UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

ESCUELA POLITÉCNICA SUPERIOR DE ELCHE

GRADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

"APLICACIÓN DE TÉCNICAS DE APRENDIZAJE ESTADÍSTICO PARA LA DETECCIÓN Y CLASIFICACIÓN DE LA MIOCARDIOPATÍA HIPERTRÓFICA DE ORIGEN HEREDITARIO."

TRABAJO FIN DE GRADO

Septiembre - 2023

Autor: Miguel Luis Alonso

Director: Francisco Javier Gimeno Blanes

Codirector: Antonio Casañez Ventura

Índice

1 Abstract

Este Trabajo de Fin de Grado (TFG) se centra en la aplicación de algoritmos de aprendizaje automático y aprendizaje profundo para abordar la detección de la Enfermedad de Cardiomiopatía Hipertrófica (HCM), una afección cardíaca hereditaria. La investigación se basa en una base de datos matricial que incluye información clínica y médica de pacientes con y sin HCM.

En este estudio, se explora la capacidad de diversos algoritmos de machine learning y técnicas de deep learning para analizar la base de datos y encontrar una separación efectiva entre los individuos afectados por la HCM y aquellos que no lo están. Se investigan enfoques de preprocesamiento de datos, selección de características y ajuste de hiperparámetros para optimizar el rendimiento de los modelos. Además, se evalúan métricas de evaluación de rendimiento, como precisión, sensibilidad y especificidad, para determinar la eficacia de los modelos propuestos.

Los resultados obtenidos sugieren que los algoritmos de machine learning y deep learning pueden desempeñar un papel crucial en la detección temprana y precisa de la HCM a partir de datos matriciales, lo que podría tener un impacto significativo en el diagnóstico y tratamiento de esta enfermedad cardíaca. Este TFG contribuye al campo de la medicina computacional y resalta el potencial de las técnicas de aprendizaje automático en la mejora de la atención médica y la identificación de enfermedades hereditarias.

2 Lista de abreviaturas

ML: Machine Learning

DL: Deep Learning

HCM: Miocardiopatía Hipertrófica

ECG: Electrocardiograma

IA: Inteligencia Artificial

PCA: Principal component análisis

SVM: Support vector machine

SVDD: Support vector data description

UMAP: Uniform Manifold Approximation and Projection

MCAR: Missing Completely at Random

MSC: Muerte súbita cardiaca

ECO: Ecocardiograma

t-SNE: t-Distributed Stochastic Neighbor Embedding

AE: Auto Encoder

3 Introducción

3.1 Motivación

Existen diversos motivos para la realización de este trabajo de fin de grado:

Necesidad Médica y de Salud Pública: La HCM es una enfermedad cardíaca grave que puede ser hereditaria y difícil de diagnosticar tempranamente. La motivación puede surgir de la necesidad de desarrollar herramientas más efectivas para identificar y diagnosticar esta afección, lo que puede llevar a una atención médica más temprana y mejores resultados para los pacientes.

Avances en la Ciencia de Datos: El interés por aplicar técnicas de machine Learning y Deep Learning en el campo médico está en constante crecimiento debido a los avances tecnológicos y la disponibilidad de grandes conjuntos de datos. La motivación podría ser aprovechar estas tecnologías para mejorar el diagnóstico médico.

Contribución a la Investigación Médica: El trabajo puede tener como objetivo contribuir al cuerpo de conocimientos en medicina computacional y análisis de datos en salud. Puede ser motivado por el deseo de desarrollar modelos efectivos que los profesionales de la salud puedan utilizar para tomar decisiones más informadas.

Impacto en la Práctica Clínica: La capacidad de detectar la HCM de manera temprana y precisa puede tener un impacto significativo en la práctica clínica, mejorando la calidad de atención y el tratamiento de los pacientes. Esta mejora en la atención médica podría ser una fuente de motivación.

Interés en la Intersección de la Medicina y la Tecnología: La motivación podría provenir de un interés personal o profesional en la intersección de la medicina y la tecnología. Los investigadores pueden estar motivados por la posibilidad de aplicar técnicas de vanguardia para abordar problemas médicos complejos.

