la construcción de la variable objetivo consolidada (*cardiopatía*). Asimismo, se realiza la extracción de características fisiológicas complejas directamente a partir de las señales de ECG en formato de onda. El cuaderno aborda también la imputación de valores ausentes y la reducción de dimensionalidad.

Cuaderno relativo al capítulo 5 (Modelado): TFG-Modelado.ipynb.

En este cuaderno se implementa el ciclo completo de entrenamiento y validación de modelos, comenzando con un modelo base creado con árboles de decisión y continuando con modelos de *ensamble*. Se gestiona el desbalance de clases y se realiza la optimización de los hiperparámetros. La selección final del modelo se fundamenta en métricas de sensibilidad por clase, su capacidad de generalización y su idoneidad para ofrecer explicaciones interpretables y mecanismos de cuantificación de la incertidumbre.

Cuaderno relativo al capítulo 6 (XAI): TFG-XAI.ipynb.

En este cuaderno se aplican técnicas de Inteligencia Artificial Explicable para analizar el comportamiento del modelo seleccionado. Se realizan análisis de importancia global de características, así como el estudio de dependencias entre variables. Adicionalmente, se generan explicaciones locales de predicciones individuales, incluyendo verdaderos positivos y falsos negativos, y métodos basados en contrafactuales.

Cuaderno relativo al capítulo 7 (Estrategias de mitigación de sesgo): TFG-XAI_Mitigacion.ipynb.

[3] https://github.com/rchecam/TFG_UOC

En este cuaderno se evalúan posibles disparidades en el rendimiento del modelo, en particular en la Tasa de Falsos Negativos entre distintos grupos étnicos. Se implementan y comparan estrategias de mitigación de sesgo, y se reevalúa el rendimiento y la equidad del sistema tras su aplicación. El análisis se apoya en métricas desagregadas y técnicas de explicabilidad para garantizar que la corrección de sesgos no compromete la validez clínica del diagnóstico.

Cuaderno relativo al capítulo 8 (UQ): TFG-UQ.ipynb.

En este cuaderno se implementan técnicas de Cuantificación de Incertidumbre para evaluar la confianza asociada a las predicciones del modelo. Se aplican técnicas de Predicción Conforme para generar conjuntos de predicción con garantías estadísticas, y se diseña un protocolo de derivación clínica basado en el nivel de certeza del modelo.

1.9. Breve descripción de otros capítulos de la memoria

Capítulo 2: Comprensión del contexto.

Este capítulo expone los fundamentos electrofisiológicos del corazón y la morfología del electrocardiograma (ECG), describiendo cómo la actividad eléctrica cardíaca puede utilizarse como herramienta de diagnóstico indirecto en patologías estructurales. Asimismo, se detallan las principales alteraciones cardíacas que pueden reflejarse en el ECG y su relevancia para el desarrollo de modelos de Inteligencia Artificial aplicados al diagnóstico clínico.

Capítulo 3: Análisis exploratorio de datos (EDA)

Este capítulo aborda la fase de Comprensión de los datos de la metodología CRISP-DM. En él se analiza en profundidad el conjunto de datos EchoNext (v1.1.0), describiendo su estructura, las variables demográficas y electrocardiográficas, y el preprocesamiento aplicado. Asimismo, se define la variable objetivo (**Cardiopatía**), evaluando su distribución y la falta de balance en las clases existentes. Se estudian las correlaciones entre predictores, la independencia de las variables y la estrategia de particionado en los subconjuntos de entrenamiento, validación, calibración y prueba. Finalmente, se presentan las principales conclusiones del análisis exploratorio.

Capítulo 4: Preparación de datos

Este capítulo describe la transformación del dataset EchoNext en una estructura tabular adecuada para el aprendizaje automático. Se detalla la limpieza y depuración del fichero de metadatos, la exclusión de variables con riesgo de sesgo o *data leakage* y la construcción de la variable objetivo consolidada. Asimismo, se aborda la extracción de características fisiológicas a partir de las señales de ECG mediante NeuroKit2, la imputación de valores ausentes y la reducción de dimensionalidad para mitigar el sobreajuste.

Capítulo 5: Modelado

En esta fase se implementan y evalúan algoritmos de aprendizaje supervisado de ensamble (*Random Forest*, XGBoost, LightGBM y CatBoost) para la clasificación de patologías. Se expone la estrategia de optimización bayesiana de hiperparámetros con Optuna y el manejo del desequilibrio de clases. Finalmente, se justifica la selección de CatBoost como modelo final por su mejor capacidad de generalización y estabilidad clínica.

Capítulo 6: XAI

Este capítulo aplica técnicas de Inteligencia Artificial Explicable al modelo seleccionado. Se analiza la importancia global de las variables mediante PFI y SHAP, identificando la edad como factor dominante. Además, se valida el comportamiento local del modelo utilizando gráficos de dependencia (PDP/ICE), explicaciones individuales con LIME y contrafactuales, lo que permite validar decisiones clínicas concretas y detectar posibles sesgos o barreras de decisión.

Capítulo 7: Estrategias de mitigación de sesgo

Se aborda la auditoría de equidad del modelo tras detectar disparidades en la tasa de falsos negativos por grupo étnico. Se evalúan y comparan dos estrategias de mitigación: la eliminación de variables sensibles (*Blind*) y el reajuste de pesos (*Weighted*). El análisis concluye que la estrategia de ponderación ofrece mejoras controladas sin comprometer la validez clínica, a diferencia de la estrategia de ceguera que degrada severamente la precisión del sistema.

Capítulo 8: UQ

Este capítulo aborda la incorporación de mecanismos de fiabilidad y seguridad en las predicciones del modelo. Se implementa un enfoque de Predicción Conforme para generar conjuntos de predicción con garantías, proponiendo un protocolo de derivación clínica que actúa como filtro de seguridad ante casos con alta incertidumbre.

1.10. Uso ético de la IA en el presente trabajo

En la elaboración del presente Trabajo Fin de Grado se ha hecho un uso responsable y éticamente justificado de herramientas de Inteligencia Artificial generativa, siempre como apoyo al proceso de redacción y verificación, y no como sustitución del razonamiento, el análisis ni la toma de decisiones propias del autor.

Concretamente, la IA ha sido utilizada de manera puntual para la revisión lingüística y estilística del texto, con el objetivo de mejorar la claridad, coherencia y corrección formal de la redacción. Asimismo, se ha empleado como herramienta de apoyo en la corrección de errores de código, así como para el contraste conceptual de determinadas explicaciones técnicas, actuando siempre como un recurso complementario.

En todos los casos, la información proporcionada por las herramientas de IA ha sido contrastada con bibliografía académica y fuentes fiables, garantizando así la veracidad, rigor y adecuación de los contenidos incluidos en el trabajo. Ninguna información generada por IA ha sido incorporada de manera automática o acrítica.

Es importante destacar que en ningún momento se han utilizado herramientas de Inteligencia Artificial para el diseño del trabajo, la interpretación de resultados, ni la toma de decisiones metodológicas, técnicas o conceptuales. Todas las decisiones, así como el análisis y las conclusiones presentadas, son fruto exclusivo del razonamiento y criterio del autor.

De este modo, el uso de la Inteligencia Artificial en este trabajo se ha limitado a funciones de apoyo, respetando los principios de integridad académica, autoría y responsabilidad intelectual exigidos en el desarrollo de un Trabajo Fin de Grado.

2. Fundamentos del ECG y su correlación con patologías cardíacas estructurales

El presente capítulo describe los fundamentos electrofisiológicos del corazón, la obtención e interpretación del electrocardiograma (ECG) y su relación con diversas patologías cardíacas estructurales. Comprender esta base fisiológica resulta esencial para justificar el uso del ECG como fuente de datos en el desarrollo del modelo de inteligencia artificial propuesto en este trabajo.

2.1. Fundamentos electrofisiológicos del corazón y del ECG

El electrocardiograma (ECG) es el registro gráfico de la actividad eléctrica del corazón. Esta actividad se origina en el sistema de conducción especializado y se propaga por el miocardio, gobernando el ciclo cardíaco.

Este ciclo está compuesto por dos fases eléctricas fundamentales [3]:

- **Despolarización**: Activación eléctrica de las células cardíacas, que conduce a su contracción (sístole).
- **Repolarización**: Recuperación eléctrica que permite la relajación (diástole) y prepara el siguiente latido.

La señal del ECG de superficie representa la suma de estos potenciales eléctricos. Un latido normal se compone de los elementos descritos en la Figura 3 [4]:

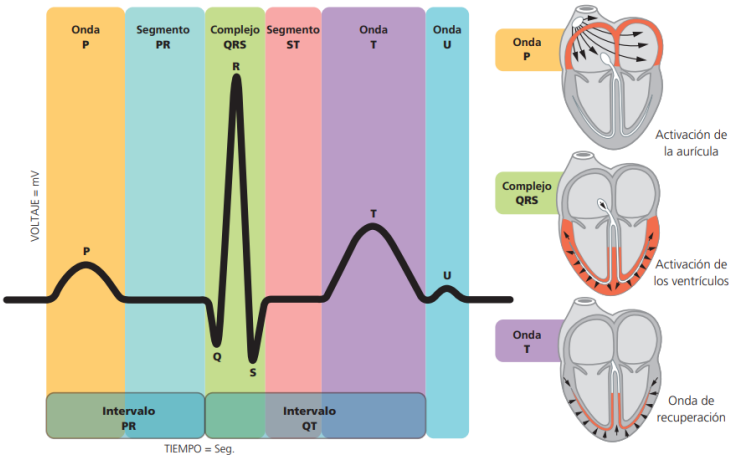

Figura 3 - Componentes de un ECG [60].[4]

- **Onda P**: Despolarización auricular.
- **Intervalo PR**: Periodo entra la despolarización auricular y la ventricular.
- **Onda Q**: Primera deflexión descendente.
- **Onda R**: Primera deflexión ascendente.
- **Onda S**: Segunda deflexión descendente en presencia de onda Q o la primera deflexión descendente cuando no se encuentra onda Q.
- **Complejo QRS**: Despolarización ventricular.
- **Segmento ST**: Despolarización completa del miocardio ventricular.
- **Onda T**: Repolarización de los ventricular.
- **Intervalo RR**: Periodo entre dos complejos QRS consecutivos.
- **Intervalo QT**: Periodo entre el comienzo de la despolarización ventricular y el final de la repolarización ventricular. El intervalo QT debe corregirse en función de la frecuencia cardíaca mediante la siguiente fórmula: $QT_c = \frac{QT}{\sqrt{RR}}$

[4] Imagen obtenida de la Fundación Española del Corazón.

- **Onda U**: Puede observarse en pacientes con hipopotasemia, hipomagnesemia o isquemia, aunque también en individuos sanos [61].

Comprender esta morfología normal es fundamental, ya que las patologías cardíacas estructurales (como la hipertrofia de las cámaras o la dilatación) alteran la masa y la geometría del miocardio, modificando así la propagación eléctrica y, por consiguiente, la forma, duración y amplitud de las ondas e intervalos.

2.2. El ECG como herramienta de diagnóstico indirecto

El ECG es la herramienta fundamental para el diagnóstico de arritmias y síndromes isquémicos agudos. Sin embargo, su uso para diagnosticar patologías estructurales (como las valvulopatías) o la función mecánica (como la función sistólica) es indirecto, ya que dichas patologías suelen confirmar con un ecocardiograma [3].

La interpretación visual del ECG para detectar estas patologías estructurales tiene una sensibilidad limitada. Un ECG puede parecer normal incluso en presencia de una enfermedad severa. No obstante, la señal del ECG contiene información mucho más sutil de la que el ojo humano puede captar.

Aquí es donde cobra valor la Inteligencia Artificial (IA). Los algoritmos de aprendizaje profundo pueden analizar patrones en los 12 canales del ECG para inferir la presencia e incluso la severidad de patologías estructurales [18]. De este modo, la IA permite detectar precozmente alteraciones que de otro modo requerirían pruebas complementarias que se suelen usar en una segunda fase dentro de los protocolos de sanitarios [62], para confirmar el diagnostico, como los ecocardiogramas.

2.3. El ECG de 12 derivaciones

Un ECG de 12 derivaciones registra la actividad eléctrica cardíaca desde 12 puntos de vista distintos, midiendo diferencias de potencial entre electrodos o entre un punto virtual y un electrodo. La disposición de estos electrodos se muestra en la Figura 4.

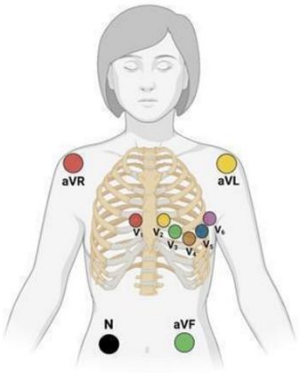

Figura 4 - Colocación de los electrodos en un ECG de 12 derivaciones estándar [63].[5]

Las derivaciones se dividen en derivaciones de las extremidades y precordiales.

2.3.1. Derivaciones de las extremidades

Se obtienen de los electrodos en las extremidades y analizan el plano frontal del corazón.

Se subdividen en:

- **Derivaciones estándar bipolares:** DI (brazo derecho-izquierdo), DII (brazo derecho-pierna izquierda) y DIII (brazo izquierdo-pierna izquierda).
- **Derivaciones monopolares aumentadas:** aVR (brazo derecho), aVL (brazo izquierdo) y aVF (pierna izquierda).

[5] Imagen obtenida de Salusplay.

2.3.2. Derivaciones precordiales

Son las derivaciones torácicas (V1 a V6). Son monopolares y miden el potencial eléctrico en el plano horizontal, permitiendo identificar lesiones en las paredes anteriores y/o laterales del corazón.

2.3.3. Análisis de las derivaciones plano frontal

- **II, III y aVF**: Cara inferior.
- **I y aVL**: Cara lateral alta.
- **aVR**: Vista superior derecha.

2.3.4. Análisis de las derivaciones plano horizontal

- **V1-V4**: Cara anterior y septal.
- **V5 y V6**: Cara lateral baja.

La Tabla 1 muestra la zona cardiaca de estudio de cada derivación.

Tabla 1 - Zona cardiaca representada por las derivaciones de un ECG.

Derivación	Zona cardiaca
V1, V2, V3 y V4	Anteroseptal
V5 y V6	Lateral baja
I y aVL	Lateral alta
II, III y aVF	Inferior
aVR	Superior derecha

La Figura 5 muestra las zonas basándose en los colores de asignación de la Tabla 1 en la impresión de un ECG típico [63] [64].

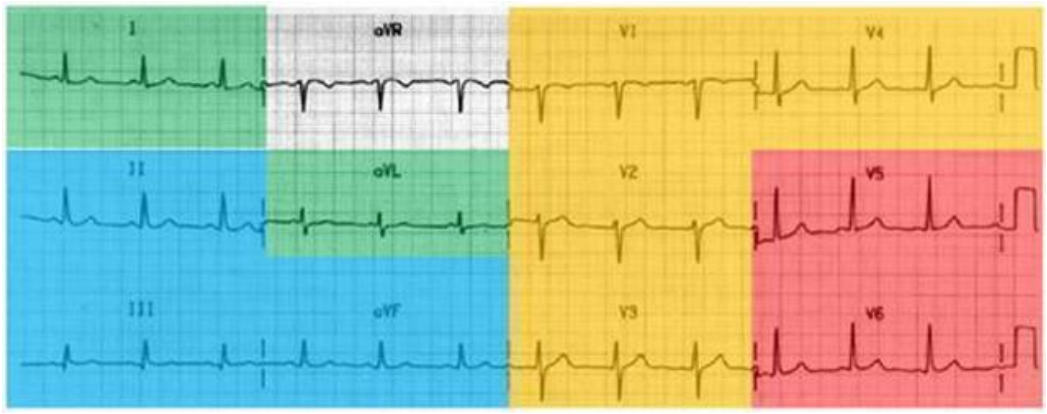

Figura 5 - Representación de zonas por derivación en un ECG [63].[6]

2.4. Patologías específicas que se pueden detectar

A continuación, se describen las patologías estructurales objetivo de este estudio y los hallazgos electrocardiográficos clásicos asociados a ellas [3]. Es importante notar que estos hallazgos pueden ser sutiles o estar ausentes, lo que justifica el uso de técnicas de Inteligencia Artificial, ya que los modelos de aprendizaje automático están diseñados para identificar patrones y correlaciones complejas en grandes conjuntos de datos [18].

[6] Imagen obtenida de Salusplay.

2.4.1. Estenosis de la válvula aórtica

Fisiopatología: La obstrucción al flujo produce una sobrecarga de presión crónica en el ventrículo izquierdo, que debe generar presiones elevadas para eyectar la sangre.

Hallazgos en el ECG:

- Patrón de sobrecarga sistólica y retraso en la activación ventricular.
- Modificación del vector inicial, con ausencia de precordiales derechas y onda Q en DI, aVL, V5 y V6.
- Desnivel del segmento ST e inversión de la onda T en derivaciones izquierdas.
- Onda P bifásica y aumento del componente negativo en V1.

2.4.2. Insuficiencia valvular

Fisiopatología: Las insuficiencias valvulares provocan sobrecarga de volumen en las cámaras cardíacas. Según la válvula afectada:

- **Insuficiencia aórtica**: Sobrecarga de volumen del ventrículo izquierdo.
- **Insuficiencia mitral**: Sobrecarga de volumen de la aurícula izquierda y del ventrículo izquierdo.
- **Insuficiencia tricúspide**: Sobrecarga de volumen de la aurícula derecha y del ventrículo derecho.
- **Insuficiencia pulmonar**: Sobrecarga de volumen del ventrículo derecho.

Hallazgos en el ECG:

- **Insuficiencia aórtica**: Ondas Q en DI, aVL, V5 y V6; complejos QRS con S profundas en V1–V2 y R altas en V5–V6; ondas T acuminadas y de base estrecha.
- **Insuficiencia mitral**: Agrandamiento de aurícula izquierda, hipertrofia ventricular izquierda, eje cardíaco normal o desviado a la izquierda y alteraciones del segmento ST–T; posible fibrilación auricular. En casos agudos, el ECG puede ser normal salvo si deriva de un infarto.
- **Insuficiencia tricúspide**: Posible fibrilación auricular, bloqueo de rama derecha y signos de crecimiento del ventrículo derecho.
- **Insuficiencia pulmonar**: Signos de sobrecarga del ventrículo derecho con patrón rSr o rsR en V1 y V3.

2.5. Conclusiones del capítulo

El ECG constituye una herramienta no invasiva y de bajo coste que refleja, de manera indirecta, alteraciones estructurales del corazón. No obstante, su interpretación manual presenta limitaciones importantes. El análisis automatizado mediante modelos de inteligencia artificial podría detectar patrones complejos en las señales eléctricas del corazón, ofreciendo una vía para el diagnóstico temprano y asistido de patologías estructurales. Eso permitiría dar herramientas a una persona encargada en el diagnóstico para clasificar si una persona paciente pudiera tener ambas patologías, solo una de ellas o si no se detecta ninguna de ellas.

3. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos (EDA) es una fase fundamental de la metodología CRISP-DM, correspondiente a la etapa de comprensión de los datos.

Su propósito es adquirir un conocimiento profundo sobre la estructura, distribución y calidad del dataset EchoNext (versión 1.1.0) [17], así como caracterizar la variable objetivo, evaluar la coherencia de las particiones y detectar potenciales problemas que puedan afectar el posterior modelado predictivo.

Los objetivos específicos de este capítulo son:

- Comprender el origen, estructura interna y relación entre los componentes del dataset EchoNext.
- Describir las variables demográficas y electrocardiográficas del fichero de metadatos e interpretar sus distribuciones.
- Explicar el tratamiento previo aplicado por los autores a los ficheros tabulares y de forma de onda.
- Definir la variable objetivo (**Cardiopatía**). Será calculada de acuerdo con las etiquetas disponibles y estará definida por los siguientes valores:
    - "Estenosis".
    - "Insuficiencia".
    - "Ambas" (en caso de padecer la persona paciente estenosis e insuficiencia).
    - "Ninguna" (en caso de ser una persona sana).
- Analizar el desbalance de clases y las implicaciones para el modelado.
- Validar estadísticamente la coherencia del particionado en TRAIN, CAL, VAL, TEST y NO_SPLIT.

3.1. Origen y estructura del dataset

El conjunto EchoNext (v1.1.0) contiene 100.000 registros multimodales distribuidos en tres componentes principales. Cada uno de los registros corresponde a los datos de una única persona.

En las matrices NumPy definidas a continuación, N representa el número de elementos (registros = personas) que tiene dicha matriz.

- **Ficheros de características tabulares** (*_tabular_features.npy): Matriz NumPy (N×7) con **predictores continuos y categóricos** preprocesados.
- **Ficheros de forma de onda** (*_waveforms.npy): Matriz NumPy (**N×1×2500×12**) que almacena los registros del ECG de 12 derivaciones en formato serie temporal.
  El fichero contiene para cada persona paciente (**N**), **un canal** de datos que corresponde al nivel de señal del ECG, **2500** medidas temporales de dichas señales, para cada una de las **12** derivaciones del ECG.
- **Fichero de metadatos** (EchoNext_metadata_100k.csv): Contiene la información **demográfica**, **variables clínicas**, **etiquetas** (*ground truth*) y asignaciones de particionado.

3.1.1. Tratamiento previo de los ficheros

Las personas responsables del dataset aplican un preprocesamiento previo que afecta tanto a las variables tabulares como a las series temporales del ECG. Este tratamiento es fundamental para comprender el comportamiento de los datos y condiciona el flujo posterior del modelado. No obstante, es necesario señalar ciertas limitaciones teóricas de este tratamiento previo realizado. En primer lugar, la eliminación de valores extremos fuera de los percentiles 0,1 y 99,9, aunque reduce ruido, conlleva el riesgo de suprimir información relevante sobre la incertidumbre del sistema o anomalías fisiológicas genuinas, críticas en la detección de anomalías. En segundo lugar, la imputación generalizada por la media introduce sesgos y elimina la incertidumbre asociada a la ausencia del dato (*missingness*), perdiendo la posible información causal de por qué

dicho dato no fue registrado. Finalmente, aunque la estandarización facilita la convergencia de ciertos algoritmos, cabe destacar que los modelos basados en árboles (objetivo principal de este TFG) son robustos ante datos no escalados.

Ficheros de características tabulares:

- **Estandarización**: Todas las variables continuas ("*age_at_ecg*", "*qrs_duration*", "*qt_corrected*", etc.) se normalizaron con la media y la desviación estándar del conjunto.
- **Imputación de valores ausentes**: Los valores faltantes se reemplazaron con la media (para variables continuas) o con 0 (para variables binarias o discretas).
- **Codificación binaria**: La variable "*sex*" fue transformada a formato binario (0="*female"*, 1="*male*"*).*
- **Corte de edad**: La variable "*age_at_ecg*" se acotó a 90 años, agrupando bajo este valor a todas las personas pacientes de edad igual o superior.

Ficheros de forma de onda:

- **Filtrado:** Se aplicó un filtro de mediana para eliminar ruido de alta frecuencia.
- **Recorte de valores extremos**: Se eliminaron los valores fuera de los percentiles 0,1 y 99,9 para reducir la influencia de artefactos.
- **Normalización**: Se normalizaron todas las señales utilizando la media y desviación estándar global del conjunto, garantizando amplitudes comparables entre derivaciones y pacientes.

Este preprocesamiento previo asegura que tanto las variables tabulares como las de forma de onda estén listas para su uso en los modelos de *machine learning*, reduciendo la necesidad de transformaciones adicionales durante la fase de modelado.

3.1.2. Relación entre los ficheros

El fichero maestro EchoNext_metadata_100k.csv actúa como eje central del dataset, conteniendo el identificador de la persona paciente ("patient_key") y la variable "split", que indica la asignación de cada registro a una de las particiones: TRAIN, VAL, TEST o NO_SPLIT. Al no definir las personas autoras del *dataset* una partición de calibración (CAL), esta se obtendrá de una partición existente (ver apartado 3.3).

Los ficheros .npy de características tabulares y de forma de onda no contienen identificadores explícitos, por lo que la correspondencia entre los tres tipos de fichero se establece por el orden de las filas (índice).

> **Ejemplo**: Una persona paciente con "*patient_key*" = 1234 ocupa la misma posición en los ficheros EchoNext_train_tabular_features.npy y EchoNext_train_waveforms.npy que en el fichero EchoNext_metadata_100k.csv.

3.1.3. Variables del fichero de metadatos

El fichero EchoNext_metadata_100k.csv actúa como la fuente principal de información demográfica y clínica. Contiene las variables de contexto de la persona paciente, las características electrocardiográficas en bruto y las etiquetas de referencia utilizadas para la definición de la variable objetivo (**Cardiopatía**).

3.1.3.1. Variables demográficas y de contexto.

Estas variables describen las características básicas de la persona paciente (ver Tabla 2).

Tabla 2 - Variables demográficas y de contexto.

Variable	Tipo de datos	Valores ausentes	Porcentaje ausentes	Valores únicos	Muestra de valores
"*patient_key*"	int64	0	0,00%	36286	[981617986, 3509735447, 4799489351, 1808045426...
"*acquisition_year*"	int64	0	0,00%	15	[2015, 2017, 2019, 2022, 2008, 2021]
"*age_at_ecg*"	int64	0	0,00%	73	[28, 49, 38, 18, 46, 64]

"race_ethnicity"	object	0	0,00%	6	["*other*", "*black*", "*white*", "*hispanic*", "*asian*", "*unknown*"]
"sex"	object	0	0,00%	2	["*male*", "*female*"]
"split"	object	0	0,00%	4	["*train*", "*val*", "*test*", "*no_split*"]

El conjunto no presenta valores ausentes en estas variables.

"Patient_key"
Identifica a una persona paciente de forma anónima.

"Adquisition_year"
Fecha de realización de la toma de la información. No se considera relevante para el análisis al no centrarse este en la evolución de las personas tras la observación.

"Age_at ecg"
Representa la edad.

La distribución de la edad (Figura 6) no es una campana de Gauss simple, ya que está fuertemente sesgada hacia pacientes de edad avanzada.

Pico de la cohorte: Se observa una alta concentración de pacientes en la década de los 60 años.

Pico por el acotado de datos: Es importante resaltar el pico artificial en los 90 años. Como se indicó en el catálogo de variables, esto se debe al capado realizado, donde la edad 90 significa 90 o más. Esto es un reto para el modelo: debe aprender que 90 no es una edad, sino una categoría que agrupa a todos las personas muy ancianas, cuyas características de ECG pueden ser distintas.

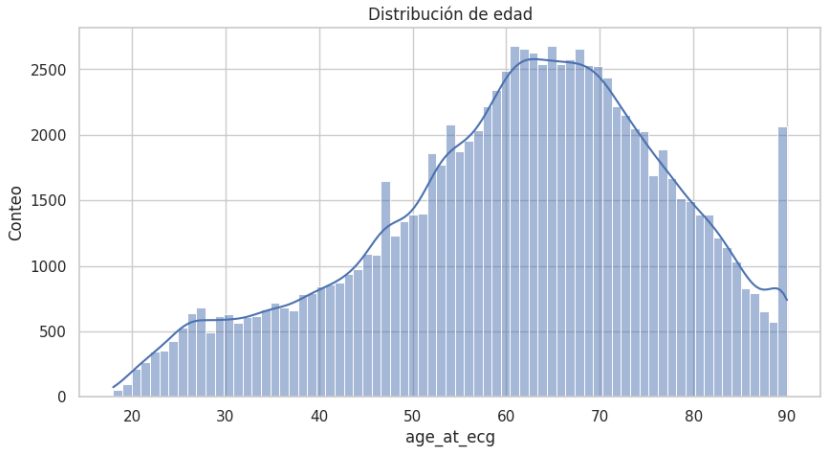

Figura 6 - Distribución de la edad.

Esto se corresponde con diversos estudios ([65] [66]) que indican que las cardiopatías valvulares son más frecuentes en personas superior a los 60 años. En este caso, la mayoría de los datos se encuentran en dicho rango.

"Race_ethnicity"
Es la variable representativa de la etnia.

La cohorte es étnicamente diversa (ver Figura 7), pero con tres particularidades importantes:

- **Grupos Mayoritarios:** El grupo más relevante es "*hispanic*" (31.013 registros, 31.01%), seguido de cerca por "*white*" (29.211 registros, 29,21%).

- **Datos Faltantes:** Es notable la alta proporción de registros "*unknown*" (12.458, 12,46%). Esto representa una cantidad significativa de datos de etnicidad faltantes que se han de asumir como una categoría en sí misma.
- **Grupos Minoritario:** El grupo formado por "*asian*" representa un valor mínimo en comparación con el resto (3,42%). Indican los autores del *dataset* que esto se debe a que los datos son proporcionados por una única institución. Esto puede afectar al modelo creando un sesgo basándose en esta variable.

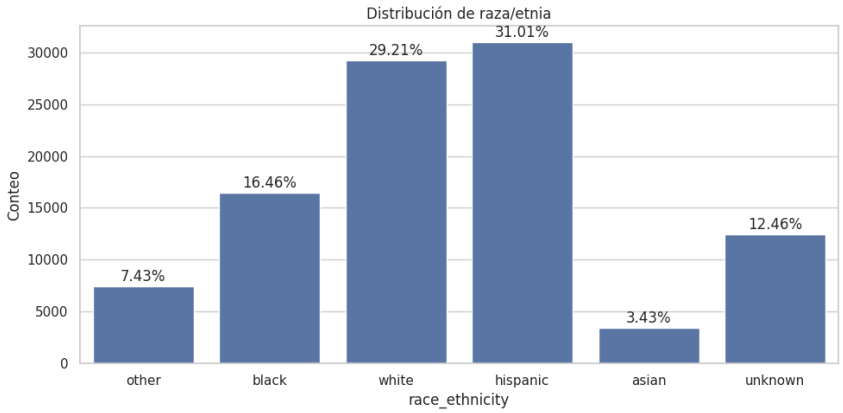

Figura 7 - Distribución por etnia.

"Sex"

Informa del sexo.

La cohorte presenta un balance de sexo razonable, con una ligera predominancia masculina, como se muestra en la Figura 8.

- **Hombres:** 53.581 registros (53,58%)
- **Mujeres:** 46.419 registros (46,42%)

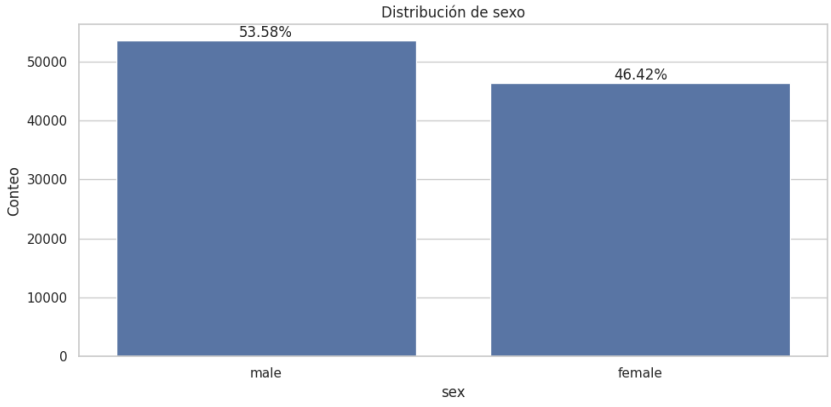

Figura 8 - Distribución por sexo.

Esta ligera desviación no se considera un desbalance severo que requiera mitigación.

"Split"

Informa del grupo al que pertenece registros ("*train*", "*val*", "*test*", "*no_split*").

3.1.3.1. Variables relativas al ECG

Las características en bruto extraídas del ECG se muestran en la Tabla 3. Estas variables representan la interpretación algorítmica de alto nivel de la morfología del ECG, resumiendo la actividad eléctrica del corazón en 5 conceptos clave.

Tabla 3 - Variables en bruto del ECG.

Variable	Tipo de datos	Valores ausentes	Porcentaje ausentes	Valores únicos	Muestra de valores
"*atrial_rate*"	float64	613	0,61%	318	[62.0, 228.0, 204.0, 112.0, 242.0, 230.0]
"*pr_interval*"	float64	10369	10,37%	205	[544.0, 224.0, 120.0, 162.0, 68.0, 215.0]
"*qrs_duration*"	float64	0	0,00%	125	[12.0, 190.0, 62.0, 24.0, 44.0, 34.0]
"*qt_corrected*"	float64	1	0,00%	481	[152.0, 707.0, 693.0, 387.0, 749.0, 519.0]
"*ventricular_rate*"	float64	0	0,00%	190	[62.0, 140.0, 30.0, 79.0, 150.0, 124.0]

"*atrial_rate*" (Frecuencia auricular)
Mide la frecuencia de las ondas P, que es el número de contracciones auriculares por minuto.

"*pr_interval*" (Intervalo PR)
Mide el tiempo (en milisegundos) desde el inicio de la onda P (activación auricular) hasta el inicio del complejo QRS (activación ventricular).

"*qrs_duration*" (Duración del QRS)
Mide el ancho (en milisegundos) del complejo QRS. Representa el tiempo total que tarda el impulso eléctrico en propagarse y despolarizar ambos ventrículos.

"*qt_corrected*" (Intervalo QT corregido)
Mide el tiempo (en milisegundos) desde el inicio del complejo QRS hasta el final de la onda T. Representa el ciclo eléctrico ventricular completo: despolarización (QRS) y repolarización (onda T). El valor se corrige (como se indicó en la Fundamentos electrofisiológicos del corazón y del ECG) para ajustarse a la frecuencia cardíaca, ya que el QT varía con la velocidad del corazón.

Las variables "*atrial_rate*" y "*pr_interval*" contienen muchos valores ausentes, pero esto no será un problema pues en los ficheros tabulares del ECG (*_tabular_features.npy) dicha información está preprocesada como se indicó al inicio de esta sección (*Origen* y estructura del dataset).

"*ventricular_rate***"** (Frecuencia Ventricular)
Mide la frecuencia de los complejos QRS, es decir, el número de latidos ventriculares por minuto. Comúnmente se denomina frecuencia cardíaca.

3.1.4. Variables de los ficheros de características tabulares

Este fichero es una matriz donde cada fila representa a una persona paciente y cada una de las 7 columnas es un predictor. Estas variables corresponden a las mismas variables existentes en el fichero de metadatos, pero como se indicó al inicio de este capítulo, ya están preprocesadas (ver Tabla 4). Serán estos valores los que usarán en los modelos.

Tabla 4 - Preproceso realizado a las variables los ficheros tabulares.

Columna	Variable	Preproceso realizado			
		Conversión a booleano	Estandarizado	Imputación de valores	Restricción de valores
1	"*sex*"	Si			
2	"*ventricular_rate*"		Si		
3	"*atrial_rate*"		Si	Con valor 0	
4	"*pr_interval*"		Si	Con valor 0	
5	"*qrs_duration*"		Si		
6	"*qt_corrected*"		Si	Con la media	
7	"*age_at_ecg*"		Si		A partir de 90 años, el valor pasa a ser 90.

La figura 9 muestra las correlaciones de los grupos de ficheros, ya que los datos vienen separados por los autores del *dataset*. No se aprecian correlaciones fuertes en ningún par, indicando independencia de todas las variables, previniendo así problemas de colinealidad que podrían degradar el rendimiento predictivo o sesgar la interpretación de los modelos. También existen variables con correlaciones medias, lo que aportará informa complementaria al modelo.

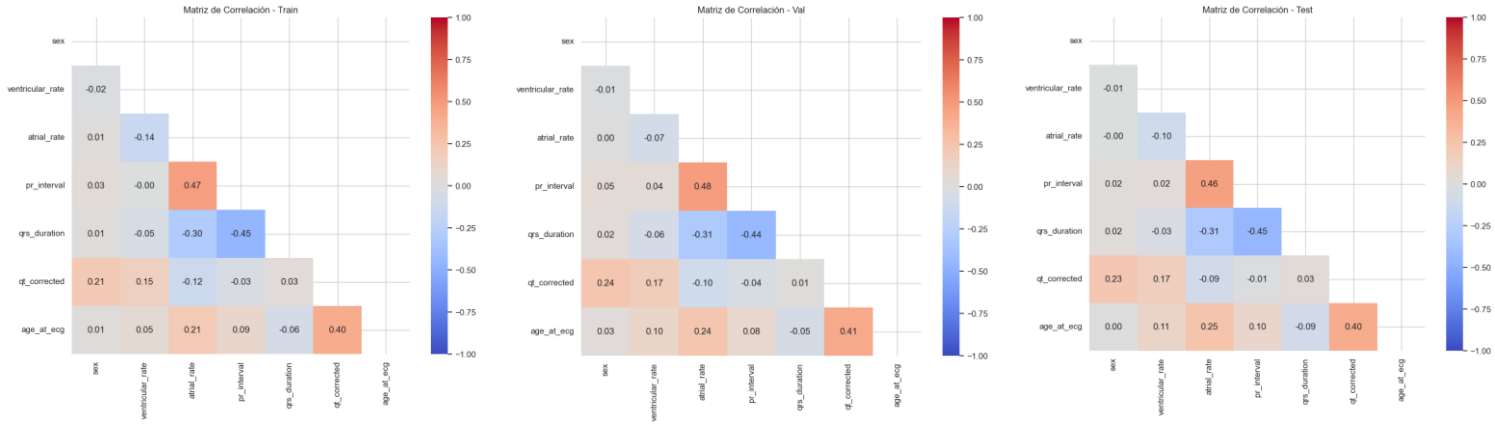

Figura 9 - Matrices de correlación

Para comparar más visualmente las diferencias, las Figura 10 muestra claramente una variación mínima entre los tres grupos de datos, siendo el valor más elevado inferior a |0,073|, por lo que no se aprecia ningún motivo que impida utilizarlos.

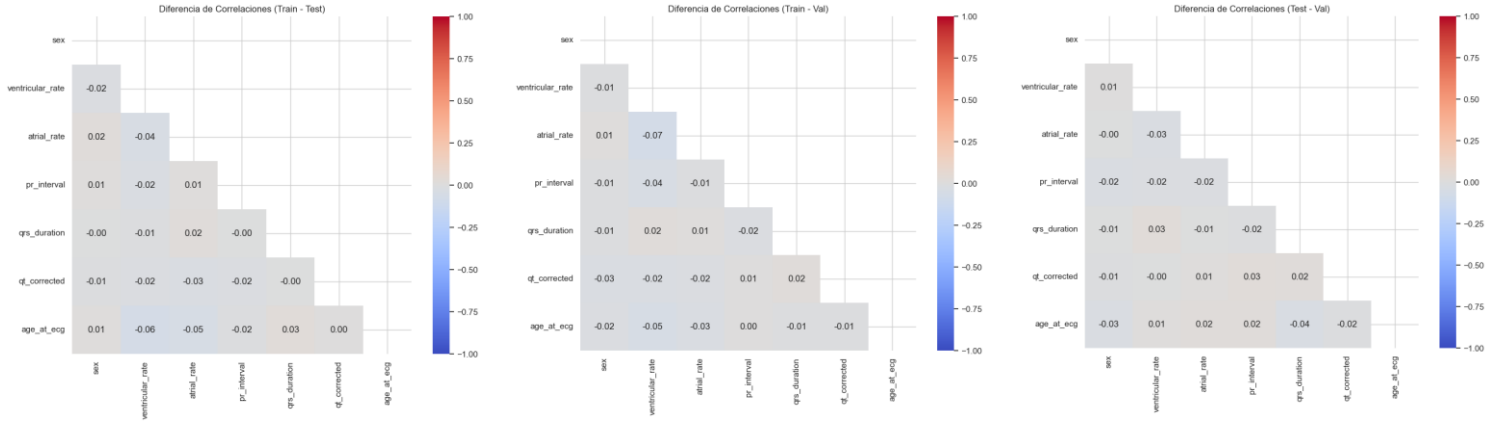

Figura 10 - Diferencias entre las correlaciones

3.1.5. Serie de datos temporales

Los ficheros de forma de onda (*_waveforms.npy) contienen los resultados de los ECG en sus 12 derivaciones. Eso permitirá hacer un estudio de ellos dando un valor añadido al estudio. Una muestra de los ECG para cada uno de los tipos de variable objetivo se muestra en la Figura 11.

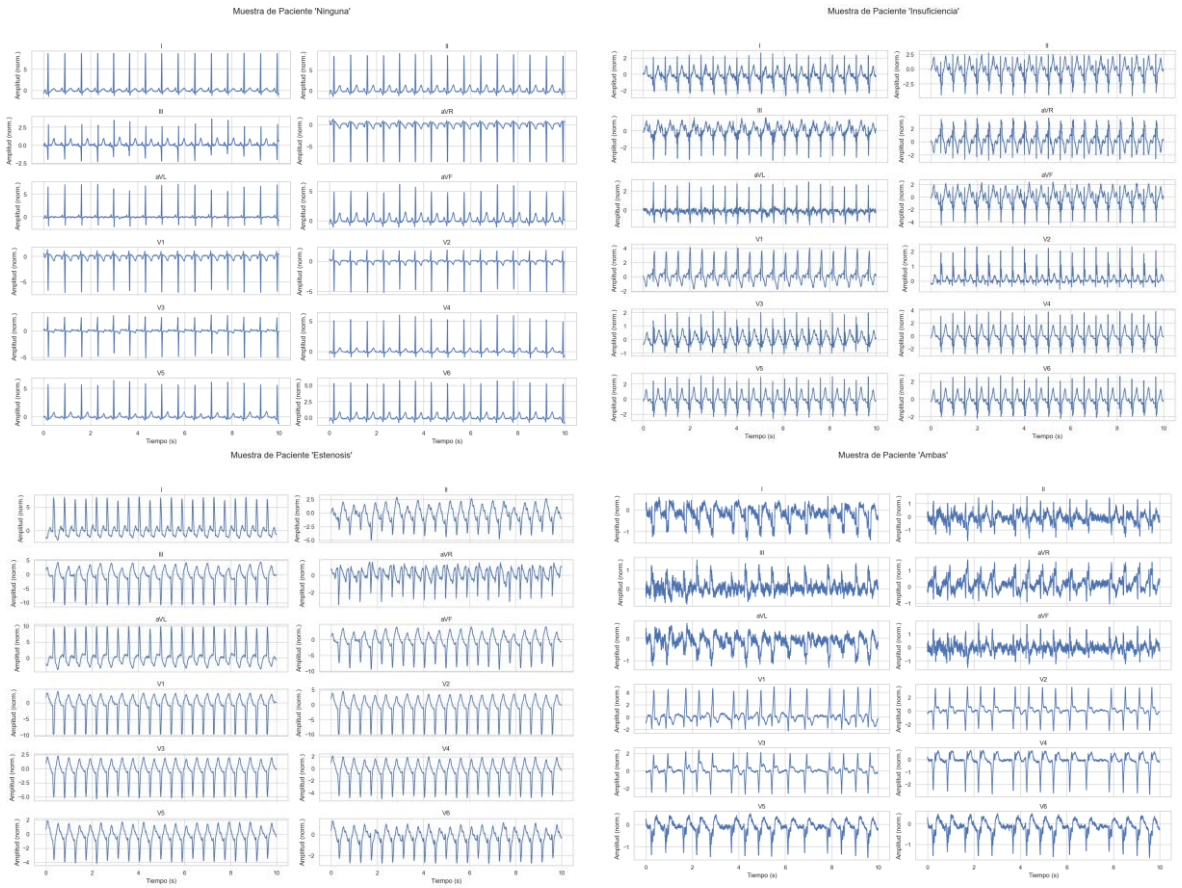

Figura 11 - ECG de 12 derivaciones del fichero de onda.

3.2. Creación y análisis de la variable objetivo

El fichero de metadatos presenta diversas variables objetivo. El propósito de este apartado es unificar dichas variables en una sola que nos permita identificar posibles cardiopatías.

3.2.1. Verificación de las variables objetivo.

De las variables objetivo disponibles se han elegido las indicadas en la Tabla 5.

Tabla 5 - Variables targets.

Variable	Tipo de datos	Valores ausentes	Porcentaje ausentes	Valores únicos	Muestra de valores
"*aortic_stenosis_value*"	object	9003	9%	5	["*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*"]
"*aortic_regurgitation_value*"	object	9003	9%	5	["*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*"]
"*mitral_regurgitation_value*"	object	9000	9%	5	["*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*"]
"*tricuspid_regurgitation_value*"	object	9036	9%	5	["*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*"]
"*pulmonary_regurgitation_value*"	object	8944	9%	5	["*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*"]

3.2.1.1. "*aortic_stenosis_value*" (Estenosis aórtica)

Índica el grado de estenosis aórtica detectado durante el ecocardiograma de confirmación.

3.2.1.2. "*aortic_regurgitation_value*" (Insuficiencia aórtica)

Índica el grado de insuficiencia aórtica detectado durante el ecocardiograma de confirmación.

3.2.1.3. "*mitral_regurgitation_value*" (Insuficiencia mitral)

Índica el grado de insuficiencia mitral detectado durante el ecocardiograma de confirmación.

3.2.1.4. "*tricuspid_regurgitation_value*" (Insuficiencia tricúspide)

Índica el grado de insuficiencia tricúspide detectado durante el ecocardiograma de confirmación.

3.2.1.5. "*pulmonary_regurgitation_value*" (Insuficiencia pulmonar)

Índica el grado de insuficiencia pulmonar detectado durante el ecocardiograma de confirmación.

Todas las variables muestran los mismos cinco posibles valores: "*none*", "*presumed none*", "*mild*", "*moderate*", "*severe*".

El valor "*none*" indica que no se ha detectado patología, y cualquier otro valor indica que no se descarta una patología.

Todas las variables muestran valores ausentes. Puesto que no se va a poder asignar un valor real a cualquier registro que no disponga de todos los valores, estos serán excluidos del estudio.

Tras realizar la limpieza quedan un total de 90861 registros útiles.

3.2.2. Creación y estudio de la variable objetivo: Cardiopatía

Para el presente análisis se usará una nueva variable objetivo basándose en las variables *target* existentes con la siguiente lógica:

- Se define **tiene_estenosis** como cierto si "*aortic_stenosis_value*" es distinto de "*none*".
- Se define **tiene_insuficiencia** como cierto si alguna de las cuatro variables de insuficiencia valvular es distinta del valor "*none*".
- La variable **cardiopatia** se asigna con la siguiente condicionalidad:
  - Si **tiene_estenosis** Y **tiene_insuficiencia** se asigna el valor "Ambas".
  - Si **tiene_estenosis** Y NO **tiene_insuficiencia** se asigna el valor "Estenosis".
  - Si NO **tiene_estenosis** Y **tiene_insuficiencia** se asigna el valor "Insuficiencia".
  - Si NO **tiene_estenosis** Y NO **tiene_insuficiencia** se asigna el valor "Ninguna".

Una vez calculados los datos, se obtiene la distribución de la Tabla 6 y se muestra en la Figura 12.

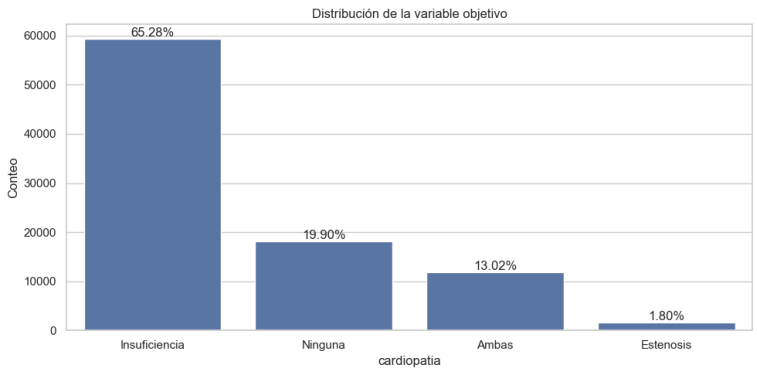

Figura 12 - Distribución de la variable objetivo.