3.2 Antecedentes de la literatura

La miocardiopatía hipertrófica (HCM o HOCM cuando es obstructiva) es una condición en la que el corazón se engrosa sin una causa obvia. Las partes del corazón más comúnmente afectadas son el tabique interventricular y los ventrículos. Esto da como resultado que el corazón sea menos capaz de bombear sangre de manera efectiva y también puede causar problemas de conducción eléctrica.

Las personas que tienen HCM pueden tener una variedad de síntomas. Las personas pueden ser asintomáticas o pueden tener fatiga, hinchazón de las piernas y dificultad para respirar. También puede resultar en dolor de pecho o desmayo. Los síntomas pueden empeorar cuando la persona está deshidratada. Las complicaciones pueden incluir insuficiencia cardíaca, latidos cardíacos irregulares y muerte cardíaca súbita.

La HCM se hereda más comúnmente con un patrón autosómico dominante. A menudo se debe a mutaciones en ciertos genes involucrados en la producción de proteínas del músculo cardíaco. Otras causas hereditarias de hipertrofia ventricular izquierda pueden incluir la enfermedad de Fabry, la ataxia de Friedreich y ciertos medicamentos como el tacrolimus. Otras consideraciones para las causas del agrandamiento del corazón son el corazón de atleta y la hipertensión (presión arterial alta). Hacer el diagnóstico de HCM a menudo implica antecedentes familiares, un electrocardiograma, un ecocardiograma y una prueba de esfuerzo. También se pueden realizar pruebas genéticas.

El diagnóstico de miocardiopatía hipertrófica se basa en una serie de características del proceso de la enfermedad. Si bien se utiliza la ecocardiografía, el cateterismo cardíaco o la resonancia magnética cardíaca en el diagnóstico de la enfermedad, otras consideraciones importantes incluyen el ECG, las pruebas genéticas (aunque no se utilizan principalmente para el diagnóstico) y cualquier antecedente familiar de MCH o muerte súbita inexplicable en personas sanas. individuos En alrededor del 60 al 70% de los casos, la resonancia magnética cardíaca muestra un engrosamiento de más de 15 mm de la parte inferior del tabique ventricular.

3.2.1 Machine learning

Hoy en día, los sistemas inteligentes que ofrecen capacidades de inteligencia artificial a menudo se basan en el aprendizaje automático. El aprendizaje automático describe la capacidad de los sistemas para aprender de los datos de entrenamiento específicos del problema para automatizar el proceso de creación de modelos analíticos y resolver las tareas asociadas.[1]

El aprendizaje automático (ML) busca aprender automáticamente relaciones y patrones significativos a partir de ejemplos y observaciones. Los avances en ML han permitido el reciente surgimiento de sistemas inteligentes con capacidad cognitiva similar a la humana que penetran nuestra vida comercial y personal y dan forma a las interacciones en red en los mercados electrónicos de todas las formas imaginables, con empresas que aumentan la toma de decisiones para la productividad, el compromiso y la retención de empleados. Sistemas asistentes que se pueden entrenar y que se adaptan a las preferencias individuales de los usuarios. La capacidad de tales sistemas para la resolución avanzada de problemas, generalmente denominada inteligencia artificial (IA), se basa en modelos analíticos que generan predicciones, reglas, respuestas, recomendaciones o resultados similares. Los primeros intentos de construir modelos analíticos se basaron en la programación explícita de relaciones, procedimientos y lógica de decisión conocidos en sistemas inteligentes a través de reglas hechas a mano (por ejemplo, sistemas expertos para diagnósticos médicos) (Russell y Norvig 2021). Impulsados por la viabilidad de los nuevos marcos de programación, la disponibilidad de datos y el amplio acceso a la potencia informática necesaria, los modelos analíticos se construyen hoy en día cada vez más utilizando lo que generalmente se conoce como ML. [1]

3.2.2 Deep Learning

El aprendizaje profundo (DL) es un concepto de aprendizaje automático basado en redes neuronales artificiales. Para muchas aplicaciones, los modelos de aprendizaje profundo superan a los modelos de aprendizaje automático (ML) poco profundos y los enfoques de análisis de datos tradicionales. [1]