Tabla 6 - Distribución de la variable objetivo.

Clase	Conteo	Porcentaje
"Insuficiencia"	59316	65,28%
"Ninguna"	18081	19,90%
"Ambas"	11833	13,02%
"Estenosis"	1631	1,80%

Se aprecia que las clases no están balanceadas, teniendo un peso muy elevado la insuficiencia cardiaca y un peso mínimo la estenosis. Este marcado desbalance constituye un desafío que se abordará en la fase de modelado mediante estrategias de ponderación algorítmica (*class weighting*). Se preservará la distribución original de los datos para garantizar que el sistema aprenda sobre la prevalencia real que encontrará en un entorno productivo, evitando técnicas de re-muestreo que podrían distorsionar la realidad clínica.

3.3. Particionado del dataset

3.3.1. Estrategia de particionado del dataset

Las personas a cargo del *dataset* han incluido 4 grupos en él. Una de entrenamiento, otra de prueba, una tercera de validación y una cuarta denominada "*no_split*" que indica que no está incluida en el proceso.

Para los propósitos de este TFG (que incluye Calibración de UQ), se define la siguiente estrategia:

- **Conjunto de entrenamiento** ("*train*"): Se utilizará el conjunto TRAIN original.
- **Conjunto de prueba** ("*test*"): Se utilizará el conjunto TEST original (sellado hasta la evaluación final).
- **Conjuntos de validación** ("val") **y calibración** ("cal"): El conjunto VAL original se subdividirá aleatoriamente (50/50) para crear:
    - **Conjunto de validación (*validation*)**: Se usará para el ajuste de los hiperparámetros.
    - **Conjunto de calibración (*calibration*)**: Se usará para calibrar la confianza de los modelos y evaluar la UQ.
- **Datos Excluidos:** El conjunto NO_SPLIT no se utilizará al no disponer información de los autores que indique el motivo de su exclusión en su trabajo.

3.3.2. Exploración del particionado del *dataset*

Se compararon las distribuciones de la variable objetivo entre los conjuntos (ver Figura 13).

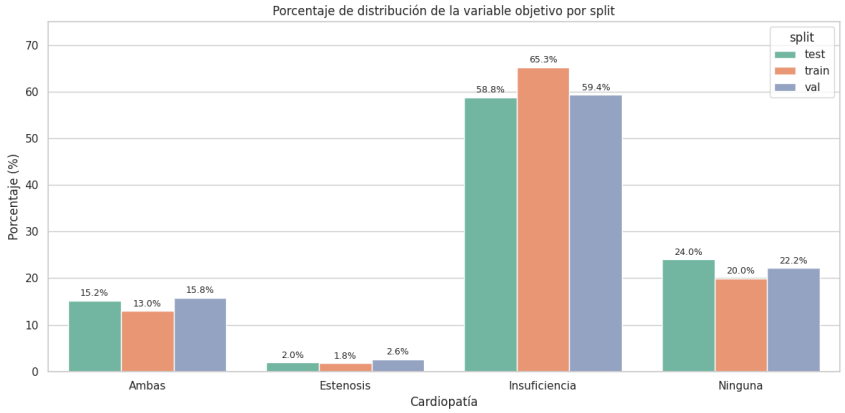

Figura 13 - Comparación de la distribución de las cardiopatías por "split".

Se observa en la Figura 13 que la distribución de clases se mantiene consistente a través de las particiones (TRAIN, VAL, TEST), respetando la proporción desbalanceada original del problema. Aunque existen pequeñas fluctuaciones porcentuales (por ejemplo, la clase "Insuficiencia" varía ligeramente entre conjuntos), la estructura general es homogénea.

Para validar estadísticamente la calidad de este particionado, se realizó una prueba de independencia Chi-cuadrado (χ^2). El objetivo es confirmar si la asignación a un grupo influye en la clase de la patología o si, por el contrario, los conjuntos son representativos.

Hipótesis:

- H_0: Cardiopatía es independiente del grupo.
- H_1: Cardiopatía es dependiente del grupo.

Resultados de la prueba:

- $\chi^2 = 143{,}04$
- $p - valor \approx 0$

Dado el elevado tamaño de la muestra (100.000 registros), la prueba es extremadamente sensible a pequeñas variaciones, resultando en un p-valor que rechaza la hipótesis nula de igualdad matemática estricta. Sin embargo, desde una perspectiva práctica de *Machine Learning*, la similitud visual de las distribuciones y la aleatoriedad del proceso de particionado permiten asumir que los datos cumplen con la propiedad IID (Independientes e Idénticamente Distribuidos).

Esto implica que las particiones son intercambiables y que el modelo entrenado en TRAIN se enfrentará en TEST a un escenario probabilístico equivalente, requisito fundamental para garantizar que las métricas de evaluación sean fiables y no fruto de un sesgo de selección en el particionado.

3.4. Conclusiones del capítulo

Tras el análisis exploratorio de datos se concluye:

- El *dataset* está limpio en sus *features*, pero existen valores ausentes en las variables que se usan para calcular la variable objetivo, por lo que se han de eliminar ya que no se obtendrían variables objetivo válidas para el estudio.
- Las variables predictoras que se van a utilizar no tienen colinealidad, por lo que no existirá redundancia al usarlas.
- La variable objetivo tiene una descompensación elevada entre las clases de esta. Existe una clase dominante ("Insuficiencia"), así como una marginal ("Estenosis"). Para gestionar este desequilibrio sin introducir artefactos, se descarta el uso de técnicas de generación de datos sintéticos (como SMOTE), las cuales podrían inflar artificialmente las métricas de rendimiento al alterar las correlaciones naturales entre clases. En su lugar, la estrategia de mitigación se centrará estrictamente en la ponderación de la función de pérdida (*loss weighting*), penalizando más severamente los errores de clasificación en las clases minoritarias durante el entrenamiento.
- El análisis estadístico y visual confirma que la estrategia de división de datos ha sido correcta. Las proporciones de las clases se mantienen estables entre los conjuntos de entrenamiento, validación y prueba, descartando sesgos de selección que pudieran invalidar la evaluación posterior del modelo.

4. Preparación de datos

La preparación de los datos (*Data Preparation*) constituye la tercera fase de la metodología CRISP-DM y representa el puente crítico entre el entendimiento de la información y el modelado efectivo. Su propósito es transformar el conjunto de datos en bruto del *dataset* EchoNext en una estructura tabular, limpia y enriquecida, apta para el entrenamiento de modelos de aprendizaje automático supervisado.

En este capítulo se detallan los procesos de limpieza, la ingeniería de características sobre las señales biomédicas y las estrategias de reducción de dimensionalidad aplicadas para mitigar el riesgo de sobreajuste.

Los objetivos específicos de este capítulo son:

- Filtrar y depurar el conjunto de metadatos, eliminando observaciones no válidas y variables que puedan introducir sesgos o fugas de información (*data leakage*).
- Construir y consolidar la variable objetivo (**Cardiopatía)** a partir de las etiquetas ecocardiográficas, generando los vectores de salida para el modelado.
- Extraer características fisiológicas interpretables directamente de las señales de onda del ECG utilizando la librería NeuroKit2 y técnicas de descomposición tiempo-frecuencia.
- Imputar valores ausentes mediante algoritmos de vecindad (KNN) para preservar la estructura local de los datos sin introducir distorsiones estadísticas.
- Reducir la dimensionalidad del conjunto de datos aplicando un *pipeline* secuencial de selección de características basado en criterios estadísticos y de colinealidad.

El procesamiento se ha realizado empleando Python con las bibliotecas, entre otras, Pandas y NumPy para la manipulación de datos, **NeuroKit2** para el procesamiento de bioseñales y Scikit-Learn para las tareas de imputación y selección de características.

4.1. Tratamiento del fichero de Metadatos y definición de variables

El fichero de metadatos original contiene la información demográfica y las etiquetas clínicas necesarias para el aprendizaje supervisado. El procesamiento de este fichero se ha centrado en la limpieza de registros no útiles, la exclusión de variables que podrían introducir sesgos y la construcción de la variable objetivo.

4.1.1. Selección de variables y exclusiones

Se realizó un filtrado inicial de las observaciones, descartando aquellas pertenecientes al grupo de datos denominado NO_SPLIT, reduciendo el análisis a los subconjuntos de TRAIN, VAL y TEST definidos por los autores del dataset. Esta decisión se tomó por la falta de documentación existente sobre el contenido del grupo NO_SPLIT.

De los subconjuntos de análisis, se seleccionaron las variables demográficas para el estudio:

- “age_at_ecg”
- “race_ethnicity”
- “sex”

Para asegurar la validez del modelo como herramienta de cribado basada en ECG, se excluyeron deliberadamente dos grupos de variables del conjunto de predictores:

- **Variables administrativas:** Identificadores como “*patient_key*”, fechas de adquisición (“*acquisition_year*”) y localización (“*location_setting*”) fueron eliminados al no aportar información fisiopatológica relevante y presentar riesgo de sesgo por lote.

- **Variables derivadas de ecocardiografía:** Se descartaron todas las mediciones provenientes de ultrasonidos definidas en la documentación del *dataset* como *Echo-Derived Binary Labels* y *Echo-Derived Features*, excepto las siguientes, ya que fueron usadas para construir la variable objetivo:
    - "*aortic_stenosis_value*".
    - "*aortic_regurgitation_value*".
    - "*mitral_regurgitation_value*".
    - "*tricuspid_regurgitation_value*".
    - "*pulmonary_regurgitation_value*".

El objetivo del TFG es evaluar la capacidad diagnóstica del ECG de forma aislada, por lo que incluir datos de la ecografía (la prueba de validación o *Gold Standard*) implicaría usar la confirmación para predecir, invalidando la utilidad del modelo como herramienta de cribado. Estas variables ecográficas auxiliares fueron eliminadas del dataset inmediatamente después de la construcción de la variable objetivo.

4.1.2. Redefinición de particiones y creación del conjunto de Calibración

Aunque la estructura original del dataset contemplaba cuatro particiones (TRAIN, VAL, TEST, NO_SPLIT), la metodología propuesta en este trabajo requiere un conjunto de datos independiente para la fase de Cuantificación de la Incertidumbre (UQ). Utilizar el mismo conjunto para ajustar hiperparámetros y para calibrar probabilidades introduciría un sesgo optimista en las estimaciones de confianza.

Al descartar la partición NO_SPLIT, se procedió a dividir el conjunto de validación original (VAL) en dos subconjuntos disjuntos del 50%, manteniendo la estratificación de la variable objetivo para preservar la distribución de clases:

- **Conjunto de Validación (*VAL*):** Destinado a la selección de modelos y ajuste de hiperparámetros.
- **Conjunto de Calibración (*CAL*):** Reservado para el ajuste de los métodos de calibración probabilística y Predicción Conforme. Se usará *Split Conformal Prediction* (originalmente *Inductive Conformal Prediction*) [67] y RAPS [68].

4.1.3. Construcción de la variable objetivo

Aunque el análisis exploratorio permitió visualizar la distribución de las patologías, es en esta fase donde se materializa la construcción de la variable objetivo "**cardiopatia**" (sin acento) que alimentará a los modelos supervisados.

Se implementó una lógica de consolidación basada en las etiquetas originales:

1. **Estenosis:** Se considera presente si "*aortic_stenosis_value*" indica cualquier grado de patología distinta a "*none*".
2. **Insuficiencia:** Se considera presente si alguna de las cuatro variables de regurgitación valvular indica cualquier grado de patología distinta a "*none*".
3. **Clasificación final:** Se asignó la etiqueta "Ambas" si coexistían las dos condiciones anteriores, "Estenosis" o "Insuficiencia" si aparecían de forma aislada, y "Ninguna" ante la ausencia de ambas.

Durante este proceso, se eliminaron los registros donde la información era insuficiente para determinar una categoría con certeza (etiquetados como "No se puede determinar"), resultando en un conjunto final de 74.799 observaciones válidas. Finalmente, se generaron los ficheros de etiquetas definitivos (y_train, y_val, y_cal, y_test) para su uso en la fase de modelado.

La variable objetivo fue transformada mediante *Label Encoding*. Esta técnica asigna un número entero único a cada categoría, permitiendo su procesamiento por algoritmos que requieren entradas numéricas estrictas sin aumentar la dimensionalidad del *dataset.*

4.1.4. Codificación de variables categóricas

Las variables cualitativas restantes, el sexo y la etnia, fueron transformadas mediante *Label Encoding*, tal y como se hizo con la variable objetivo.

4.2. Tratamiento de los ficheros de ECG y extracción de características

El dataset original proporciona características tabulares preprocesadas. Sin embargo, para garantizar la explicabilidad (XAI) del modelo y asegurar que cada variable tenga una interpretación fisiológica trazable, se ha optado por descartar estos ficheros y extraer las características directamente desde las señales de onda.

Debido a que el uso de las 12 derivaciones completas aumentaría drásticamente la dimensionalidad y el coste computacional, se realizó una selección estratégica de 4 derivaciones representativas basándonos en la literatura clínica:

- **II**: Mejor detección de R-peaks, morfología P más clara [69].
- **V1**: Mejor onda P en arritmias supraventriculares; QRS inicial [70].
- **V5**: Lateral izquierda - criterios de HVI (Cornell) [71].
- **aVL**: Detecta alteraciones sutiles en cara lateral alta [3].

Además de la información de las derivaciones, es recomendable incluir métricas globales como información complementaria [71] [72]

4.2.1. Extracción mediante NeuroKit2 y Transformada Wavelet

Para la extracción de características se ha utilizado **NeuroKit2**, una biblioteca especializada en el procesamiento de señales neurofisiológicas que ofrece algoritmos avanzados para la detección de eventos cardíacos.

Para la delineación de la señal (identificación de los puntos de inicio, pico y fin de las ondas P, QRS y T), se ha optado por el método de **Transformada Wavelet Discreta (DWT)** (*method*='dwt'). A diferencia de los métodos clásicos basados en umbrales o Fourier, que pierden la información temporal, las *wavelets* permiten analizar la señal simultáneamente en el dominio del tiempo y la frecuencia. Esta propiedad es crucial en el análisis del ECG, ya que permite distinguir con precisión componentes de alta frecuencia (como el complejo QRS) de componentes de baja frecuencia (como las ondas P y T) con la resolución temporal adecuada para cada uno.

La Tabla 7 muestra las características extraídas. Un total de 59 (14 por cada una de las 4 derivaciones y 3 métricas globales de señal).

Tabla 7 - Variables obtenidas del fichero de onda.

Grupo de información	USO	Variable por derivación
Pico R	Detección robusta [69]	*"R_Peak_MeanAmplitude"*
		"R_Peak_MaxAmplitude"
Onda P / Onda T	Amplitudes y duraciones [70]	*"P_Amplitude"*
		"P_Duration"
		"T_Amplitude"
		"T_Duration"
		"PR_Interval"
Complejo QRS	Información base [3]	*"Q_Amplitud"*
		"Atrial_rate"
		"Ventricular_rate"
		"R_Amplitude"
		"S_Amplitude"
		"QT_Corrected"
		"QRS_duration"
		Variable global

Características de señal globales	Información complementaria [71] [72]	*"Signal_Mean"*
		"Signal_Std"
		"Signal_Range"

4.3. Preparación de los datos para los modelos

Una vez unificadas las características extraídas con los metadatos, se procedió a la limpieza final y reducción de datos sobre el conjunto completo de TRAIN, aplicándose posteriormente las transformaciones realizadas sobre este conjunto a los conjuntos de validación, calibración y prueba.

4.3.1. Tratamiento de Valores Ausentes mediante K-Nearest Neighbors (KNN)

Para el tratamiento de valores ausentes derivados de procesos de extracción de características (como ondas P no detectables en señales ECG), se implementó una metodología basada en el algoritmo **K-Nearest Neighbors** (**KNN**). Este enfoque estima cada valor faltante a partir de las personas pacientes más similares en el espacio multivariante, preservando las relaciones locales entre observaciones de manera más efectiva que métodos simples como la imputación por media o mediana, los cuales pueden introducir sesgos y aplanar la variabilidad natural de los datos.

Preprocesamiento y estandarización: Previo a la imputación, todas las variables numéricas fueron estandarizadas mediante transformación z-score. Esta normalización es un requisito técnico indispensable para el cálculo correcto de distancias euclidianas en el algoritmo K-NN. Aunque los modelos finales de *ensamble* (basados en árboles) no requieren estrictamente datos estandarizados al no basarse en métricas de distancia, este paso intermedio es necesario para garantizar la calidad de la imputación previa.

Optimización multimétrica del hiperparámetro k: La elección del número de vecinos (k) es crítica: valores demasiado bajos (k=1) hacen la imputación sensible a ruido y valores atípicos, mientras que valores excesivos ($k \geq 15$) pueden sobre-suavizar las estimaciones, diluyendo patrones clínicamente relevantes y distorsionando la estructura local de los datos. Para determinar el k óptimo de manera objetiva, se diseñó un procedimiento sistemático que evalúa simultáneamente tres dimensiones esenciales para datos médicos:

- **Precisión de imputación**: Error cuadrático medio (RMSE) entre valores imputados y valores reales conocidos.
- **Preservación de la distribución**: Estadístico de Kolmogorov-Smirnov (KS) comparando la distribución estadística original con la de los valores imputados.
- **Conservación de relaciones**: Porcentaje de preservación de la matriz de correlaciones multivariantes originales.

Sobre el conjunto de entrenamiento, se generaron artificialmente patrones de ausencia en un 10% de los datos conocidos, evaluando un rango de valores $k \in \{1, 3, 5, 7, 10, 15, 20\}$ mediante validación cruzada de 5 particiones. Las tres métricas fueron normalizadas y combinadas en una puntuación global ponderada (RMSE: 40%, KS: 30%, Correlaciones: 30%), reflejando la prioridad de obtener valores precisos sin comprometer la integridad estructural de los datos.

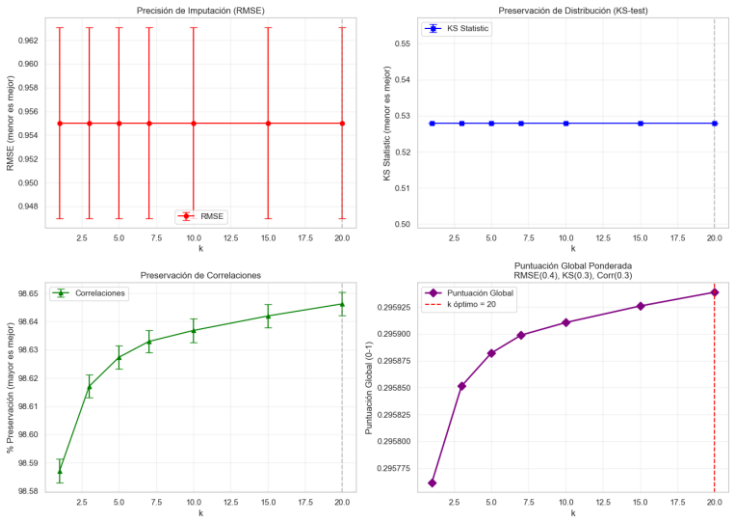

Figura 14 – Resultado del estudio para obtener un k óptimo.

La figura 14 muestra la evolución de las tres métricas de evaluación (RMSE, estadístico KS y preservación de correlaciones) junto con la puntuación global combinada, para el rango de valores de k evaluado. La línea vertical roja indica el valor de k seleccionado como óptimo.

El valor que maximizó la puntuación integral y fue seleccionado como óptimo para este estudio fue **k=20**. Los resultados completos de la evaluación se presentan en la Tabla 8.

Tabla 8 - Resultado del estudio para obtener un k óptimo.

k	RMSE	KS_Statistic	Corr_Preservation_%	Global_Score
1	0,9550 ± 0,0080	0,5280 ± 0,0000	98,5871% ± 0,0042%	0,2958
3	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6171% ± 0,0040%	0,2959
5	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6274% ± 0,0042%	0,2959
7	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6329% ± 0,0040%	0,2959
10	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6368% ± 0,0042%	0,2959
15	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6420% ± 0,0041%	0,2959
20	0,9550 ± 0,0080	0,5280 ± 0,0000	98,6463% ± 0,0041%	0,2959

Procedimiento final y prevención de *data leakage*: Con el hiperparámetro k óptimo determinado, se ajustó el imputador **KNN** final sobre el conjunto de entrenamiento completo. Para garantizar la validez de la evaluación, todo el proceso de optimización y ajuste de parámetros (escalador e imputador) se realizó exclusivamente sobre el conjunto de entrenamiento. Los conjuntos de VAL, CAL y TEST fueron transformados posteriormente utilizando únicamente los parámetros ya aprendidos, asegurando así la independencia de los datos y previniendo cualquier forma de *data leakage*.

4.3.2. Reducción de la dimensionalidad y selección de características

A las 62 variables iniciales, se aplicó un proceso secuencial de reducción para eliminar redundancia y ruido.

- **Análisis de valores atípicos (Outliers)**

Este análisis busca identificar observaciones anómalas que se desvían de la distribución estadística normal (detectadas mediante Rango Intercuartílico). Aunque se detectaron variables con un porcentaje significativo de *outliers* (p. ej., "*V1_Q_Amplitude*" con 11,3%), se decidió conservar todos los registros. En el contexto biomédico, los valores extremos en el ECG suelen reflejar condiciones fisiológicas reales (arritmias, hipertrofias) y no errores de medición, por lo que su eliminación supondría una pérdida de información crítica [73].

- **Umbral de varianza (*Variance Threshold*)**

Esta técnica elimina características que son constantes o casi constantes entre las observaciones, ya que no aportan capacidad discriminativa al modelo. Utilizando un umbral conservador de 1e-4, se confirmó que todas las variables presentaban suficiente variabilidad, por lo que ninguna fue descartada en esta fase.

- **Eliminación por colinealidad (*Correlation Removal*)**

La multicolinealidad, que ocurre cuando dos variables están altamente correlacionadas, introduce redundancia e inestabilidad en los modelos, dificultando la interpretación de la importancia de las variables. Se calculó la matriz de correlación de Pearson para el conjunto de entrenamiento.

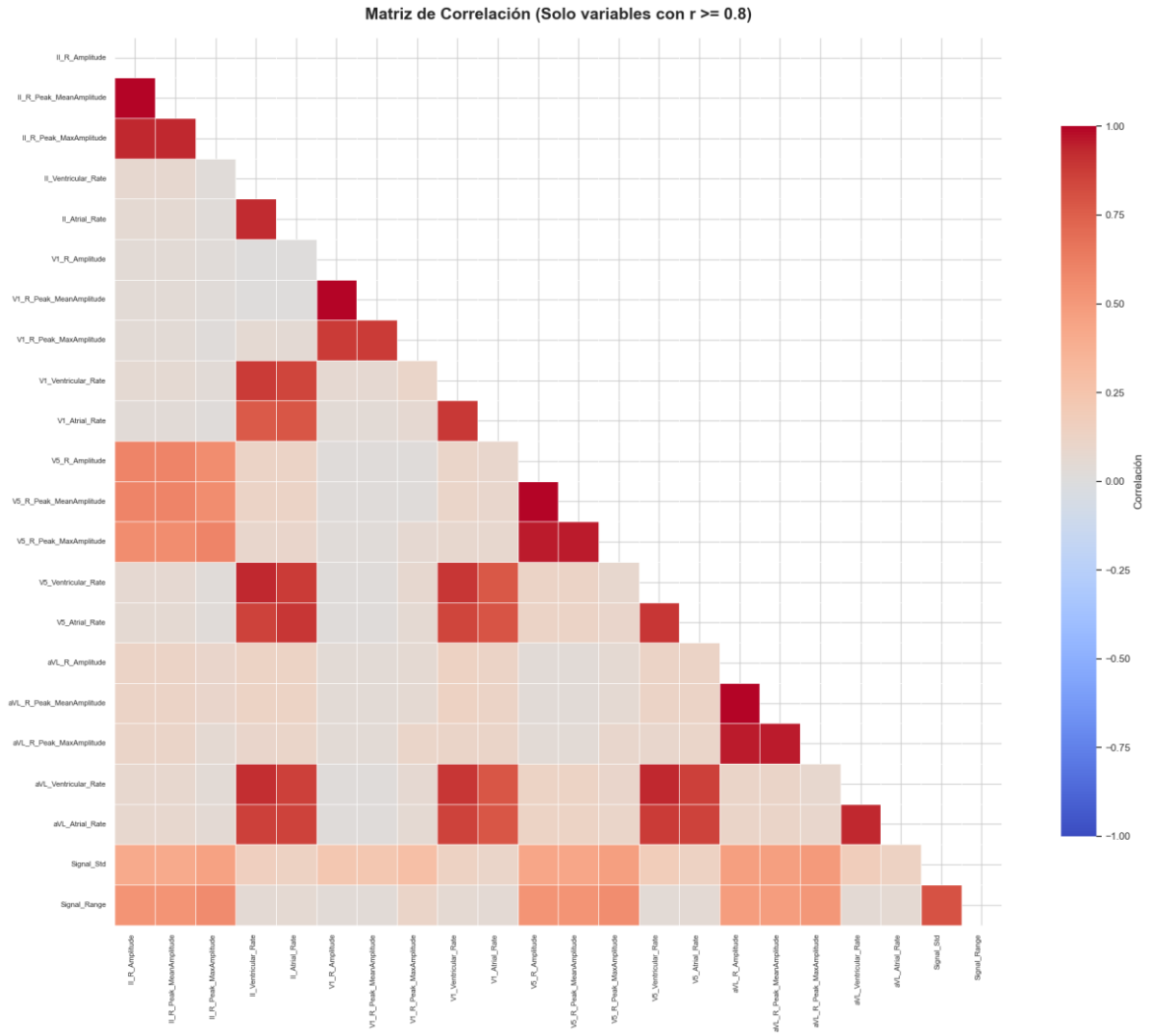

Figura 15 - Matriz de correlación de TRAIN antes de la reducción.