Recientemente, las técnicas de aprendizaje profundo han surgido como métodos poderosos para aprender representaciones de características automáticamente a partir de datos. En particular, estas técnicas han proporcionado importantes mejoras en la detección de objetos. La detección de objetos es una parte importante del diagnóstico de la miocardiopatía hipertrófica, ya que gracias a la detección de objetos es posible diagnosticar esta enfermedad si parte del tejido del corazón está hipertrofiado. [2]

3.2.3 Fisiología del corazón

El corazón es un órgano importante del cuerpo humano cuya función es bombear oxígeno y nutrientes a todas las partes del cuerpo mediante contracciones rítmicas repetidas. Consta de cuatro cámaras: dos en la parte superior, las aurículas (izquierda y derecha), y dos en la parte inferior, los ventrículos izquierdo y derecho (ver Ilustración 1).

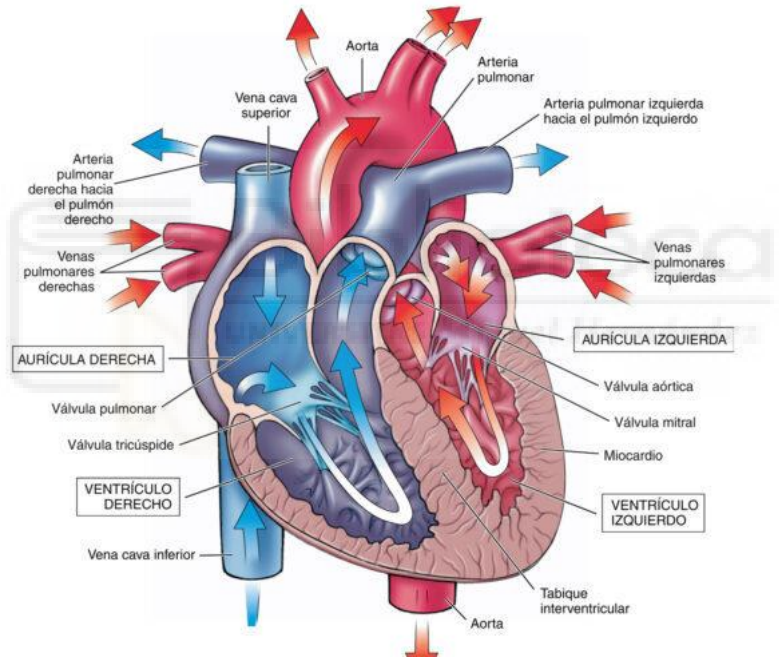

Ilustración 1 Estructura del corazón

La sangre desoxigenada regresa al corazón desde el resto del cuerpo a través de la vena cava superior (VSC) y la vena cava inferior (VCI), que llevan la sangre de regreso al corazón. La sangre desoxigenada ingresa a la aurícula derecha (RA) y la sangre fluye a través de la válvula tricúspide (VT) hacia el ventrículo derecho (VD), que es responsable de bombear sangre desoxigenada a través de la válvula pulmonar (VP) hacia la arteria pulmonar principal (PPA). Desde allí, la sangre es transportada a los pulmones a través de las arterias pulmonares derecha e izquierda, donde recibe oxígeno. La sangre oxigenada llega a la aurícula izquierda (LA) a través de las cuatro venas pulmonares. Luego, la sangre oxigenada fluye a través de la válvula mitral (VM) hacia el ventrículo izquierdo (VI). El ventrículo izquierdo (VI) bombea sangre oxigenada a través de la válvula aórtica (AoV) hacia la aorta (Ao), que es la arteria principal que transporta sangre oxigenada al resto del cuerpo.

Este ciclo rítmico del corazón, llamado frecuencia cardíaca, se repite entre 60 y 100 veces por minuto y varía según las necesidades del cuerpo. La frecuencia cardíaca varía según la actividad física de una persona. Actividades como el ejercicio o el estrés hacen que este ritmo aumente y disminuya por la noche o bajo la influencia de ciertos medicamentos [3].