En la figura 15 se observan clústeres de alta correlación que indican redundancia.

Se estableció un umbral de correlación de |0,85| para la eliminación de variables redundantes. Este valor corresponde a un Factor de Inflación de Varianza (**VIF**) teórico de 3,6 veces, por debajo del límite conservador de **VIF** > 5 recomendado en modelado estadístico [74], garantizando estabilidad numérica sin eliminación excesiva. Para cada par de variables que superaba este límite, se eliminó sistemáticamente aquella con mayor correlación promedio con el resto del dataset, priorizando así la retención de variables con más información única. Este proceso resultó en la eliminación de 14 variables altamente redundantes, principalmente medidas duplicadas de frecuencia cardíaca y amplitudes repetitivas, reduciendo la dimensionalidad mientras se preservaba la integridad clínica de los predictores restantes.

- **Selección de características (ANOVA F-test)**

Finalmente, se aplicó un análisis de varianza (ANOVA) univariante para evaluar la capacidad de cada característica individual para discriminar entre las clases de la variable objetivo. Esta prueba calcula un estadístico F y un p-valor asociado, permitiendo filtrar variables irrelevantes.

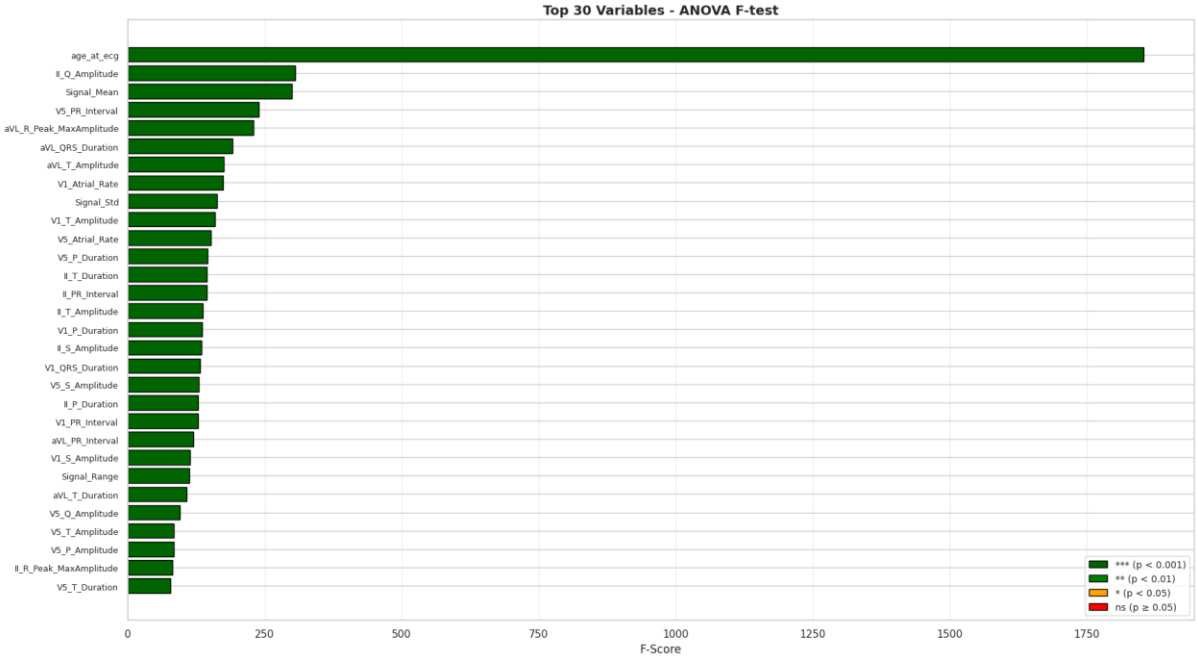

Figura 16 - Variables más significativas para la clasificación según ANOVA F-test.

La figura 16 muestra que todas las variables seleccionadas tienen una alta relevancia estadística.

Como resultado final, se conservaron únicamente las variables con significancia estadística *p-value* < 0,05. El análisis confirmó que las 48 variables restantes eran estadísticamente significativas para la predicción, por lo que se mantuvieron todas para el modelado final.

4.3.3. Resultado final

La evolución de los registros en base a las fases del preprocesado queda mostrada en la Tabla 9, manteniendo un 74,80% del *dataset* original.

Tabla 9 - Evolución de los registros de las particiones del dataset.

Partición	*Dataset* original	Descarte del grupo NO_SPLIT	División de VAL original	Eliminación de registros sin cardiopatía válida	Registros finales
TRAIN	72.475	72.475	72.475	65.878	**65.878**
VAL	4.626	4.626	2.313	2.053	**2.053**
CAL	No existe	No existe	2.313	2.053	**2.053**
TEST	5.442	5.442	5.442	4.815	**4.815**
NO_SPLIT	17.457	No existe	No existe	No existe	**No existe**
TOTAL	100.000	82.543	82.543	74.799	**74.799**

Tras el proceso de preparación, se redujo la dimensionalidad de 62 a 48 variables (**reducción del 22,6%**). El *dataset* final quedó estructurado en las cuatro particiones definidas, garantizando la robustez del modelado y la validación de la incertidumbre.

Las variables finalmente seleccionadas se muestran en la Tabla 10 y su composición en la Figura 17.

Tabla 10 - Variables seleccionadas para el estudio.

Grupo de características	Variable
Demográficas	"age_at_ecg"
	"sex"
	"race_ethnicity"
Derivación II	"II_T_Amplitude"
	"II_PR_Interval"
	"II_QT_Corrected"
	"II_R_Peak_MaxAmplitude"
	"II_Q_Amplitude"
	"II_P_Amplitude"
	"II_T_Duration"
	"II_S_Amplitude"
	"II_P_Duration"
	"II_QRS_Duration"
Derivación V1	"V1_P_Duration"
	"V1_R_Peak_MaxAmplitude"
	"V1_QT_Corrected"
	"V1_S_Amplitude"
	"V1_P_Amplitude"
	"V1_Q_Amplitude"
	"V1_QRS_Duration"
	"V1_PR_Interval"
	"V1_T_Duration"
	"V1_Atrial_Rate"
	"V1_T_Amplitude"
Derivación V5	V5_PR_Interval"
	V5_QRS_Duration"
	V5_R_Peak_MaxAmplitude"
	V5_P_Amplitude"
	V5_S_Amplitude"
	"V5_P_Duration"
	"V5_QT_Corrected"
	"V5_T_Amplitude"
	"V5_Q_Amplitude"
	"V5_Atrial_Rate"
	"V5_T_Duration"
Derivación aVL	"aVL_T_Amplitude"
	"aVL_S_Amplitude"
	"aVL_Q_Amplitude"
	"aVL_R_Peak_MaxAmplitude"
	"aVL_PR_Interval"
	"aVL_QRS_Duration"
	"aVL_T_Duration"
	"aVL_P_Duration"
	"aVL_QT_Corrected"
	"aVL_P_Amplitude"
Características de señal globales	*"Signal_Mean"*
	"Signal_Std"
	"Signal_Range"

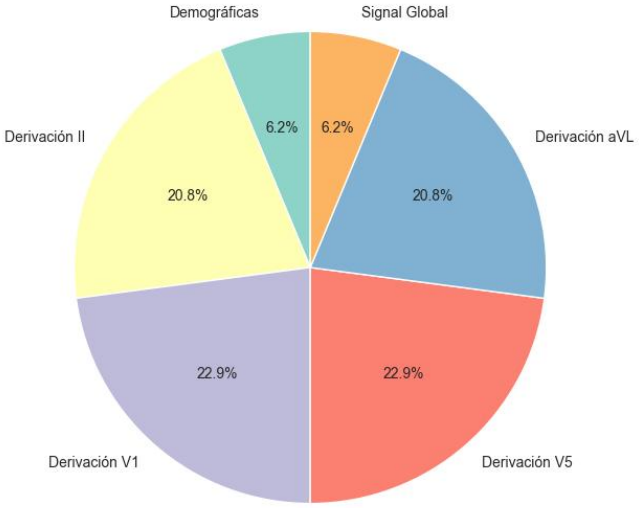

Figura 17 - Representación de variables por grupo de características.

5. Modelado

Esta fase del ciclo CRISP-DM se centra en la aplicación de técnicas de aprendizaje automático para construir modelos capaces de clasificar las patologías estructurales a partir de las 48 características seleccionadas en la etapa anterior. Dado el contexto médico y el severo desbalance de clases identificado en el EDA, la estrategia de modelado se ha diseñado priorizando la capacidad de generalización y, críticamente, la capacidad de detección de las clases minoritarias para su posterior análisis clínico.

5.1. Estrategia de modelado y configuración experimental

Para garantizar la reproducibilidad y la robustez de los resultados, se estableció un entorno controlado con una semilla aleatoria fija (*random_state*=1976). Se utilizaron los conjuntos de datos definidos en el capítulo anterior: entrenamiento (TRAIN, n=65.878) para el ajuste de parámetros y validación (VAL, n=2.053) para la selección de modelos y optimización de hiperparámetros.

5.1.1. Gestión del desbalance de clases

Como se observó en el capítulo 3, la clase "Estenosis" representa apenas un 1,80% de las muestras frente al 65,28% de "Insuficiencia". Para mitigar el sesgo del modelo hacia la clase mayoritaria, se aplicó una estrategia de ponderación de clases (*class weighting*) [75].

Se calcularon pesos inversamente proporcionales a la frecuencia de cada clase utilizando la heurística *balanced* de Scikit-Learn, asignando una penalización drásticamente mayor a los errores en las patologías menos frecuentes:

- "Insuficiencia"**:** peso ≈ 0,38 (baja penalización).
- "Estenosis"**:** peso ≈ 13,90 (muy alta penalización).

5.1.2. Selección de métricas de evaluación

Siguiendo las recomendaciones de TRIPOD-AI, se seleccionaron métricas que evalúan tanto la discriminación como la calibración, descartando la exactitud (*Accuracy*) como métrica única debido al desbalance:

- ***Macro F1-Score*:** Métrica principal para la optimización. Penaliza a los modelos que ignoran las clases minoritarias.
- **Sensibilidad por clase (*Recall*)** [76]**:** Crítica en el ámbito clínico para minimizar los falsos negativos.
- **AUC-ROC (*One-vs-Rest*)** [77]**:** Evalúa la capacidad de discriminación global independiente del umbral de decisión. Esta métrica mide la calidad de las probabilidades predichas a través de todos los posibles puntos de corte de clasificación, sin depender de un valor fijo para asignar la clase.
- ***Log-Loss*** [78] **y *Brier Score*** [40]**:** Métricas probabilísticas esenciales para evaluar la calidad de las probabilidades predichas, requisito indispensable para la fase posterior de Cuantificación de Incertidumbre (UQ).

Cabe destacar que para obtener una visión única del rendimiento del sistema, estas métricas se reportan como agregados no ponderados (*Macro-Averaging*). Esta estrategia otorga el mismo peso a la clase "Estenosis" (minoritaria) que al resto de clases, asegurando que el modelo no sea validado como optimo si falla sistemáticamente en las patologías menos representadas.

5.2. Establecimiento de la línea base (*Baseline*)

Antes de emplear algoritmos complejos, se entrenó un Árbol de Decisión (*Decision Tree Classifier*) para establecer una línea base. Este modelo ofrece alta interpretabilidad y permite validar la dificultad intrínseca del problema.

Se evaluó una versión con parámetros por defecto y otra optimizada mediante búsqueda en rejilla (*GridSearchCV*), explorando profundidades máximas y criterios de división. Los resultados se muestran en la Tabla 11.

Tabla 11 - Rendimiento del modelo baseline.

Modelo	Macro F1	AUC-ROC	Log-*Loss*	***Brier Score***	Gap Train-Val (F1)
BASELINE	**0,3118**	**0,5371**	**18,9435**	**0,2628**	**68,82%** (Muy Alto)
BASELINE optimizado	0,2822	0,5203	19,1191	0,2652	71,78% (Muy Alto)

El árbol de decisión, incluso tras la optimización, mostró limitaciones severas. Se observó un sobreajuste masivo, con un F1-Score de 1,0 en entrenamiento que caía aproximadamente al 0,28 en la validación. Además, la sensibilidad para la clase "Estenosis" fue del 1,9%, como se muestra en la Figura 18, lo que confirma la necesidad de modelos de ensamble más sofisticados.

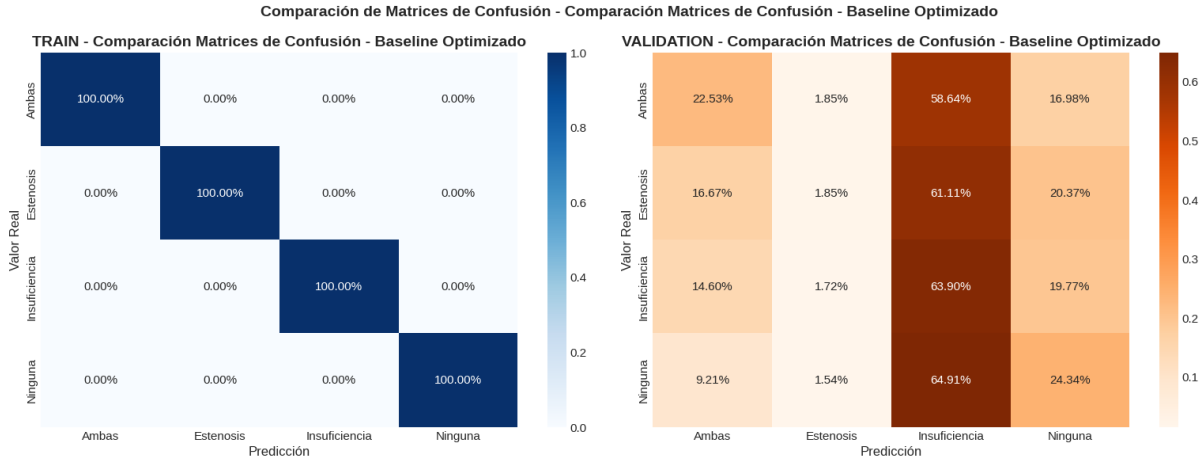

Figura 18- Matriz de confusión del modelo Baseline.

5.3. Selección y entrenamiento de modelos de ensamble

Se seleccionaron cuatro algoritmos de aprendizaje supervisado basados en ensamble, reconocidos por su rendimiento en datos tabulares:

- ***Random Forest*** [23]**:** Algoritmo de *Bagging*.
- **XGBoost** [24] [25]**:** Algoritmo de *Gradient Boosting* optimizado.
- **CatBoost** [28] [29]**:** Variante de *Boosting* que maneja nativamente características categóricas.
- **LightGBM** [26] [27]**:** Algoritmo basado en histogramas, enfocado en eficiencia.

5.3.1. Evaluación inicial (Parámetros por defecto)

En una primera fase, se entrenaron los modelos con sus hiperparámetros por defecto (manteniendo el *class_weight*). Los resultados preliminares (Tabla 12) mostraron que los modelos de *Boosting* superaban significativamente a *Random Forest* en generalización.

Tabla 12 Métricas de rendimiento y calibración de los modelos ensamble.

Modelo	Macro F1	AUC-ROC	Log-*Loss*	***Brier Score***	Gap Train-Val (F1)
BASELINE	0,3118	0,5371	18,9435	0,2628	68,82% (Muy Alto)
Random Forest	0,1981	0,6357	**0,9783**	**0,1363**	80,19% (Muy Alto)
XGBoost	0,3435	**0,6577**	1,1456	0,1626	29,92% (Alto)
CatBoost	0,3184	0,6407	1,2418	0,1735	**8,93% (Moderado)**
LightGBM	**0,3487**	0,6555	1,1574	0,1650	26,81% (Alto)

Aunque XGBoost y LightGBM obtuvieron métricas brutas ligeramente superiores, CatBoost demostró desde el inicio una capacidad de generalización muy superior, con una brecha entre entrenamiento y validación de solo el 8,9%, frente al más de un 25% de sus competidores.

5.4. Optimización bayesiana de hiperparámetros

Con el objetivo de maximizar el rendimiento de los modelos de ensamble seleccionados y mitigar el sobreajuste detectado en las fases preliminares, se procedió a la optimización sistemática de sus hiperparámetros.

Para esta tarea se utilizó **Optuna** [79], un *framework* de *software* de optimización automática de hiperparámetros de nueva generación. Optuna se distingue por su enfoque "*define-by-run*", que permite construir espacios de búsqueda dinámicos, y por el uso del algoritmo TPE (*Tree-structured Parzen Estimator*). A diferencia de la búsqueda aleatoria (*Random Search*) o la búsqueda en rejilla (*Grid Search*), el TPE es un método de optimización bayesiana que modela la probabilidad de que un conjunto de hiperparámetros mejore la métrica objetivo basándose en los resultados de las iteraciones previas, concentrando la búsqueda en las regiones más prometedoras del espacio de hiperparámetros.

5.4.1. Configuración del espacio de búsqueda

Los rangos de los hiperparámetros han sido definidos considerando las particularidades del problema y literatura en cardiología computacional:

- **Características del dataset:** Dado que se cuenta con 65.878 muestras en el conjunto de entrenamiento, el volumen de datos permite el uso de modelos complejos. Sin embargo, al tratarse de un espacio de 48 características y existir un desbalance severo (ratio 36:1), es imperativo priorizar parámetros de regularización.
- **Valores por defecto y restricciones:** Los rangos se centran alrededor de los valores por defecto de las bibliotecas, ampliándose para permitir la exploración, pero acotándose para evitar extremos computacionalmente inviables.
- **Consideraciones computacionales:** Se ha limitado el número de estimadores ("n_*estimators*") a 500 para mantener tiempos de entrenamiento razonables y se ha utilizado una escala logarítmica para la tasa de aprendizaje ("*learning_rate"*), dado que el modelo es más sensible a variaciones en valores bajos.

La Tabla 13 resume la configuración de los espacios de búsqueda definidos para los modelos de *ensamble*.

Tabla 13 - Resumen de hiperparámetros y espacios de búsqueda por modelo.

Hiperparámetro	Descripción	Rango	Justificación
"n_estimators"	Número de árboles	100-500	Un mayor número mejora el rendimiento, pero con rendimientos decrecientes [23].
"max_depth"	Profundidad máxima	3-20 (*Gradient Boosting*) 5-30 (*Bagging*)	Controla la complejidad del modelo. Los algoritmos de *Gradient Boosting* suelen preferir árboles menos profundos (aprendices débiles).
"learning_rate"	Tasa de aprendizaje	0,01-0,3 (log)	Valores más bajos requieren más árboles para converger, pero suelen generalizar mejor.
"min_samples"	Muestras mínimas (hoja/división)	2-50 / 1-20	Actúa como regularizador para evitar el sobreajuste (*overfitting*) en nodos terminales.
"Subsample"	Submuestreo (filas/columnas)	0,6-1,0	Introduce aleatorización en el entrenamiento para reducir la varianza del modelo.
"reg_alpha"/"lambda"	Regularización L1/L2	1e-8 a 10 (log)	Penaliza la complejidad de los pesos. El uso de escala logarítmica mejora la sensibilidad de la búsqueda.

5.4.2. Resultados de la optimización

El proceso de optimización se ejecutó realizando 50 intentos independientes para cada algoritmo. Para garantizar la robustez de los resultados, la función objetivo a maximizar fue el Macro F1-*Score* evaluado mediante una validación cruzada estratificada de 5 particiones (*Stratified 5-Fold CV*), asegurando que la proporción de clases minoritarias se mantuviera constante en cada pliegue.

La Tabla 14 muestra el rendimiento de los modelos con la configuración óptima encontrada, comparando las métricas obtenidas en validación cruzada (TRAIN), conjunto de validación (VAL) y conjunto de prueba sellado (TEST).

Tabla 14 – Métricas de rendimiento y generalización de los modelos ensamble optimizados.

Modelo	TRAIN (CV)	VAL F1	TEST F1	AUC-ROC TEST	Gap CV-VAL	Gap VAL-TEST
BASELINE	0,3240±0,0041	0,3180	0,3165	0,5371	68,8%	+0,5%
Random Forest	0,8267±0,0035	0,3487	0,3510	0,6646	47,8%	-0,2%
XGBoost	0,9758±0,0003	0,3527	0,3551	0,6456	62,3%	-0,2%
CatBoost	0,6453±0,0026	0,3454	0,3445	0,6563	30,0%	+0,1%
LightGBM	1,0000±0,0000	0,3377	0,3219	0,6265	66,2%	+1,6%

- **Análisis de generalización**:
  El análisis de los resultados revela un *trade-off* crítico entre el rendimiento en validación cruzada y la capacidad de generalización:
  - **Sobreajuste severo en modelos de *Boosting*:**
    - **LightGBM** alcanzó un rendimiento perfecto en CV (F1 = 1,0000), indicativo de memorización completa del conjunto de entrenamiento. Su *gap* CV-VAL del 66,2% es el más alto, evidenciando que su estrategia *leaf-wise growth* resultó excesivamente agresiva para este *dataset* desbalanceado.
    - **XGBoost** mostró un comportamiento similar con F1 = 0,9758 en CV pero un colapso a 0,3527 en VAL (*gap* del 62,3%). A pesar de la optimización bayesiana, los hiperparámetros encontrados priorizaron el ajuste a los datos de entrenamiento sobre la generalización.
    - ***Random Forest***, aunque con menor sobreajuste que los modelos de *Boosting* (gap del 47,8%), también evidenció una brecha considerable entre CV y *hold-out*.
  - **CatBoost como modelo más generalizable:**
    CatBoost demostró ser el modelo más estable, con un gap CV-VAL de solo 30,0%, reduciendo a menos de la mitad el sobreajuste observado en XGBoost y LightGBM. Esta ventaja se atribuye a:
    - Su técnica de *ordered Boosting*, diseñada específicamente para prevenir *overfitting* en presencia de ruido y desbalance.
    - Regularización implícita mediante *random permutations* durante el entrenamiento.
  - **Excelente estabilidad VAL→TEST:**
    Todos los modelos mostraron *gaps* mínimos entre validación y test (<2%), validando:
    - La representatividad del conjunto de validación.
    - La ausencia de *data leakage* entre particiones.
- **Análisis de métricas complementarias:**
  - **Capacidad discriminativa. AUC-ROC en TEST** (ver Tabla 14):
    - CatBoost lidera con 0,6563, seguido de *Random Forest* (0,6646) y XGBoost (0,6456).
    - Todos los modelos superan el BASELINE, pero se mantienen en el rango de discriminación moderada (0,6-0,7).
    - LightGBM presenta el peor AUC-ROC (0,6265), coherente con su colapso hacia la clase mayoritaria.
  - **Calibración de probabilidades. *Log-Loss* y *Brier Score* sobre datos de VAL** (ver tabla 15):

- XGBoost ofrece la mejor calibración inicial (menor *Brier Score*), aunque este beneficio debe interpretarse con cautela dado su alto *overfitting*
- Los valores absolutos de *Log-Loss* (0,83-1,05) indican calibración moderada.