Cada latido del corazón es estimulado por señales eléctricas que pasan a través del músculo cardíaco (o miocardio). Las señales eléctricas siguen un camino específico a través del corazón, comenzando en el nódulo sinoauricular (SA), ubicado en el ventrículo superior derecho (o aurícula). Luego, la señal se ramifica a través de las aurículas izquierda y derecha, que se contraen y fuerzan la sangre hacia los ventrículos. Al mismo tiempo, la señal eléctrica a los ventrículos se conduce a través del nódulo auriculoventricular (AV), que se encuentra en la pared entre la aurícula y el ventrículo derechos, y viaja lentamente con un retraso de 0,1 segundos. Finalmente, la señal viaja a través del haz de His y a lo largo de sus ramas a través de las fibras de Purkinje, haciendo que los ventrículos se contraigan y bombeen sangre a los pulmones y al cuerpo. [4]

Durante los latidos del corazón, las aurículas y los ventrículos pasan por dos fases: sístole (sístole), que corresponde al momento en que recibe estimulación eléctrica de los nódulos SA y AV, respectivamente, y empuja la sangre hacia los ventrículos, el cuerpo, respectivamente. y los pulmones, y un período de relajación (diástole) cuando el corazón se llena de sangre. La fase de contracción es causada por ondas de despolarización cargadas positivamente desde las uniones SA y AV y comienza cuando las fibras musculares están cargadas (en un estado de reposo eléctricamente neutro). Los nódulos SA y AV están despolarizados por iones Ca2 y la forma de onda en estos puntos es más larga, lo que explica el retraso de 0,1 s entre la contracción auricular y ventricular mencionado anteriormente. A esto le sigue la repolarización (causada por iones K), que conduce a la relajación del miocardio.

3.2.3.1 Membrana cardiaca

En reposo, existe un equilibrio entre las cargas eléctricas dentro y fuera de las células del corazón. Dentro de la célula, el K+ es el catión más importante; El anión es principalmente fosfato y la base conjugada de un ácido orgánico, mientras que fuera de la célula el Na+ y el Cl- son los más importantes. Esta condición da como resultado una diferencia de potencial de membrana de -85 a 95 mV.

Todas las células cardíacas, musculares y de los tejidos conductores disparan periódicamente potenciales de acción para establecer el ritmo cardíaco mediante la despolarización espontánea, que implica tres cambios en la conductancia de los iones Na+, Ca2+ y K+ [5]. El nodo SA, también conocido como "marcapasos", es el único nodo que genera impulsos (60-70 veces por minuto) en condiciones normales. Si no funciona, el nodo AV produce una frecuencia cardíaca de 40 a 60 latidos por minuto, pero más lenta. Sus células musculares y ventriculares producen de 30 a 40 y de 20 a 30 latidos por minuto, respectivamente [5].

La membrana de las células del corazón consta de bombas y canales de intercambio iónico. La bomba es selectiva al discriminar entre el tamaño y la carga de los iones, lo que permite que los iones se muevan a través del canal, creando un flujo de iones según su gradiente electroquímico. Dependiendo de la diferencia de potencial se pueden configurar 3 estados: abierto, cerrado e inactivo. Las bombas permiten el movimiento opuesto por un intercambio.

Utilizando estos canales y bombas, obtenemos las siguientes 5 fases como se muestra en la Ilustración 2:

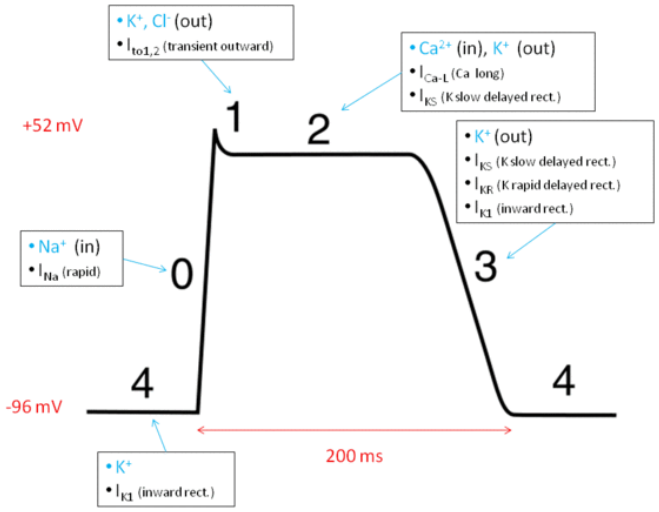

Ilustración 2 Ciclo potencial de acción de la membrana

Etapa 0, despolarización rápida: la apertura de los canales de Na+ produce corrientes internas de muy alta amplitud que provocan la despolarización. El potencial de membrana en reposo es crucial para esta fase: si el potencial es lo suficientemente bajo, todos los canales rápidos de Na+ se cerrarán y habrá una gran afluencia después de la excitación (Vmax tenderá a E[Na+] = 60 mV), pero si el potencial inicial es mayor, es más probable que el canal esté inactivo, lo que resulta en una menor respuesta a los estímulos (lo que resulta en una Vmax más baja, lo que aumenta el riesgo de arritmias).

Etapa 1, Repolarización inicial: Los canales Na+ ahora están inactivos y los canales K+ se abren causando una pequeña desviación por debajo del potencial de acción.

Etapa 2, Meseta: El estado se mantiene con un equilibrio de flujos de Ca2+ hacia afuera y K+. También los intercambiadores Na+/Ca2+ y las bombas Na+/K+ están interfiriendo, pero en un nivel inferior.

Etapa 3, Repolarización rápida: Durante esta fase, los canales Ca2+ se cierran mientras que los canales K+ permanecen abiertos, asegurando un cambio negativo en el potencial de la membrana, y como resultado la apertura de más K+ canales.

Etapa 4, Descanso: Unos pocos canales K+ permanecen abiertos para restaurar el potencial de la membrana en reposo y esperar una nueva despolarización.

3.2.4 Ecocardiograma

La ecocardiografía es una técnica de diagnóstico médico que utiliza ondas de ultrasonido para obtener imágenes en tiempo real del corazón y las estructuras circundantes. Estas imágenes proporcionan información importante sobre la anatomía, la función y el flujo sanguíneo del corazón. La ecocardiografía es una herramienta esencial en cardiología utilizada para evaluar diversas enfermedades cardíacas. [6]

El principio de un ecocardiograma se basa en la transmisión y recepción de ondas sonoras de alta frecuencia (ultrasonido). Estas ondas sonoras son generadas por un transductor, un dispositivo

colocado en la superficie de la piel del paciente. El transductor emite una serie de pulsos de ultrasonido que penetran en el tejido mamario y rebotan en varias estructuras del corazón. Las ondas sonoras reflejadas son captadas por transductores y convertidas en imágenes en la pantalla de una computadora en tiempo real. Estas imágenes, llamadas ecocardiograma, muestran cómo se mueve el corazón, lo que permite a los médicos evaluar su función y detectar cualquier anomalía. La velocidad y la fuerza de la reflexión de las ondas sonoras se utilizan para calcular la distancia entre el transductor y la estructura del corazón, lo que da como resultado imágenes detalladas del corazón y sus cámaras.

3.2.4.1 Información proporcionada por un ecocardiograma

La ecocardiografía es una herramienta de diagnóstico extremadamente versátil y valiosa que proporciona información completa sobre la anatomía y función cardíaca. Los siguientes son los principales aspectos de la información que proporciona la ecocardiografía: [6]

Anatomía del corazón: un ecocardiograma puede mostrar la anatomía del corazón en detalle, incluidas las cuatro cámaras del corazón (aurículas y ventrículos), las válvulas cardíacas (válvulas mitral, aórtica, tricúspide y pulmonar), las paredes musculares del corazón y el pericardio (membrana). alrededor del corazón). Esto es esencial para identificar anomalías, defectos de nacimiento y problemas estructurales.