Tabla 15 - Calibración de probabilidades (Log-Loss y Brier Score - datos de VAL)

Modelo	Log-Loss	Brier Score
Random Forest	1,0263	0,1538
XGBoost	0,8267	0,1482
CatBoost	1,0358	0,1629
LightGBM	1,0535	0,1521

- **Análisis de las matrices de confusión**:
  Las matrices de confusión de la Figura 19 (VAL vs TEST para cada modelo) confirman que el principal desafío es la correcta identificación de "Estenosis":

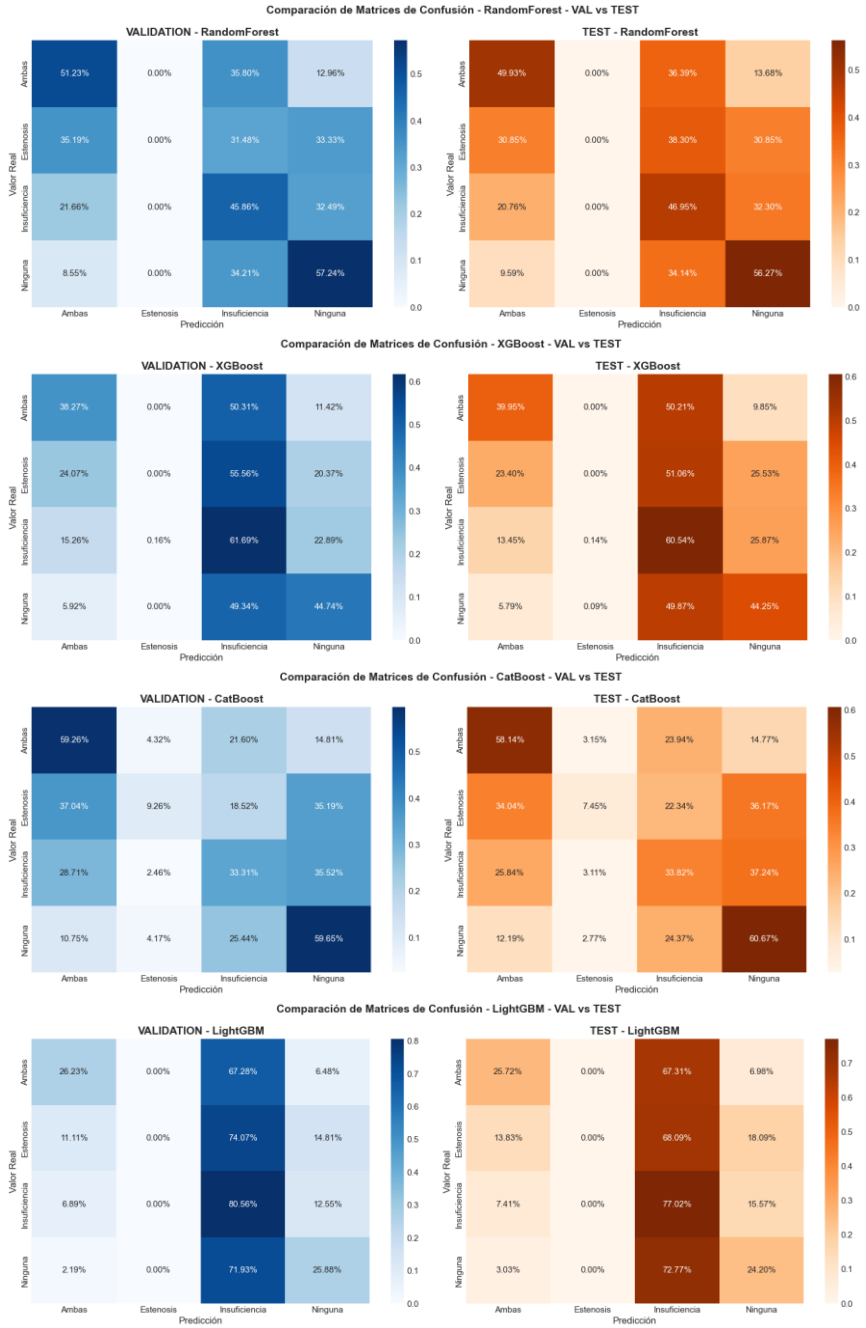

Figura 19 - Matrices de confusión VAL - TEST de los modelos generados.

Observaciones por modelo en TEST:

- ***Random Forest***: Colapso predictivo hacia "Ninguna" e "Insuficiencia". No detecta ningún caso "Estenosis" (sensibilidad = 0%). Queda descartado para análisis XAI por esta clase crítica.
- **XGBoost**: Similar a *Random Forest*, con sensibilidad nula en "Estenosis". Aunque optimiza bien la clase "Insuficiencia" (60,54%), ignora completamente la clase minoritaria crítica.
- **CatBoost**: Se muestra como el modelo más equilibrado, logrando:
  - Única detección de Estenosis (7,45%).
  - Discriminación más homogénea visible en la distribución de errores de su matriz de confusión.
- **LightGBM**: Maximiza agresivamente "Insuficiencia" (77,02%) a costa de todas las demás clases.

- **Análisis de importancia de variables (Feature Importance)**:
  Para comprender los criterios de decisión de los modelos y evaluar su coherencia clínica, se ha extraído la importancia relativa de las variables para cada algoritmo tras el proceso de optimización (ver Figura 20).

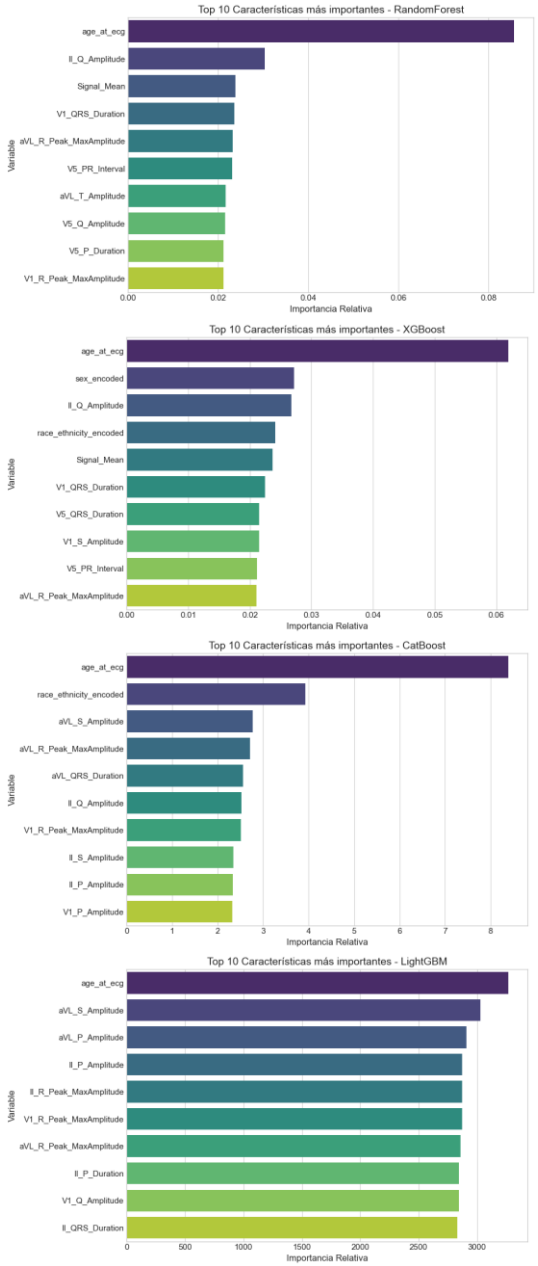

Figura 20 - Importancia de las variables en los modelos optimizados.

- **Predominancia de la edad:** De manera unánime, la variable "*age_at_ecg*" se posiciona como la característica más influyente en los cuatro modelos. Este hallazgo es consistente con la literatura médica, donde la edad es el factor de riesgo demográfico principal para el desarrollo de valvulopatías degenerativas [80].
- **Variables de amplitud y morfología (ECG):**
  - En **XGBoost** y ***Random Forest***, variables como "*II_Q_Amplitude*" y "*Signal_Mean*" ocupan puestos elevados, sugiriendo que la energía total de la señal y las deflexiones iniciales del complejo QRS son críticas para sus predicciones.
  - En **CatBoost**, destaca la inclusión de "*race_ethnicity_encoded*" como segunda variable en importancia, lo que podría indicar que este modelo está capturando sesgos demográficos o prevalencias específicas por etnia que otros modelos omiten.
  - **LightGBM** muestra una distribución de importancia mucho más plana entre sus variables top (como "*aVL_S_Amplitude*" y "*aVL_P_Amplitude*"), lo que explica su comportamiento más errático y su tendencia al sobreajuste al intentar extraer información de múltiples características con pesos muy similares.
- **Consistencia en parámetros eléctricos:** Es notable que medidas de duración y amplitud en derivaciones específicas (especialmente V1, V5 y aVL) aparecen recurrentemente. Por ejemplo, "*V1_QRS_Duration*" y "*aVL_R_Peak_MaxAmplitude*" son consistentes entre XGBoost y *Random Forest*, lo que refuerza la idea de que la hipertrofia ventricular y los retrasos de conducción (reflejados en estas derivaciones) son los principales predictores que la IA utiliza para identificar anomalías valvulares.

5.5. Selección del modelo final

La selección del modelo definitivo no se ha basado únicamente en las métricas de rendimiento global (como el *Macro F1-Score o AUC-ROC*), sino que se ha realizado un análisis pormenorizado de la utilidad clínica del modelo. En un sistema de soporte al diagnóstico, es crítico evaluar cómo se comporta el algoritmo frente a cada patología específica, especialmente aquellas menos frecuentes, pero de mayor gravedad clínica.

La Tabla 16 presenta el desglose de la sensibilidad (*Recall*) por clase para los cuatro modelos evaluados en el conjunto de TEST.

Tabla 16 – Sensibilidad por clase y modelo en el conjunto de TEST.

Modelo	"Ambas"	"Estenosis"	"Insuficiencia"	"Ninguna"
Random Forest	49,93%	0,00%	46,95%	56,27%
XGBoost	39,95%	0,00%	60,54%	44,25%
CatBoost	58,14%	7,45%	33,82%	60,67%
LightGBM	25,72%	0,00%	77,02%	24,20%

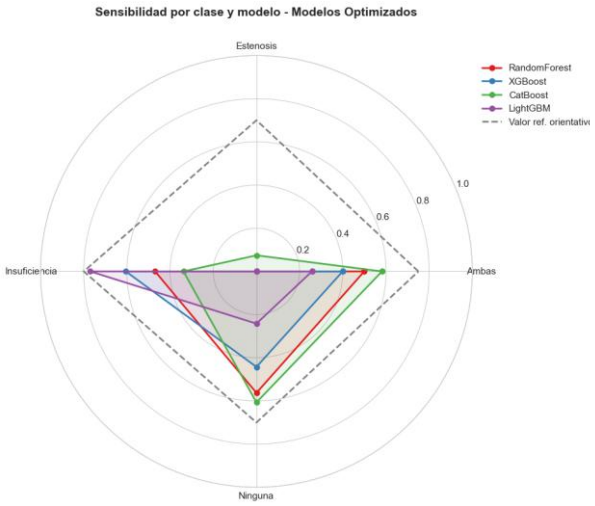

Figura 21 - Comparativa radial de la sensibilidad por clase.

La Figura 21 ilustra comparativamente la sensibilidad por clase mediante un gráfico radial. En ella se evidencia cómo CatBoost ofrece el perfil de detección más equilibrado, logrando una cobertura superior en las categorías críticas ("Ambas" y "Estenosis") frente al sesgo hacia la clase mayoritaria que presentan los demás modelos.

5.5.1. Análisis de la capacidad de detección

El análisis de los resultados revela comportamientos muy dispares entre los algoritmos:

- **Detección de la clase crítica ("Estenosis"):** Este es el factor discriminante más importante. CatBoost es el único modelo capaz de identificar casos de estenosis aórtica (sensibilidad del 7,45%). El resto de los modelos (*Random Forest*, XGBoost y LightGBM), a pesar de la optimización y el uso de pesos de clase, obtuvieron una sensibilidad nula (0,00%) en esta patología, lo que los invalida para un análisis posterior de explicabilidad sobre esta clase.
- **Rendimiento en patología combinada ("Ambas"):** CatBoost también lidera la detección en pacientes con patología doble, alcanzando una sensibilidad del 58,14%, superando notablemente a *Random Forest* (49,93%) y quedando muy por encima de XGBoost (39,95%) y LightGBM (25,72%).
- **Comportamiento frente a la clase mayoritaria:** LightGBM y XGBoost mostraron una clara tendencia a maximizar el rendimiento en la clase "Insuficiencia" (77,02% y 60,54% respectivamente). Por el contrario, CatBoost sacrificó sensibilidad en esta clase mayoritaria (33,82%) para redistribuir su capacidad predictiva hacia las clases más difíciles ("Estenosis" y "Ambas") y la detección de casos sanos ("Ninguna"), donde también obtuvo el mejor resultado con 56,27%.

5.5.2. Análisis de estabilidad VAL vs TEST

La Tabla 17 muestra las variaciones de sensibilidad entre los conjuntos de validación y test.

Tabla 17 - Estabilidad de sensibilidad por clase (Δ = TEST - VAL).

Modelo	"Ambas"	"Estenosis" (Clase crítica)	"Insuficiencia" (Clase mayoritaria)	"Ninguna"
Random Forest	-1,30 pp	+0,00 pp	+1,09 pp	-0,97 pp
XGBoost	+1,67 pp	+0,00 pp	-1,15 pp	-0,48 pp
CatBoost	-1,12 pp	-1,81 pp	+0,51 pp	+1,03 pp
LightGBM	-0,52 pp	+0,00 pp	-3,54 pp	-1,68 pp

- **Variaciones mínimas**: Confirman que el conjunto VAL es representativo de TEST, validando la estrategia de partición estratificada.
- **Valor más elevado de LightGBM en "Insuficiencia"**: (Δ = -3,54 pp), sugiriendo posible sobreajuste específico a patrones de esta clase en VAL que no generalizan a TEST.
- **Consistencia de CatBoost:** Todas sus variaciones están en el rango entre -2 y +1 pp, evidenciando la mayor estabilidad de sus predicciones.

5.5.3. Justificación de la elección de CatBoost

Tras ponderar los resultados, se ha seleccionado **CatBoost** como el modelo final para las fases de Explicabilidad (XAI) y Cuantificación de Incertidumbre (UQ). Esta decisión se fundamenta en tres pilares:

- **Maximización de Verdaderos Positivos en clases complejas:** CatBoost demostró la mayor capacidad para detectar las clases más difíciles ("Estenosis" y "Ambas"). Para cumplir el objetivo del TFG de aplicar técnicas de **XAI** (como SHAP o LIME), es estrictamente necesario que el modelo sea capaz de predecir correctamente instancias de estas patologías; de lo contrario, no existirían casos de éxito sobre los cuales generar explicaciones clínicas.
- **Mitigación del sesgo hacia la mayoría:** Aunque la sensibilidad de CatBoost en la clase "Insuficiencia" quedó en un 33,82%, este comportamiento refleja un modelo que ha priorizado el aprendizaje de patrones distintivos de las clases minoritarias en lugar de recurrir a la probabilidad a priori de la clase dominante. En un contexto de cribado, se prioriza la capacidad de alertar sobre patologías severas o combinadas.
- **Generalización y estabilidad:** Como se observó en el apartado 5.4, CatBoost presentó la menor brecha entre entrenamiento y validación, sugiriendo que sus predicciones son más fiables y menos propensas al sobreajuste que las de XGBoost o LightGBM.
- **Coherencia en la importancia nativa de características:** CatBoost presenta el equilibrio más adecuado para el análisis de XAI y UQ en las próximas fases. A diferencia de LightGBM, cuya importancia de

variables es excesivamente plana y sugiere una dependencia de correlaciones débiles, o de XGBoost, que ignora variables demográficas clave, CatBoost asigna pesos significativos tanto a factores de riesgo clínicos conocidos ("*age_at_ecg*") como a hallazgos eléctricos específicos ("*aVL_S_Amplitude*", "*aVL_QRS_Duration*"). Esta distribución indica un sentido clínico en la importancia nativa y confirma que el trabajo subsiguiente sea biológicamente plausible.

En consecuencia, el modelo **CatBoost optimizado** será el objeto de estudio en los capítulos subsiguientes.

6. XAI

Tras la selección del modelo CatBoost en la fase de modelado (Capítulo 5), debido a su capacidad superior para generalizar y detectar las clases minoritarias críticas (“Estenosis” y “Ambas”), es fundamental facilitar la interpretación de los resultados obtenidos. Dado que CatBoost es un modelo de conjunto basado en árboles (caja opaca), sus decisiones no son intrínsecamente interpretables por el personal clínico.

En este capítulo se aplican técnicas de XAI post-hoc agnósticas y específicas del modelo para desentrañar la lógica interna del algoritmo. Este análisis responde al doble objetivo de validar que el modelo atiende a biomarcadores electrofisiológicos coherentes y cumplir con los requisitos de transparencia del Reglamento Europeo de IA (AI Act).

6.1. Análisis de importancia global

El primer nivel de explicabilidad busca comprender qué variables dominan el comportamiento general del modelo. Para ello, se han contrastado tres enfoques: la importancia nativa del modelo, la Importancia por Permutación (PFI) y los valores SHAP globales.

6.1.1. Importancia Nativa vs. *Permutation Feature Importance* (PFI)

El análisis de importancia nativa (basado en la reducción de impureza en los árboles) identificó a la variable “*age_at_ecg*” (edad) como el predictor dominante, seguido por “*race_ethnicity_encoded*” y amplitudes específicas del ECG como “*aVL_S_Amplitude*”, como se muestra en la Figura 22.

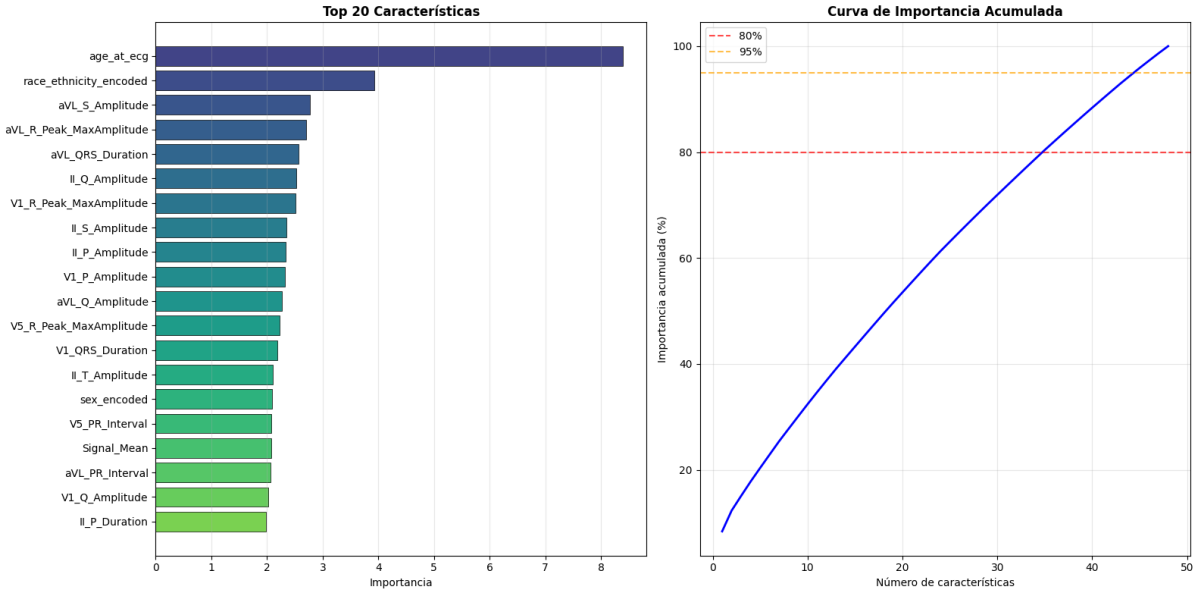

Figura 22 - Resultado del análisis de importancia nativa.

Sin embargo, dado el sesgo conocido de la importancia nativa hacia variables de alta cardinalidad, se contrastó con la **PFI (*Permutation Feature Importance*)** [45] [81] calculada sobre el conjunto de validación utilizando la métrica F1-Macro.

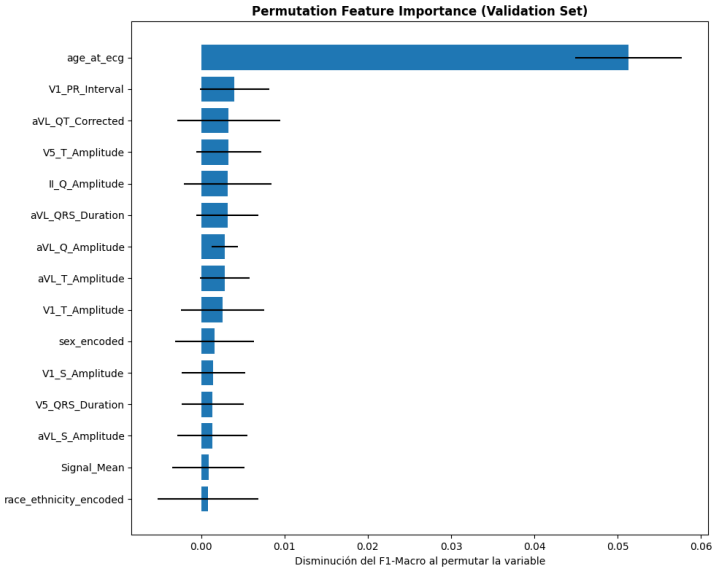

Figura 23 - Resultado del análisis PFI.

Los resultados del PFI (Figura 23) confirman la predominancia de la edad ("*age_at_ecg*") como el factor más discriminante, provocando una caída media del 5% en el F1-Macro al ser permutada. No obstante, el PFI revela que variables electrofisiológicas como "*V1_PR_Interval*", "*aVL_QT_Corrected*" y "*V5_T_Amplitude*" son críticas para la generalización del modelo, superando en relevancia a las variables demográficas secundarias que el método nativo sobrevaloraba.

6.1.2. Importancia global: SHAP

Para una comprensión más profunda, se utilizaron los valores **SHAP (SHapley Additive exPlanations)** [46] [47], que consideran las interacciones entre variables basándose en la teoría de juegos. El gráfico *beeswarm* (Figura 24) muestra no solo la magnitud del impacto, sino la dirección de este.

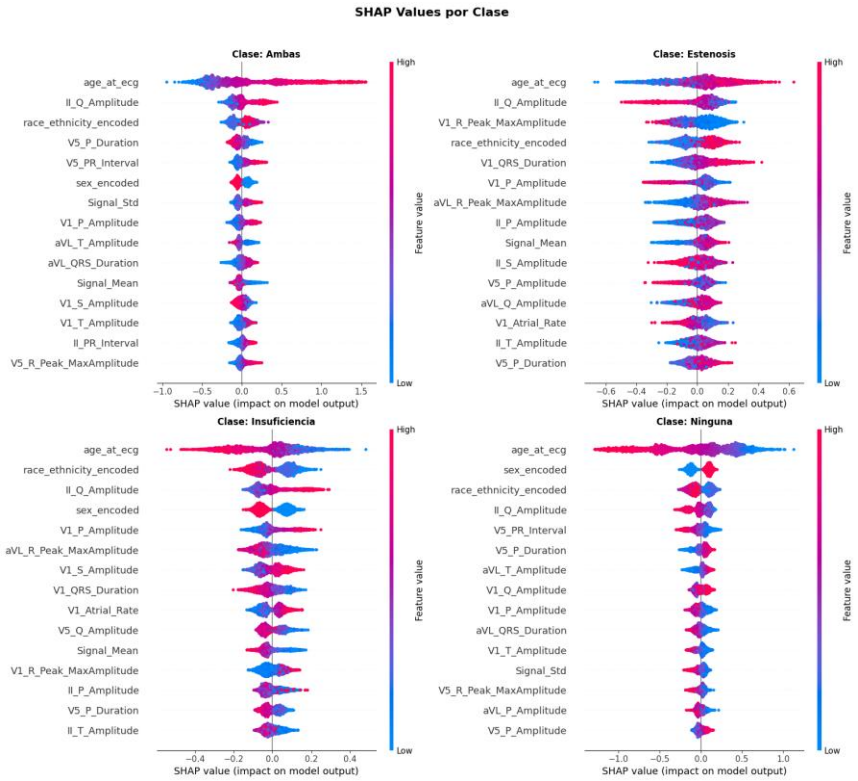

Figura 24 - Resultado del análisis SHAP global.

El análisis global de SHAP revela hallazgos clínicamente coherentes, pero también riesgos potenciales:

- **Edad:** Es el predictor más fuerte. Valores altos de edad (color rojo) se asocian positivamente (mitad derecha del gráfico) con el riesgo de patología estructural, lo cual es consistente con la epidemiología de enfermedades cardiacas valvulares [80].
- **Biomarcadores ECG:** Variables como "*II_Q_Amplitude*", "*V1_P_Amplitude*" y "*V5_P_Duration*" aparecen en el top 10, indicando que el modelo está detectando alteraciones en la despolarización ventricular y auricular características de sobrecargas de presión o volumen [3] [71]. Respecto a la direccionalidad, la Figura 23 confirma que las clases "Ambas" e "Insuficiencia" tienen valores elevados (color rojo) en el lado derecho aumentando el riesgo predicho, pero en "Ninguna" se encuentran en el lado opuesto, penalizando dicha clase.
- **Factores Demográficos:** La variable "*race_ethnicity_encoded*" aparece como la tercera más influyente, lo que levanta una alerta sobre posibles sesgos algorítmicos que se analizarán en la sección 6.4.

6.2. Análisis de comportamiento (PDP e ICE)

Para entender la relación funcional entre las variables críticas y la predicción, se emplearon gráficos de Dependencia Parcial (PDP) [82] y Expectativa Condicional Individual (ICE) [11] (Figura 25).

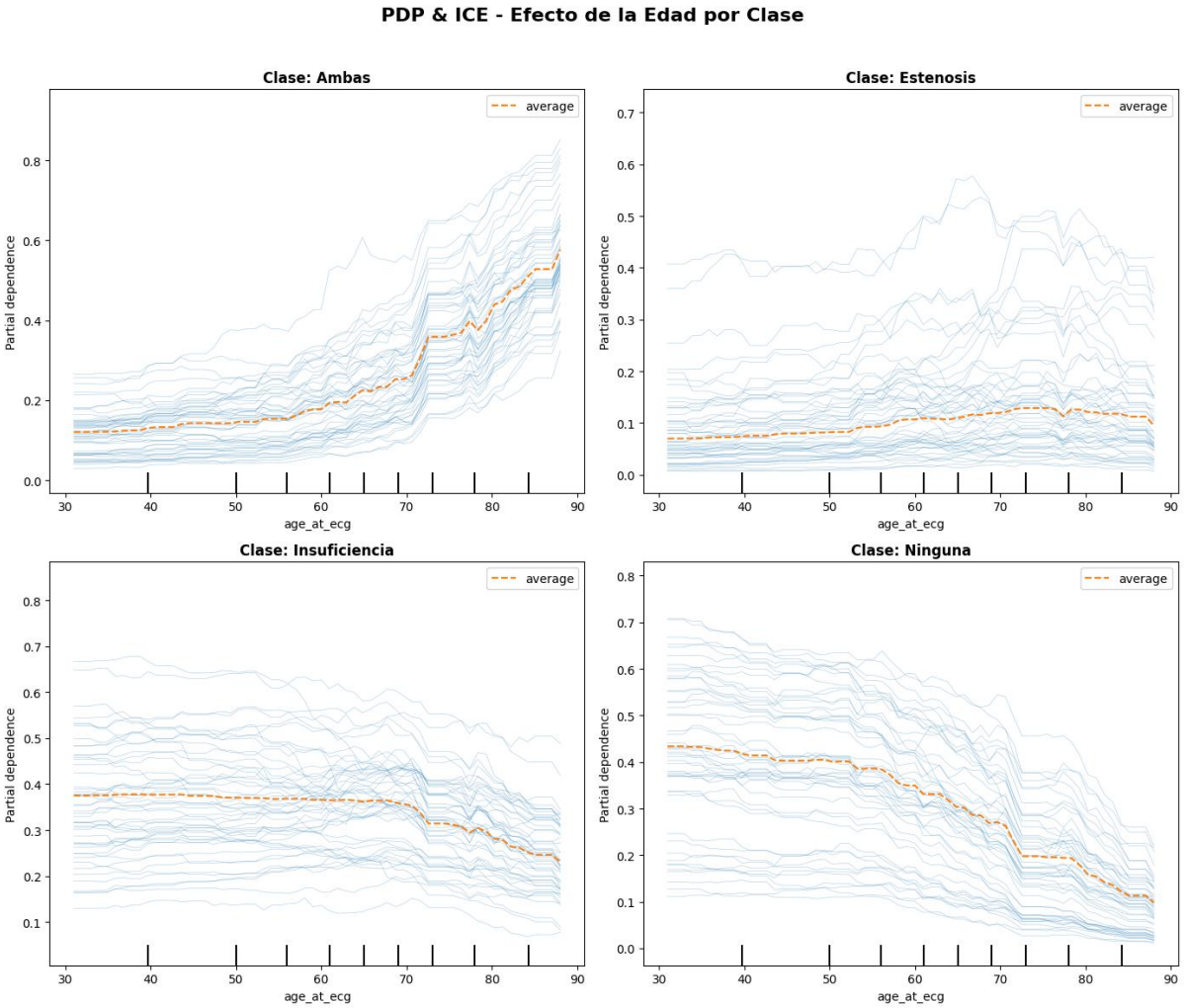

Figura 25 - Resultado del análisis de sensibilidad sobre "age_at_ecg".

- **Efecto de la Edad:** El gráfico PDP muestra una relación no lineal. El riesgo predicho por el modelo se mantiene estable hasta los 60 años, punto a partir del cual asciende abruptamente. Esto mimetiza la prevalencia clínica de la estenosis aórtica en población de edad avanzada [80].

- **Interacciones:** La divergencia de las líneas ICE (azul) respecto al promedio (naranja) sugiere que el impacto de la edad no es uniforme para todas las personas pacientes, sino que depende de interacciones con otras variables del ECG.

6.3. Explicabilidad local: validación diagnóstica

La validación clínica requiere entender decisiones individuales. A continuación, se analizan un caso de éxito (Verdadero Positivo) y un error crítico (Falso Negativo) para la clase "Estenosis".

6.3.1. Análisis de un Verdadero Positivo (TP)

Se analizó al Paciente 1000, diagnosticado correctamente con "Estenosis" con una confianza del 56,72%.

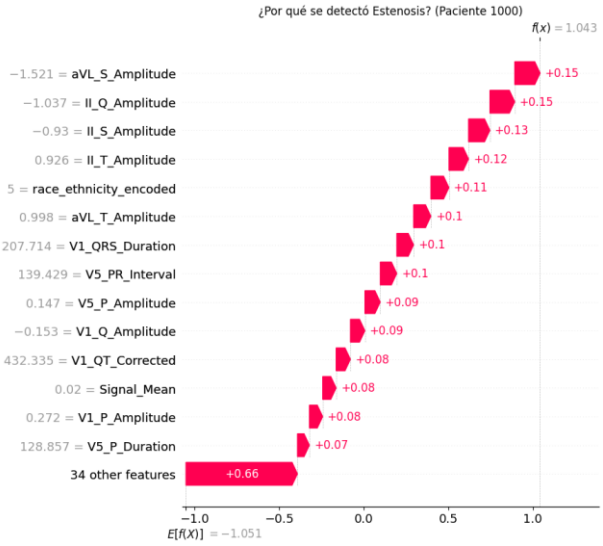

Figura 26 - Resultado de análisis de un TP de la clase minoritaria.

El gráfico *waterfall* de la Figura 26 desglosa la decisión mediante la aditividad de SHAP:

- La predicción base E[f(x)] era negativa (-1,051).
- Las variables *"aVL_S_Amplitude"* (-1,521), *"II_Q_Amplitude"* (-1,037) y la duración del QRS ("*V1_QRS_Duration*" = 207 ms) aportaron valores positivos (+0,15, +0,10) que empujaron la decisión hacia la clase Estenosis.

Validación Cruzada con LIME: Se aplicó LIME [49] [50] a este mismo paciente, confirmando que la duración del QRS (> 201,23) y *"aVL_S_Amplitude"* (<= 1,35) fueron los factores determinantes locales, lo que otorga robustez a la explicación (Figura 27).

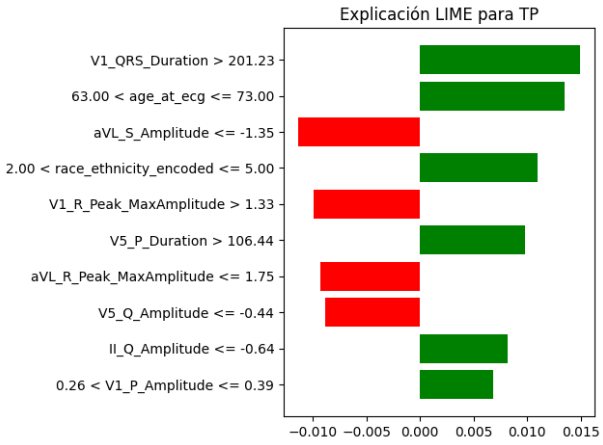

Figura 27 - Validación cruzada con LIME al mismo TP de clase minoritaria.

Estudio DiCE [51]: Se generaron unas explicaciones contrafactuales para mostrar que cambios debería haber en los resultados para diagnosticar al paciente como sano.

De los posibles escenarios (ver Tabla 18), el único realista fue el que indicó que la persona paciente debería tener una disminución de "*V5_PR_Interval*", puesto que los otros indicaban que la persona debería ser más joven para estar sana.

Tabla 18 - Resultados DiCE para el Caso 1000.

Resultado	**"V5_PR_*Interval*"**	**"*age_at_ecg*"**	**"cardiopatia"**
Modelo	139,4	70,0	"Estenosis"
Contrafactual 1	-	42,1	"Ninguno"
Contrafactual 2	-	19,8	"Ninguno"
Contrafactual 3	51,8	-	"Ninguno"

6.3.2. Análisis de un Falso Negativo (FN)

Se examinó al **Paciente 48**, quien padece "Estenosis", pero fue clasificado como "Ninguna" (Sano).

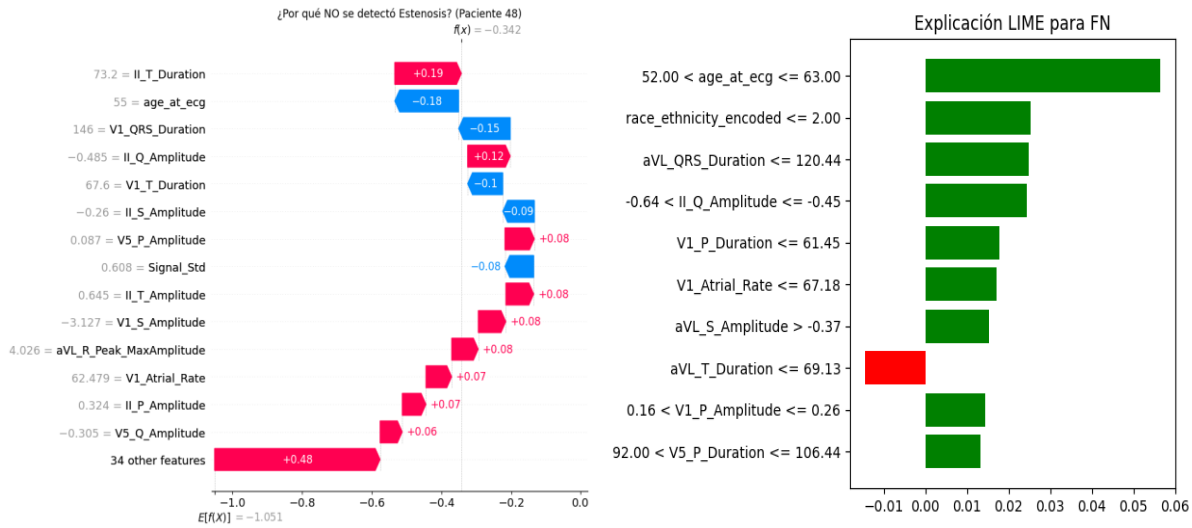

Figura 28 - Resultado de análisis SHAP y LIME de un FN de la clase minoritaria.

El análisis revela la causa raíz del error (Figura 28):

- **Enmascaramiento por Edad:** La persona paciente tiene 55 años. El modelo penalizó fuertemente el riesgo (-0,18 en valor SHAP y principal variable contribuyente en LIME) debido a que esta edad se encuentra en la zona "segura" según el perfil aprendido en el PDP.
- **Señales Contradictorias:** A pesar de tener alteraciones en la duración de la onda T, otros biomarcadores actuaron reduciendo la probabilidad de patología.

Para profundizar en el análisis, se utilizó **DiCE** [51] para generar explicaciones contrafactuales. Se identificó que para predecir "Estenosis" (ver Tabla 19), la persona debería (principalmente) tener más edad. Esto refuerza lo detectado previamente, indicando que el modelo "asume" que en ese rango de edad es muy difícil tener estenosis aórtica.

Tabla 19 - Resultados DiCE para el Caso 48.

Resultado	"V5_PR_*Interval*"	"*age_at_ecg*"	"cardiopatía"
Modelo	86,4	55,0	"Ninguno"
Contrafactual 1	-	75,3	"Estenosis"
Contrafactual 2	-	75,5	"Estenosis"
Contrafactual 3	611,8	78,4	"Estenosis"

6.4. Auditoría de equidad

Dado que la variable "*race_ethnicity*" apareció consistentemente como un predictor importante en SHAP, se procedió a una auditoría de equidad en el conjunto de validación para evaluar el cumplimiento de los principios de no discriminación del *AI Act*. Los resultados de esta auditoría se muestran en la Tabla 20.

Tabla 20 – Desempeño con base en la variable "race_ethnicity".

Grupo Étnico	Muestras	*Accuracy*	Tasa Falsos Negativos (FNR)
"asian"	62	16,1%	0,0%
"black"	314	4,1%	**100,0%**
"hispanic"	582	5,8%	85,7%
"white"	615	14,0%	62,5%
"unknown"	285	9,1%	83,3%

Hallazgo Crítico: Se ha detectado una disparidad severa. El modelo presenta una Tasa de Falsos Negativos del 100% en el grupo *"black"* y del 85,7% en el grupo *"hispanic"*. Esto indica que el modelo ha aprendido sesgos presentes en el desbalance de los datos de entrenamiento, lo que representa un riesgo de discriminación algorítmica inaceptable.

6.5. Plan de mitigación y futuras iteraciones

Ante la detección de este sesgo crítico, y dado que la eliminación de la variable *"race_ethnicity"* no corregiría el problema debido a la existencia de variables redundantes en el ECG (*proxies*), se define la siguiente estrategia para la siguiente iteración del modelo:

- **Estrategias a nivel de datos (Preprocesamiento):**
    - Dada la inviabilidad de recolectar nuevos datos multicéntricos en el corto plazo para balancear las clases demográficas, se aplicará **Ponderación** (*Re-weighting*). Se asignarán pesos mayores a las muestras de los grupos minoritarios ("*black*" y "*asian*") en la función de coste durante el entrenamiento para forzar al modelo a aprender sus patrones específicos.
- **Estrategias a nivel de modelo (In-procesamiento):**
    - Se implementarán **restricciones de equidad** (*Fairness Constraints*) directamente en el algoritmo de aprendizaje. Se añadirá un término de regularización a la función de pérdida que penalice las diferencias significativas en la Tasa de Falsos Negativos entre grupos étnicos.
- **Estrategias a nivel de validación (Posprocesamiento):**
    - **Ajuste de umbrales por grupo:** Se calibrarán umbrales de decisión específicos para cada grupo étnico con el fin de igualar la oportunidad (*Equalized Odds*), asegurando que el riesgo clínico sea equivalente para todas las personas pacientes.

- **Auditoría contrafactual:** Se sistematizará el uso de DiCE para verificar que la alteración sintética de la etiqueta de etnia no modifique la predicción clínica.

- **Análisis causal e interdisciplinar:**
  - Se establecerá un grupo de trabajo con profesionales sanitarios para investigar el origen clínico o sociodemográfico del sesgo étnico detectado. Es fundamental discernir la disparidad en la tasa de error, ya que la causa raíz determinará la validez ética de las correcciones estadísticas aplicadas.
- **Supervision humana (Human-in-the-loop):**
  - Independientemente de las mejoras técnicas, se establece como requisito obligatorio que cualquier predicción sobre grupos demográficos con alta incertidumbre (ver Capítulo 8 sobre UQ) o pertenecientes a los grupos afectados por el sesgo sea derivada automáticamente para revisión por una persona especialista en medicina, garantizando la seguridad de las personas pacientes.

7. Estrategias de mitigación de sesgo

7.1. Introducción y contexto normativo

El desarrollo de sistemas de soporte a la decisión clínica basados en Inteligencia Artificial conlleva un imperativo ético, clínico y legal: garantizar que sus predicciones sean no solo precisas, sino también equitativas y seguras para todos los grupos poblacionales.

De acuerdo con el Reglamento (UE) 2024/1689 (AI Act)**,** los sistemas de IA utilizados en contextos de triaje médico, diagnóstico o apoyo a decisiones clínicas se clasifican como sistemas de alto riesgo, lo que implica obligaciones explícitas de gestión de riesgos y control de sesgos discriminatorios [10]. En este marco regulatorio, la evaluación de disparidades en métricas críticas como la Tasa de Falsos Negativos (FNR) no constituye únicamente una buena práctica técnica, sino un requisito alineado con los artículos relativos a calidad de datos, evaluación de riesgos y mitigación de impactos adversos.

En el capítulo anterior se identificó que el modelo base elegido presentaba una disparidad significativa en la FNR, afectando de manera desproporcionada al grupo demográfico *Black*. Dado que, en un contexto cardiovascular, un falso negativo implica la omisión de una patología potencialmente grave, la minimización de la FNR se considera clínicamente prioritaria frente a otros tipos de error.

En este capítulo se evalúa la eficacia de dos estrategias de mitigación del sesgo implementadas mediante la librería Fairlearn [83]:

- ***Blind* (Ceguera)**: eliminación explícita de la variable sensible "*race_ethnicity*", bajo el paradigma de *Fairness through Unawareness*.
- ***Weighted* (Ponderación)**: reajuste de los pesos de entrenamiento para penalizar errores (FNR elevados) en combinaciones grupo–clase históricamente desfavorecidas.

El objetivo no es únicamente reducir disparidades numéricas, sino determinar si dicha reducción es compatible con la validez clínica y la seguridad del sistema, evaluando los modelos mediante métricas de equidad [84] [85], explicabilidad global (SHAP) [82] y análisis local (LIME y DiCE) [49] [86].

7.2. Evaluación de equidad: la paradoja de las métricas agregadas

7.2.1. Resultados globales

Las métricas agregadas iniciales sugieren una mejora sustancial en la FNR del grupo *Black* al aplicar la estrategia *Blind*:

- **Modelo Base**: FNR (Black) = 33,98%
- **Modelo *Weighted***: FNR (Black) = 26,62%
- **Modelo *Blind***: FNR (Black) = 0,53%

Si el análisis se limitara a métricas de igualdad de oportunidades (*Equality of Opportunity*) [87], el modelo *Blind* podría interpretarse erróneamente como el más equitativo. Sin embargo, la literatura advierte que la eliminación de variables sensibles raramente elimina el sesgo real debido a la presencia de variables *proxy*, y puede incluso degradar el rendimiento clínico global [56] [88].

Este resultado ilustra un fenómeno bien documentado: una métrica de equidad favorable no garantiza un modelo justo ni clínicamente útil.

7.2.2. Análisis desagregado por clase patológica

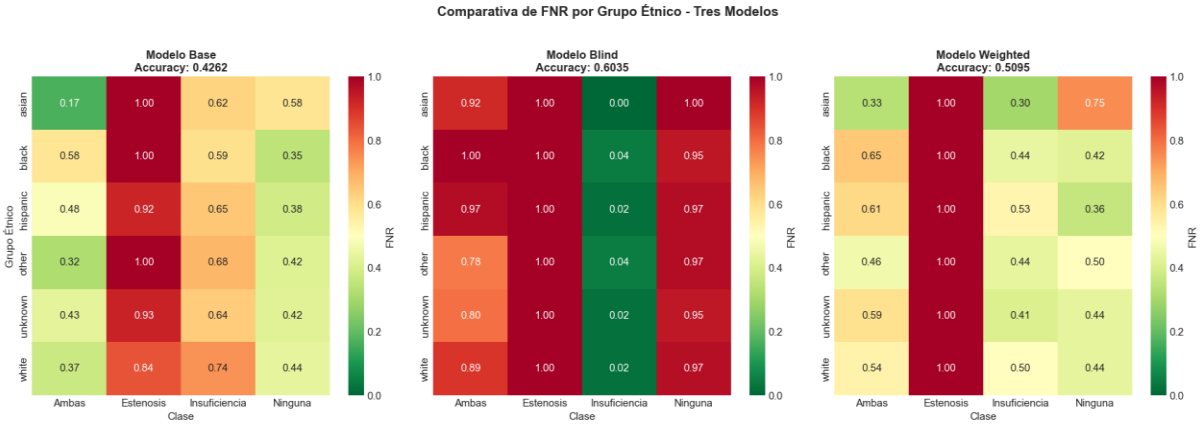

Figura 29 - Mapa de calor de los FNR por grupo étnico y clase patológica para los tres modelos.

El análisis desagregado del rendimiento por clase patológica (Figura 29) revela comportamientos que las métricas agregadas ocultan:

- **Colapso del modelo *Blind***
  El mapa de calor correspondiente al modelo *Blind* muestra una FNR cercana a 0,00 en la clase "Insuficiencia" para todos los grupos, pero simultáneamente una FNR cercana a 1,00 en la clase "Ninguna". Este patrón indica que el clasificador ha perdido su capacidad discriminatoria, prediciendo sistemáticamente la presencia de patología. Aunque este comportamiento minimiza matemáticamente los falsos negativos, genera una tasa de falsos positivos clínicamente inaceptable, lo que en un entorno real saturaría los recursos médicos y violaría el principio de no maleficencia [89].
- **Invisibilidad sistémica de la estenosis**
  En los tres modelos, la clase "Estenosis" presenta una FNR cercana a 1,00 para prácticamente todos los grupos. Ninguna de las estrategias de mitigación logra que el modelo aprenda patrones discriminativos para esta patología, lo que sugiere un problema estructural de datos (insuficiencia de muestra o ruido en las etiquetas) más que un sesgo algorítmico corregible mediante ponderación [88].
- **Estabilidad del modelo *Weighted***
  A diferencia del modelo *Blind*, la estrategia *Weighted* mantiene la estructura general de predicción. Se observa una mejora legítima en la detección de Insuficiencia para el grupo "*black*" (reducción de FNR de 0,59 a 0,47), sin un colapso equivalente de la especificidad en pacientes sanos.

7.2.3. Cuantificación de disparidades

Tabla 21 - Comparativa de FNR y disparidad máxima por clase y modelo.

Clase	Disp_Base	Disp_*Blind*	Disp_*Weighted*	Mejora_*Blind*	Mejora_*Weighted*
"Ambas"	0,414	0,220	0,312	0,194	0,102
"Estenosis"	0,158	0,000	0,000	0,158	0,158
"Insuficiencia"	0,146	0,043	0,238	0,103	-0,092
"Ninguna"	0,238	0,055	0,387	0,183	-0,149

La Tabla 21 resume las disparidades máximas observadas entre grupos étnicos para cada clase patológica. Es importante destacar que una baja disparidad no implica necesariamente un modelo justo o clínicamente válido, como ilustra el caso del modelo *Blind*, donde la aparente equidad es consecuencia directa del colapso del clasificador y no de una mejora real en el aprendizaje.

7.3. Importancia global: SHAP

Para comprender los mecanismos internos de decisión, se utilizaron valores SHAP (ver Figura 30), que cuantifican la contribución marginal de cada característica a la predicción del modelo.

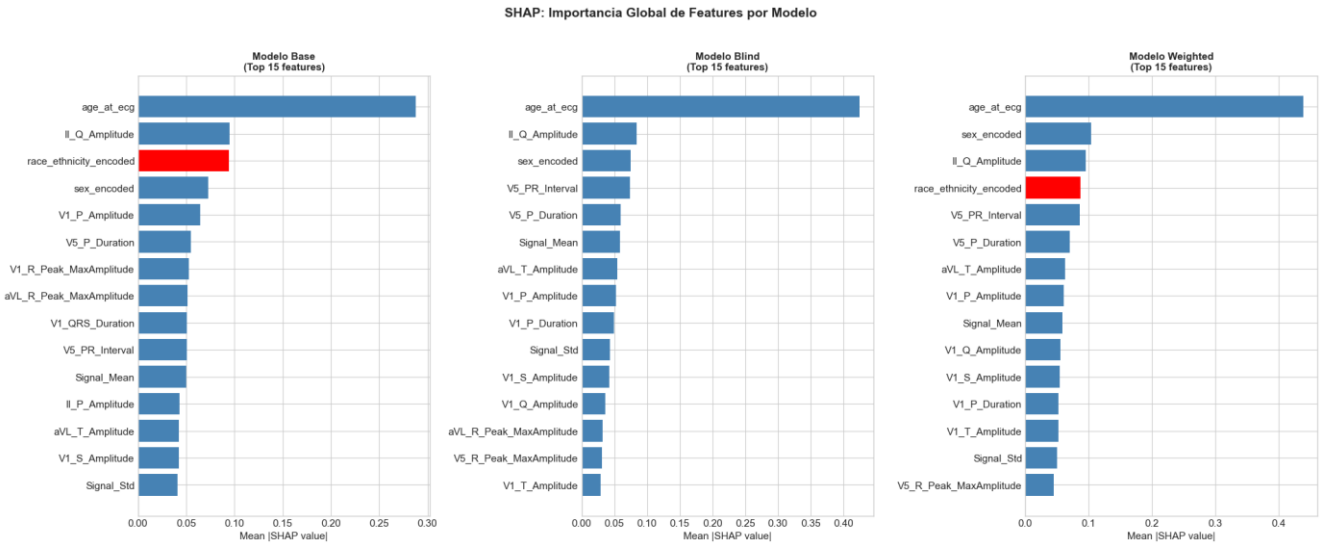

Figura 30 - Gráficos de importancia global SHAP comparativos.

En los modelos Base y *Weighted*, la variable "*race_ethnicity_encoded*" aparece consistentemente como la tercera y cuarta característica más influyente respectivamente, confirmando que el modelo utiliza activamente la etnia como factor de decisión. Dado el contexto histórico de desigualdades en datos médicos, esta dependencia es problemática si se basa en correlaciones espurias más que en relaciones clínicamente causales [90].

En el modelo *Blind*, la eliminación explícita de etnia provoca una redistribución de la importancia hacia variables del ECG y otras como sexo, lo que confirma la hipótesis de Barocas et al. [56]: los modelos "ciegos" tienden a explotar variables *proxy* para recuperar información sensible perdida, potencialmente exacerbando otros sesgos, como el sesgo etario.

Cabe destacar que los valores SHAP reflejan dependencias estadísticas aprendidas por el modelo y no relaciones causales, por lo que una alta importancia de una variable no debe interpretarse como evidencia de causalidad médica directa.

7.4. Análisis local y diagnóstico de fallos

7.4.1. El coste de la corrección: Caso 1000

La persona paciente 1000 ("*white*", "Estenosis") fue diagnosticada correctamente por el modelo Base, pero clasificada erróneamente por el modelo *Weighted* ("Ninguna") y por el modelo *Blind* ("Insuficiencia").

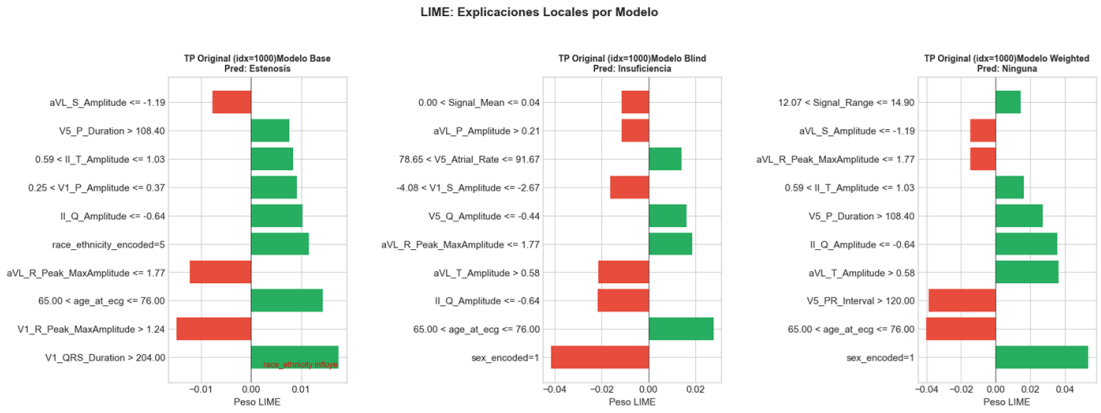

Figura 31 - Explicación LIME del Caso 1000.

El análisis LIME muestra que, en el modelo *Weighted* (Figura 31), variables como "*sex_encoded*" y "*aVL_T_Amplitude*" empujan la predicción hacia "Ninguna", anulando señales patológicas detectadas previamente. Este caso ilustra el *trade-off* entre equidad y exactitud descrito en la literatura [88] [85]: al forzar mejoras para un grupo, pueden introducirse errores en otros si no se aborda el problema de manera estructural.

7.4.2. La barrera de la edad: Caso 48

La persona paciente 48 (*"hispanic", "*Estenosis") representa un falso negativo persistente en todos los modelos.

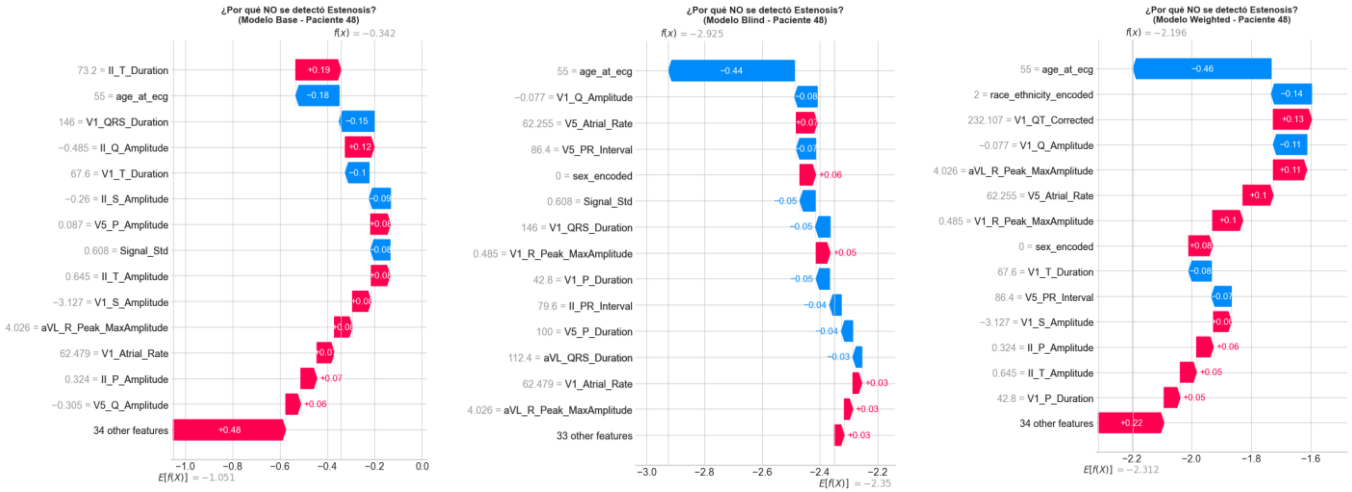

Figura 32 - Explicación SHAP local del caso 48.

El análisis SHAP (Figura 32) revela que la edad ("*age_at_ecg*" = 55) tiene una contribución negativa significativa a la probabilidad de estenosis, sugiriendo que el modelo ha aprendido un patrón implícito del tipo "personas pacientes relativamente jóvenes no presentan estenosis". Este sesgo etario impide el diagnóstico correcto incluso cuando existen señales fisiológicas relevantes.

7.5. Análisis contrafactual (DiCE)

Para profundizar en la causa raíz del error del Caso 48, se generaron explicaciones contrafactuales mediante **DiCE** [86].

Tabla 22 - Resultados DiCE para el Caso 48 en el modelo Base.

Resultado	"II_PR_Interval"	"age_at_ecg"	"cardiopatia"
Modelo	79,6	55,0	"Ninguno"
Contrafactual 1	-	73,2	"Estenosis"
Contrafactual 2	-	70,3	"Estenosis"
Contrafactual 3	687,8	72,4	"Estenosis"

En el modelo Base DiCE indica que, para diagnosticar Estenosis, la persona paciente debería tener aproximadamente 73 años y una amplitud de onda R significativamente mayor (Tabla 22), confirmando la existencia de un sesgo por edad severo.

En el modelo *Weighted*, DiCE no logra generar ningún contrafactual factible para diagnosticar únicamente "Estenosis", aunque si son capaces de crearlos para diagnosticar el valor "Ambas". Este hallazgo es crítico: indica que la clase "Estenosis" ha quedado tan marginada en el espacio de decisión del modelo que no existe una combinación fisiológicamente plausible de variables cercanas al paciente que active dicho diagnóstico. Desde una perspectiva clínica, esto compromete seriamente la aplicabilidad del modelo en escenarios reales.

7.6. Conclusiones

El análisis multidimensional realizado permite extraer las siguientes conclusiones:

- La estrategia ***Blind*** **debe descartarse**. Aunque reduce numéricamente la disparidad, lo hace a costa de la validez clínica, incurriendo en una forma de equidad ilusoria incompatible con sistemas médicos de alto riesgo.
- La estrategia ***Weighted*** **ofrece una mejora real y controlada** en la detección de "Insuficiencia" para todos los grupos, manteniendo una estructura de predicción estable.
- Ninguna estrategia algorítmica corrige los problemas estructurales asociados a la "Estenosis", lo que indica que la raíz del problema reside en la calidad y representatividad de los datos.
- Estos resultados refuerzan la necesidad de combinar mitigación algorítmica con **intervenciones en la fase de preprocesamiento**, validación estratificada y auditorías de equidad continuas, en línea con las recomendaciones de Chen et al. [89].

8. UQ

8.1. Introducción y objetivos

Hasta este punto, la evaluación del modelo se ha centrado en métricas de rendimiento puntual (*Accuracy*, F1-*Score*) y equidad. Sin embargo, para un despliegue clínico seguro, el AI Act y las guías TRIPOD-AI exigen que el sistema no solo sea preciso, sino que ha de implementar protocolos de validación y de detección de anomalías [11] [37].

El objetivo de este capítulo es operacionalizar la incertidumbre predictiva del modelo final (CatBoost) para transformar sus puntuaciones "*raw scores*" en medidas de confianza accionables. Para ello, se implementa un *pipeline* de Cuantificación de Incertidumbre (UQ) para obtener Predicción Conforme (*Conformal Prediction*) y así generar conjuntos de predicción con garantías estadísticas de cobertura [42].Finalmente, se propone un protocolo de derivación clínica (semáforo) que decide cuándo el sistema podría diagnosticar automáticamente y cuándo requiere intervención humana, aunque siempre se recomienda la revisión por una persona especialista en el ámbito sanitario concreto.

8.2. Predicción Conforme (*Conformal Prediction*)

Para garantizar estadísticamente la fiabilidad, se aplica *Split Conformal Prediction* utilizando la librería MAPIE. A diferencia de la clasificación tradicional que devuelve una sola etiqueta, este método devuelve un conjunto de predicciones, por ejemplo, ("Estenosis", "Insuficiencia"), que garantiza contener la clase real con una probabilidad predefinida (1-α) [42].

8.2.1. Configuración experimental

Método: *Split Conformal* con puntuación de conformidad RAPS (*Regularized Adaptive Prediction Sets*).

Nivel de significancia (**α**): Se estableció un α = 0,20 para fines ilustrativos en este TFG. Esto implica que se busca una garantía teórica del 80% de que el diagnóstico real esté dentro del conjunto devuelto.

8.2.2. Resultados de cobertura y tamaño del Set

Los resultados obtenidos sobre el conjunto TEST se resumen a continuación:

Cobertura empírica global: 94,97%.

A pesar de configurar un α=0,20 (meta del 80%), la cobertura real fue superior al 95%. Esto indica que el modelo conforme es conservador y robusto, superando las garantías mínimas exigidas.

Tamaño medio del set: 2,82 clases.

Dado que el problema tiene 4 clases, un tamaño medio de 2,82 indica una alta ambigüedad. El modelo rara vez se atreve a dar un diagnóstico único con las garantías exigidas.

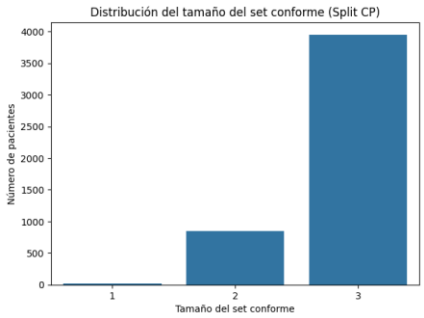

Figura 33 - Histograma del tamaño de los conjuntos de predicción (Cardinalidad).

La Figura 33 muestra que la mayoría de las predicciones resultan en conjuntos de 3 clases (barra más alta), lo que refleja la dificultad del modelo para discriminar con certeza entre las patologías valvulares.

8.2.3. Validación de cobertura en la clase crítica

Aunque la cobertura global superó ampliamente el objetivo (94,97% vs 80%), es imperativo verificar si esta garantía se distribuye equitativamente entre las distintas patologías. El análisis desagregado para la clase crítica Estenosis reveló una disparidad significativa:

- **Cobertura esperada**: 80,0%
- **Cobertura real ("Estenosis")**: 75,0%

Interpretación del déficit:

Este resultado evidencia el problema de la cobertura marginal vs. condicional. El algoritmo de calibración calculó un umbral de conformidad único ($\hat{q}$) basado en el promedio de la población. Dado que las clases mayoritarias ("Ninguna", "Insuficiencia") son más fáciles de predecir, el umbral global resultó ser demasiado exigente para la clase "Estenosis", que presenta mayor incertidumbre intrínseca.

Como consecuencia, para el 25% de las personas con Estenosis real, el modelo no incluyó la patología correcta dentro del conjunto de predicción, violando el principio de seguridad en el subgrupo más vulnerable.

Implicación Clínica:

Este hallazgo sugiere que, para futuras iteraciones, no basta con *Split Conformal Prediction* estándar. Se requiere implementar *Mondrian Conformal Prediction* (calibración estratificada por clase), lo cual forzaría al algoritmo a calcular un umbral $\hat{q}$ específico para cada patología, garantizando el 80% de cobertura independientemente del tamaño o dificultad de la clase [42].

8.3. Integración clínica: Sistema de derivación

Finalmente, se diseñó un algoritmo de decisión clínica basándose en la Predicción Conforme (tamaño del set) para clasificar a las personas pacientes en dos flujos: Diagnóstico "automático" o Revisión humana obligatoria.

Las reglas definidas fueron:

- **Revisión (set vacío)**: Si el modelo no encuentra ninguna clase compatible (Cardinalidad = 0).
- **Revisión (ambigüedad)**: Si el modelo devuelve más de 1 clase posible (Cardinalidad > 1).
- **Automático**: Solo si el set contiene una única clase.

8.3.1. Resultados del *pipeline*

Al aplicar estas reglas al conjunto TEST, la distribución de la carga de trabajo se muestra en la Figura 34:

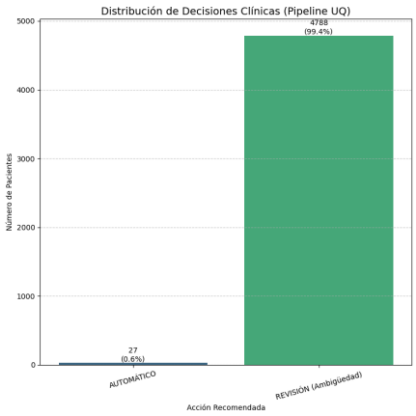

Figura 34 - Porcentaje de personas pacientes derivadas a revisión vs. diagnosticados "automáticamente".

- **Revisión (ambigüedad)**: 99,4% (4788 pacientes).
- **Automático**: 0,6% (27 pacientes).

8.4. Conclusiones

La aplicación de técnicas de Cuantificación de Incertidumbre (UQ) revela una realidad crítica sobre la viabilidad del modelo actual para la automatización completa:

- **Alta seguridad, baja eficiencia**: El sistema prioriza la seguridad de la persona paciente. Al exigir garantías estadísticas (*Conformal Prediction*), el sistema detecta correctamente que no tiene suficiente evidencia para distinguir entre clases en la gran mayoría de los casos.
- **El costo de la garantía**: Para asegurar (con un α=0,20) que no se comenten errores, el modelo se ve obligado a devolver conjuntos de predicción muy amplios (promedio de 2,82 clases), lo que deriva al 99,6% de las personas pacientes a revisión manual.
- **Utilidad como triaje**: Aunque el modelo no sirve como herramienta definitiva en soporte al diagnóstico médico, funciona eficazmente como una herramienta de segunda opinión o triaje, identificando un 0,6% de casos donde la certeza es elevada, y alertando sobre la ambigüedad en el resto.

9. Conclusiones, discusión y líneas de trabajo futuras

9.1. Conclusiones generales

El presente Trabajo de Final de Grado ha completado con éxito el ciclo de vida de un proyecto de Inteligencia Artificial y Ciencia de Datos aplicado al ámbito clínico, siguiendo la metodología CRISP-DM. Se ha desarrollado, validado y auditado un sistema de soporte a la decisión clínica capaz de clasificar patologías cardíacas estructurales a partir de datos tabulares de ECG, superando el enfoque de "caja opaca" mediante la integración de capas de explicabilidad y fiabilidad.

Las principales conclusiones extraídas del estudio son:

1. **Superioridad del *Gradient Boosting* en datos tabulares**: Tras evaluar múltiples arquitecturas, CatBoost se consolidó como el modelo con mejor rendimiento, demostrando mejor manejo de variables categóricas y datos desbalanceados sin necesidad de preprocesamiento agresivo. Su capacidad de generalización superó a los modelos de referencia (*Decision Tree*) y a otras técnicas de ensamble (*Random Forest*), especialmente en métricas de calibración (*Brier Score*).
2. **Gestión efectiva del desbalance sin datos sintéticos**: Se ha demostrado que es posible abordar el severo desequilibrio de clases mediante ponderación de la función de pérdida (*class weighting*), sin recurrir a técnicas de generación de datos sintéticos como SMOTE. Esto asegura que el modelo aprenda la prevalencia real de las patologías sin introducir artefactos estadísticos que inflen artificialmente el rendimiento.
3. **Explicabilidad alineada con la fisiología**: El análisis mediante SHAP y Contrafactuales ha revelado que el modelo basa sus decisiones en marcadores fisiológicos coherentes, validando que el algoritmo ha aprendido patrones médicos reales y no sesgos espurios de los datos.
4. **La fiabilidad como prioridad clínica**: La implementación de Predicción Conforme (*Conformal Prediction*) ha permitido transformar las puntuaciones crudas del modelo en conjuntos de predicción con garantías estadísticas formales. Con una cobertura objetivo del 80%, el sistema es capaz de admitir su incertidumbre entregando conjuntos de etiquetas múltiples en casos ambiguos, en lugar de forzar una única predicción errónea. Esto es crítico en medicina, donde el error de omisión (falso negativo) puede ser fatal.

9.2. Alcance del sistema

El sistema desarrollado se define como una herramienta de triaje y segunda opinión, y nunca como una solución de diagnóstico autónomo. Su alcance operativo se circunscribe a:

- **Entorno de uso**: Servicios de cardiología o atención primaria con acceso a electrocardiógrafos digitales que proporcionen datos tabulares procesados o con conexión a equipos informáticos capaces de realizar el procesado.
- **Función**: Dar soporte a las personas especialistas en cardiología para así reducir el tiempo de diagnóstico para poder priorizar dentro de las listas de espera y ofrecer un sistema de alerta temprana.
- **Población**: Personas pacientes adultas con características similares a la demografía del *dataset* de origen (principalmente población occidental, con distribución de edad y sexo equilibrada en la muestra general).

9.3. Limitaciones del estudio

Es imperativo reconocer las limitaciones inherentes al enfoque adoptado, tal como exigen los estándares de ética y rigor científico:

- **Pérdida de información en datos tabulares**: Al trabajar con características extraídas y preprocesadas en lugar de la señal de onda bruta, el modelo está limitado por dicho preprocesamiento. Matices sutiles de la morfología de la onda podrían haberse perdido.
- **Imputación**: Entre el 5-8% de los datos en ciertas variables son inferidos y fueron imputados mediante KNN.

- **Confusión entre clases patológicas**: El modelo muestra dificultad para discriminar entre "Estenosis" e "Insuficiencia". Esto es esperable dada la comorbilidad frecuente entre estas condiciones, pero limita su especificidad diferencial.
- **Validación céntrica única**: El modelo ha sido entrenado y validado con particiones de un mismo origen de datos. No se ha realizado una validación externa con datos de fuentes diferentes, lo cual es el "estándar de oro" para garantizar la portabilidad del sistema.

9.4. Líneas de trabajo futuras

Para evolucionar este TFG hacia un producto clínico desplegable, se proponen las siguientes líneas de investigación:

- **Validación externa**: Evaluar el modelo con un *dataset* totalmente nuevo para medir la degradación del rendimiento ante cambios de dominio (*domain shift*).
- **Aprendizaje profundo sobre señal cruda**: Entrenar una Red Neuronal Convolucional (1D-CNN) directamente sobre las series temporales del ECG y fusionar sus predicciones con el modelo tabular actual (*Ensamble* Híbrido).
- **Mejora de la interfaz Humano-Máquina**: Desarrollar un *dashboard* interactivo para el médico que no solo muestre la predicción, sino que permita simular escenarios ("¿Qué pasaría si la duración del QRS para el estudio actual fuera menor?") utilizando el motor de contrafactuales.
- **Estudio de equidad (*Fairness*)**: Realizar una auditoría exhaustiva de sesgos demográficos, evaluando tanto la calibración del modelo en cada subgrupo (*Brier Score*) como métricas de equidad (paridad demográfica, *equalized odds*) independientemente del sexo o etnia de la persona paciente, cumpliendo con los requisitos de alto riesgo del Reglamento Europeo de IA.
- **Extensión de la Predicción Conforme** mediante *Mondrian Conformal Prediction*: Definir conjuntos de predicción con garantías de cobertura condicionadas por clase o subgrupo clínico, lo que podría mejorar el control de la incertidumbre en escenarios con fuerte desbalance de clases o requisitos diferenciados de riesgo clínico.

9.5. Cumplimiento de estándares: Informe TRIPOD+AI

Con el objetivo de garantizar la transparencia y reproducibilidad de este modelo predictivo, se presenta a continuación un resumen (ver Tabla 23) de cumplimiento basado en la lista de verificación **TRIPOD+AI**.

Tabla 23 - Informe TRIPOD+AI

Sección TRIPOD	Elemento evaluado	Estado en el TFG	Justificación / Ubicación en memoria
Título y Resumen	Identificación como desarrollo de modelo	Cumplido	Título explícito y *Abstract* estructurado.
Antecedentes	Justificación médica y objetivos	Cumplido	Capítulos 1 y 2. Epidemiología cardiovascular.
Datos	Fuente, participantes y resultados	Cumplido	Capítulo 3 y 4. Dataset público documentado.
Predictores	Definición y preprocesamiento	Cumplido	Capítulo 4. Limpieza, Imputación K-NN.
Desarrollo del Modelo	Métodos de modelado e hiperparámetros	Cumplido	Capítulo 5. Selección de CatBoost, Optuna.
Evaluación	Métricas de rendimiento	Cumplido	Capítulo 8. Uso de F1-macro, AUC y *Brier Score*.
	Calibración e Incertidumbre	Cumplido	Capítulo 8. Inclusión de Curvas de Calibración y *Conformal Prediction*.
Resultados	Flujo de participantes	Parcial	Se describe el tamaño del dataset, pero no el flujo clínico original.
	Rendimiento del modelo	Cumplido	Tablas detalladas por clase y matrices de confusión.
Interpretación	Limitaciones y Generalización	Cumplido	Detallado en el presente Capítulo 9.3.
	Interpretación del modelo (XAI)	Cumplido	Capítulo 7. SHAP, Feature Importance.
Implicaciones	Uso clínico potencial	Cumplido	Discusión sobre triaje y soporte a la decisión.

10. Acrónimos

ANOVA: ***An****alysis* ***o****f* ***Va****riance* [91].

AUC-ROC: ***A****rea* ***U****nder the* ***R****eceiver* ***O****perating* ***C****haracteristic* ***C****urve* [92].

CatBoost: ***Cat****egorical* ***Boost****ing* [29].

CRISP-DM: ***CR****oss-****I****ndustry* ***S****tandard* ***P****rocess for* ***D****ata* ***M****ining* [20].

DWT: ***D****iscrete* ***W****avelet* ***T****ransform* [93].

ECG: **E**lectro**C**ardio**G**rama.

EDA: ***E****xploratory* ***D****ata* ***A****nalysis*.

FNR: ***F****alse* ***N****egative* ***R****ate*.

IA: **I**nteligencia **A**rtificial.

ICE: ***I****ndividual* ***C****onditional* ***E****xpectation*.

KNN: **K**-***N****earest* ***N****eighbors*.

KS: **K**olmogorov-**S**mirnov [94].

LightGBM: ***Light*** ***G****radient* ***B****oosting* ***M****achine* [27].

LIME: ***L****ocal* ***I****nterpretable* ***M****odel-agnostic* ***E****xplanations* [50].

PDP: ***P****artial* ***D****ependence* ***P****lots.*

PFI: ***P****ermutation* ***F****eature* ***I****mportance*.

RAPS: ***R****egularized* ***A****daptive* ***P****rediction* ***S****ets*.

RMSE: ***R****oot* ***M****ean* ***S****quare* ***E****rror*.

SHAP: ***SH****apley* ***A****dditive ex****P****lanations* [47].

TFG: **T**rabajo **F**in de **G**rado.

TPE: ***T****ree-structured* ***P****arzen* ***E****stimator*.

UQ: ***U****ncertainty* ***Q****uantification* [12].

VIF: ***V****ariance* ***I****nflation* ***F****actor***.**

XAI: *E****X****plainable* ***A****rtificial* ***I****ntelligence* [11].

XGBoost: *E****X****treme* ***G****radient* ***Boost****ing* [24].

11. Glosario

Calibración probabilística: Evaluación de si las probabilidades predichas por el modelo reflejan la realidad clínica, asegurando que la confianza del sistema corresponda con la frecuencia observada de la patología.

Contrafactuales: Explicaciones que indican qué cambios mínimos serían necesarios en las variables de entrada (por ejemplo, en el ECG) para que el modelo modificara su predicción original.

Derivaciones (ECG): Puntos de vista distintos desde los cuales se registra la actividad eléctrica cardíaca. Un ECG estándar consta de 12 derivaciones que miden diferencias de potencial en los planos frontal y horizontal.

EchoNext: *Dataset* utilizado en el estudio, compuesto por 100.000 registros multimodales que incluyen señales de ECG, variables demográficas y etiquetas ecocardiográficas.

Estenosis aórtica: Patología estructural caracterizada por la obstrucción al flujo sanguíneo, lo que produce una sobrecarga de presión crónica en el ventrículo izquierdo.

Explicabilidad Local: Análisis enfocado en entender decisiones individuales del modelo, validando clínicamente casos de éxito o diagnosticando fallos específicos mediante técnicas como LIME o SHAP local.

Hiperparámetros: Configuraciones del modelo (como la profundidad de un árbol o la tasa de aprendizaje) que se ajustan antes del entrenamiento para optimizar su rendimiento. En este trabajo se optimizan mediante Optuna.

Imputación KNN: Método avanzado de tratamiento de valores ausentes que estima el dato faltante basándose en las personas pacientes más similares (vecinos cercanos) en el espacio multivariante, preservando la estructura local de los datos.

Insuficiencia valvular: Patología que provoca una sobrecarga de volumen en las cámaras cardíacas (ventrículos o aurículas) debido a un cierre defectuoso de las válvulas.

Modelos de Ensamble: Algoritmos de aprendizaje supervisado que combinan múltiples modelos base para mejorar la generalización. Se han utilizado *Random Forest*, XGBoost, LightGBM y CatBoost.

Predicción Conforme (*Conformal Prediction*): Técnica que transforma las predicciones puntuales en conjuntos de predicción que garantizan contener la clase real con una probabilidad estadística predefinida (cobertura).

Valores SHAP: Método de explicabilidad basado en la teoría de juegos que asigna un valor de importancia a cada variable, considerando tanto su magnitud como sus interacciones con otras características.

12. Bibliografía

[1] INE, «Estadística de Defunciones según la Causa de Muerte», INE, 23 Junio 2025. [En línea]. Available: https://www.ine.es/dyngs/Prensa/pEDCM2024.htm.

[2] Organización Mundial de la Salud, «Las diez causas principales de defunción», Organización Mundial de la Salud, 7 Agosto 2024. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[3] H. a. Argente y M. E. Alvarez, Semiología médica: Fisiopatología, semiotecnia y propedéutica. Enseñanza basada en el paciente., 1ª edición 3ª reimpresión ed., Buenos Aires: Editorial médica panamericana, 2008.

[4] T. Cascino y M. J. Shea, «Electrocardiography», MSD Manuals, Abril 2025. [En línea]. Available: https://www.msdmanuals.com/professional/cardiovascular-disorders/cardiovascular-tests-and-procedures/electrocardiography.

[5] Y. E. Yoon, S. Kim y H.-J. Chang, «Artificial Intelligence and Echocardiography», *Journal of cardiovascular imaging,* vol. 29, nº 3, pp. 193-204, 2021.

[6] W. J. de Jong-Watt y H. M. Arthur, «Anxiety, depression and quality of life in patients waiting for coronary angiography», *Hearth & lung,* vol. 33, nº 4, pp. 237-248, 2004.

[7] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam y et al. , «Screening for asymptomatic left ventricular dysfunction using an artificial intelligence–enabled electrocardiogram», *Nature medicine,* vol. 25, pp. 70-74, 2019.

[8] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh y et al. , «An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction.», *The Lancet,* vol. 394, nº 10201, pp. 861-867, 2019.

[9] C. D. Galloway, A. V. Valys, J. B. Shreibati, D. L. Treiman, F. L. Petterson, V. P. Gundotra y et al. , «Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram.», *JAMA Cardiology,* vol. 4, nº 5, pp. 428-436, 2019.

[10] Comisión Europea, «Ley de IA», Comisión Europea, 8 Octubre 2025. [En línea]. Available: https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai. [Último acceso: 19 Octubre 2025].

[11] C. Molnar, «Interpretable Machine Learning», 2025. [En línea]. Available: https://christophm.github.io/interpretable-ml-book/#about-the-book.

[12] T. Wang, Y. Wang, J. Zhou, B. Peng, X. Song, C. Zhang y et al. , «From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence», *arXiv,* p. 14, 2025.

[13] P. Elias, T. J. Poterucha, V. Rajaram, L. Matos Moller, V. Rodriguez, S. Bhave y et al. , «Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease.», *The Journal of the American College of Cardiology,* vol. 80, nº 6, p. 613–26, 2022.

[14] A. E. Ulloa-Cerna, L. Jing, J. M. Pfeifer, S. Raghunath, J. A. Ruhl, D. B. Rocha y et al. , «rECHOmmend: An ECG-Based Machine Learning Approach for Identifying Patients at Increased Risk of Undiagnosed Structural Heart Disease Detectable by Echocardiography», *Circulation,* vol. 146, nº 1, p. 36–47, 2022.

[15] K. C. Siontis, P. A. Noseworthy, Z. I. Attia y P. A. Friedman, «Artificial intelligence-enhanced electrocardiography in cardiovascular disease management», *Nature Reviews Cardiology,* vol. 18, nº 7, p. 465–478, 2021.

[16] T. J. Poterucha, L. Jing, R. Pimentel Ricart, M. Adjei-Mosi, J. Finer, D. Hartzel y et al. , «Detecting structural heart disease from electrocardiograms using AI», *Nature,* vol. 644, p. 221–230, 2025.

[17] P. Elias y J. Finer, «EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs», 2025. [En línea]. Available: https://doi.org/10.13026/3ykd-bf14.

[18] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart y et al. , «Automatic diagnosis of the 12-lead ECG using», *Nature comunicactions,* vol. 11, p. 1760, 9 Abril 2020.

[19] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster y et al, «TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods», BMJ, 2024. [En línea]. Available: https://www.bmj.com/content/385/bmj-2023-078378.

[20] C. Shearer, «The New Blueprint for Data Mining», *Journal of Data Warehousing,* vol. 5, pp. 13-22, 2000.

[21] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Les, C. Schölzel y S. H. A. Chen, «NeuroKit2: A Python toolbox for neurophysiological signal processing», *Behavior Research Method,* vol. 53, pp. 1689-1696, 2021.

[22] M. G. Frasch, «Comprehensive HRV estimation pipeline in Python using Neurokit2: Application to sleep physiology», *MethodsX,,* vol. 9, p. 101782, 2022.

[23] L. Breiman, «Random Forests», *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[24] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* p. 785–794, 2016.

[25] xgboost developers, «XGBoost Documentation», 2025. [En línea]. Available: https://xgboost.readthedocs.io/en/stable/.

[26] G. Ke, Q. Meng, T. Finley, T. Wang, W. Che, W. Ma, Q. Ye y T.-Y. Liu, «LightGBM: a highly efficient gradient boosting decision tree», de *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 2017.

[27] Microsoft, «Welcome to LightGBM's documentation!», LightGBM, 2025. [En línea]. Available: https://lightgbm.readthedocs.io/en/latest/.

[28] L. Prokhorenkova, G. Gusev, A. Voro, A. V. Dorogush y A. Gulin, «CatBoost: unbiased boosting with categorical features», 2019.

[29] A. Gulin, «CatBoost», [En línea]. Available: https://catboost.ai/docs/en/. [Último acceso: 2025].

[30] K. Budholiya, S. K. Shrivastava y V. Sharma, «An optimized XGBoost based diagnostic system for effective prediction of heart disease», *Journal of King Saud University - Computer and Information Sciences,* vol. 34, nº 7, pp. 4514-4523, 2022.

[31] K. Cao, C. Liu, S. Yang, Y. Zhang, L. Li, H. Jung y S. Zhang, «Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model», *Scientific Reports,* vol. 15, nº 12406, 2025.

[32] C. Bentéjac, A. Csörgő y G. Martínez-Muñoz, «A comparative analysis of gradient boosting algorithms», *Artificial Intelligence Review,* vol. 54, pp. 937–1967,, 2021.

[33] R. Shwartz-Ziv y A. Armon, «TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED», *arXiv,* 2021.

[34] L. Grinsztajn , E. Oyallon y G. Varoquaux, «Why do tree-based models still outperform deep learning on tabular data?», *arXiv,* 2022.

[35] A. Mignan y M. Broccardo, «One neuron versus deep learning in aftershock prediction», *Nature,* vol. 574, p. E1–E3, 2019.

[36] D. McElfresh, S. Khandagale, J. Valverde, C. Vishak Prasad, G. Ramakrishnan, M. Goldblum y et al. , «When do neural nets outperform boosted trees on tabular data?», de *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, 2023.

[37] M. Moreno de Castro, «Más alla de la métrica», Septiembre 2025. [En línea]. Available: https://github.com/MMdeCastro/Uncertainty_Quantification_XAI/tree/main.

[38] M. Grandini, E. Bagli y G. Visani, «METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW», arXiv, 2020.

[39] J. M. Lobo, A. Jiménez-Valverde y R. Real, «AUC: a misleading measure of the performance of predictive distribution models», *Global Ecology and Biogeography,* vol. 17, nº 2, pp. 145-151, 2008.

[40] G. W. BRIER, «VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY», *MONTHLY WEATHER REVIEW,* vol. 78, nº 1, pp. 1-3, 1950.

[41] A. Niculescu-Mizil y R. Caruana, «Predicting good probabilities with supervised learning», de *ICML '05: Proceedings of the 22nd international conference on Machine learning*, New York, 2005.

[42] A. N. Angelopoulos y S. Bates, «A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification», arXiv, 7 Diciembre 2022. [En línea]. Available: https://arxiv.org/abs/2107.07511.

[43] J. Vazquez y J. C. Facelli, «Conformal Prediction in Clinical Medical Sciences», *Journal of Healthcare Informatics Research,* vol. 6, p. 241–252, 2022.

[44] IBM, «What is explainable AI?», IBM, 2025. [En línea]. Available: https://www.ibm.com/think/topics/explainable-ai#:~:text=La%20inteligencia%20artificial%20explicable%20(XAI,los%20algoritmos%20de%20machine%20learning.. [Último acceso: 18 Noviembre 2025].

[45] A. Fisher , C. Rudin y F. Dominic, «All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously», *Journal of machine learning research,* vol. 20, p. 20:177, 2019.

[46] T. A. Assegie, «Evaluation of Local Interpretable Model-Agnostic Explanation and Shapley Additive Explanation for Chronic Heart Disease Detection», *Proceedings of Engineering and Technology Inovation,* vol. 23, p. 48–59, 2023.

[47] S. Lundberg, «Welcome to the SHAP documentation», 2018. [En línea]. Available: https://shap.readthedocs.io/en/latest/.

[48] S.-I. Lee, «shap», Github, 2025. [En línea]. Available: https://github.com/shap/shap?tab=readme-ov-file.

[49] M. T. Ribeiro, S. Singh y C. Guestrin, «"Why Should I Trust You?": Explaining the Predictions of Any Classifier», *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1135-1144, 2016.

[50] M. T. Ribeiro, «Lime», Github, 2020. [En línea]. Available: https://github.com/marcotcr/lime.

[51] S. Wachter, B. Mittelstadt y C. Russ, «COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR», *Harvard Journal of Law & Technology,* vol. 31, nº 2, pp. 842-861, 2018.

[52] T. L. Beauchamp y J. F. Childress, Principles of Biomedical Ethics. 7th Edition, Oxford University Press, 2013.

[53] A. B. Popejoy y S. M. Fullerton , «Genomics is failing on diversity», *Nature,* vol. 538, p. 161–164, 2016.

[54] D. A. Vyas , L. G. Eisenstein y D. S. Jon, «Reconsidering the Use of Race Correction in Clinical Algorithms», *The New England Journey of Medicine,* vol. 383, nº 9, pp. 874-882, 2020.

[55] Z. Obermeyer , B. Powers, C. Vogeli y S. Mullainathan, «Dissecting racial bias in an algorithm used to manage the health of populations», *Science,* vol. 366, nº 6464, pp. 447-453, 2019.

[56] S. Barocas, M. Hardt y A. Narayanan, Fairness and Machine Learning Limitations and Opportunities, The MIT Press, 2023.

[57] R. Schwartz, J. Dodge, N. A. Smith y O. Etzioni, «Green AI», *Communications of the ACM,* vol. 63, nº 12, pp. 54-63, 2020.

[58] A. Salam y A. E. Hibaoui, «Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city –», de *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, Rabat, 2018.

[59] K. Gardiner , K. Hanneman y R. Kozor , «The environmental effects of non-invasive cardiac imaging.», *American Heart Journal Plus: Cardiology Research and Practice,* vol. 46, 2024.

[60] Fundación Española del Corazón, «Ficha del paciente: Electrocardiograma», Fundación Española del Corazón, 17 Enero 2023. [En línea]. Available: https://fundaciondelcorazon.com/phocadownload/recursos-didacticos/fichas/Electrocardiograma.pdf.

[61] J. Pérez-Lescure Picarzo y O. Patiño Hernández, «El electrocardiograma», *Formación Activa en Pediatría de Atención Primaria,* vol. 4, nº 1, 2011.

[62] Consellería de Sanidad, «PROTOCOLO DE ACTUACIÓN PARA LA MEJORA», Comunitat Valenciana, 2024. [En línea]. Available: https://www.san.gva.es/documents/d/assistencia-sanitaria/protocolo_deteccion_diag_precoz_ic_ap_definit_-v8-docx.

[63] Salusplay, «TEMA 2. EL ELECTROCARDIOGRAMA», Salusplay, 2025. [En línea]. Available: https://www.salusplay.com/apuntes/cuidados-medico-quirurgicos/tema-2-el-electrocardiograma.

[64] E. Plaza Montero, «Derivaciones de los miembros», Urgencias y Emergencias, 17 Diciembre 2017. [En línea]. Available: https://www.urgenciasyemergen.com/las-derivaciones-del-electrocardiograma/.

[65] S. H. Pujari y P. Agasthi, «Aortic Stenosis», 16 Abril 2023. [En línea]. Available: https://www.ncbi.nlm.nih.gov/books/NBK557628/.

[66] A. F. Esteves, D. Brito, J. Rigueira, I. Ricardo, R. Pires, M. Mendes Pedroa y et al. , «Profiles of hospitalized patients with valvular heart disease: Experience of a tertiary center», *Revista Portuguesa de Cardiologia,* vol. 37, nº 12, pp. 991-998, 2018.

[67] V. Vovk , A. Gammerman y G. Shafer, Algorithmic Learning in a Random World, Springer Nature, 2022.

[68] Quantmetry, «Theoretical Description Classification : contents», Mapie, 2022. [En línea]. Available: https://mapie.readthedocs.io/en/latest/theoretical_description_classification.html#regularized-adaptive-prediction-sets-raps.

[69] J. P. Birkbeck, D. B. Wilson, M. A. Hall y D. G. Meyers, «P-wave morphology correlation with left atrial volumes assessed by 2-dimensional echocardiograph», *Journal of Electrocardiology,* vol. 39, nº 2, pp. 225-229, 2006.

[70] B. Brembilla-Perrot, «Study of P wave morphology in lead V1 during supraventricular tachycardia for localizing the reentrant circuit», *Am Heart J.,* vol. 6, nº 1, pp. 1714-20, 1991.

[71] I. Morrison, E. Clark y P. W. Macfarlane, «Evaluation of the electrocardiographic criteria», *The anatolian journey of cardiology,* vol. 7, nº 1, pp. 159-63, 2007.

[72] L. Sörnmo y P. Laguna, Bioelectrical Signal Processing in Cardiac and Neurological Applications, Academic Press, 2005.

[73] H. Aguinis, R. Gottfredson y H. Joo, «Best-Practice Recommendations for Defining, Identifying, and Handling Outliers», *Organizational Research Methods,* vol. 16, nº 2, pp. 270-301, 2013.

[74] G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning, Springer, 2021.

[75] L. Liu, X. Wu, S. Li, Y. Li, S. Tan y Y. Bai, «Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection», *BMC Medical Informatics and Decision Making,* vol. 22, nº 82, 2022.

[76] T. Saito y M. Rehmsmeier , «The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets», *PLOS ONE,* 2015.

[77] T. Fawcett, «An introduction to ROC analysis», *Elsevier,* vol. 27, nº 8, pp. 861-874, 2006.

[78] C. Guo, G. Pleiss, Y. Sun y K. Q. Weinber, «On Calibration of Modern Neural Networks», *34th International Conference on Machine Learning,* vol. 70, pp. 1321-1330, 2017.

[79] T. Akiba, S. Sano, T. Yanase, T. Ohta y M. Koyama, «Optuna: A Next-generation Hyperparameter Optimization Framework», *arXiv,* nº arXiv:1907.10902, p. 10, 2016.

[80] J. Li , Q. Meng , J. Gu, C. Ji y X. Pan , «Changing trends in burden of calcific aortic valve diseases and degenerative mitral valve diseases among people aged 60-89 years from 1990 to 2030», *European Journal of Medical Research ,* 2025.

[81] C. Strobl, A.-L. Boulesteix, A. Zeile y T. Hothorn , «Bias in random forest variable importance measures: Illustrations, sources and a solution», *BMC Bioinformatics,* vol. 8, nº 25, p. 21, 2007.

[82] S. Lundberg y S.-I. Lee, «A Unified Approach to Interpreting Model», de *31st Conference on Neural Information Processing Systems* , Long Beach, 2017.

[83] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan y et al. , «Fairlearn: A toolkit for assessing and», 22 Septiembre 2020. [En línea]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf.

[84] M. Hardt, E. Price y N. Srebro, «Equality of Opportunity in Supervised Learning», *arXiv,* pp. 1-22, 2016.

[85] S. Verma y J. Rubin, «Fairness definitions explained», de *FairWare '18: Proceedings of the International Workshop on Software Fairness*, Gothenburg, 2018.

[86] R. K. Mothilal, A. Sharma y C. Tan, «Explaining machine learning classifiers through diverse counterfactual explanations», de *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Montreal, 2020.

[87] M. Hardt, E. Price y N. Srebro, «Equality of Opportunity in Supervised Learning», 2016.

[88] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman y A. Galstyan, «A Survey on Bias and Fairness in Machine Learning», *ACM Computing Surveys (CSUR),* vol. 54, nº 6, pp. 1 - 35, 2021.

[89] I. Y. Chen, P. Szolovits y M. Ghassemi, «Can AI Help Reduce Disparities in General Medical and Mental Health Care?», *AMA Journal of Ethics,* vol. 21, nº 2, pp. 167-179, 2019.

[90] Z. Obermeyer , B. Powers, C. Vogeli y . S. Mullai, «Dissecting racial bias in an algorithm used to manage the health of populations», *Science,* vol. 366, nº 6464, pp. 447-453, 2019.

[91] colaboradores de Wikipedia, «Análisis de la varianza», Wikipedia, La enciclopedia libre, 23 Diciembre 2024. [En línea]. Available: https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_la_varianza&oldid=164279017.

[92] Google Developers, «Classification: ROC and AUC», developers.google.com, 2025. [En línea]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[93] W. contributors, «Discrete wavelet transform», Wikipedia, The Free Encyclopedia, 28 Octubre 2025. [En línea]. Available: https://en.wikipedia.org/w/index.php?title=Discrete_wavelet_transform&oldid=1319136491.

[94] colaboradores de Wikipedia, «Prueba de Kolmogórov-Smirnov», Wikipedia, La enciclopedia libre, 22 Octubre 2025. [En línea]. Available: https://es.wikipedia.org/w/index.php?title=Prueba_de_Kolmog%C3%B3rov-Smirnov&oldid=170114500.