Función cardíaca: un ecocardiograma proporciona una evaluación detallada de la función cardíaca, incluida la capacidad de contraerse y relajarse. Se mide por la fracción de eyección, que indica la proporción de sangre expulsada del ventrículo izquierdo con cada latido del corazón. También se evalúa la contractilidad de la pared del corazón y la presencia de hipertrofia ventricular (engrosamiento de las paredes de los ventrículos).

Flujo sanguíneo: la ecocardiografía Doppler puede evaluar el flujo sanguíneo a través de las válvulas del corazón y las arterias y venas principales del corazón. Esto es esencial para detectar obstrucciones, insuficiencia valvular y otras anomalías del flujo sanguíneo.

Presión arterial: la ecocardiografía también se utiliza para evaluar la presión arterial ventricular, que es útil para evaluar la hipertensión pulmonar o la hipertensión sistémica.

Diagnóstico de enfermedades cardíacas: La ecocardiografía es muy importante en el diagnóstico y evaluación de varias enfermedades cardíacas como valvulopatías, miocardiopatía, pericarditis, endocarditis, trombosis y cardiopatías congénitas.

Seguimiento y orientación del tratamiento: La ecocardiografía es fundamental en el seguimiento de los pacientes con enfermedades cardíacas crónicas. También gestionan procedimientos invasivos como la colocación de stent, cirugía cardíaca, reparación de válvulas y colocación de marcapasos.

Evaluación de la respuesta al tratamiento: la ecocardiografía se utiliza durante el seguimiento de pacientes con enfermedades cardíacas para evaluar su respuesta al tratamiento y si su función cardíaca ha mejorado.

3.2.4.2 Variables extraídas de un ecocardiograma

En este apartado se nombran un amplio número de variables que pueden ser obtenidas a partir de un ecocardiograma y que han sido utilizados en los procedimientos de Machine Learning y Deep Learning de este trabajo de fin de grado. Para el resumen rápido de un campo amplio como es el caso de los ecocardiogramas se ha optado por resumir la información que puede llegar a proporcionar cada una de estas variables. Por lo tanto, para cada variable se para que campo de la salud es interesante esta medida y una descripción general junto con lo que implica valores que se alejen de la normalidad.

El Diámetro Telediastólico del Ventrículo Izquierdo (DTDVI) es una medida esencial que indica el tamaño máximo del ventrículo izquierdo del corazón durante la fase de llenado completo, llamada diástole. Se obtiene en milímetros y se toma cuando las válvulas cardíacas están completamente abiertas y el ventrículo izquierdo se encuentra lleno de sangre. Esta medición es crucial para evaluar la función y el tamaño del ventrículo izquierdo, lo que proporciona información clave sobre la salud cardíaca. Un DTDVI anormalmente grande puede indicar problemas como insuficiencia cardíaca o hipertensión, mientras que uno pequeño podría señalar una disminución en la función de llenado del corazón. El DTDVI se utiliza junto con otros datos ecocardiográficos para diagnosticar y guiar el tratamiento de afecciones cardíacas.

Volumen telediastólico de ventrículo izquierdo: Esta es una medida básica que representa el volumen máximo de sangre que el ventrículo izquierdo del corazón puede contener antes de contraerse en su fase de llenado total (llamada diástole). El valor se obtiene mediante ecografía y se expresa en mililitros (ml). VTDVI es muy importante para evaluar la capacidad del ventrículo izquierdo para recibir y retener sangre antes de bombearla al resto del cuerpo.

Un VTDVI anormalmente alto puede indicar agrandamiento del ventrículo izquierdo, que a menudo se asocia con problemas cardíacos como insuficiencia cardíaca o enfermedad valvular. Por otro lado, un VTDVI anormalmente bajo puede indicar una función de llenado cardíaco alterada. Cuando se utiliza junto con otros parámetros ecocardiográficos, VTDVI proporciona una evaluación integral de la función y la salud cardíacas, y es fundamental para el diagnóstico y tratamiento de enfermedades cardíacas.

El Septo telediastólico del VI es una medida importante que se obtiene mediante ecocardiografía, que es una evaluación de la pared que separa el ventrículo izquierdo durante la fase de llenado completo (llamada diástole). Este diafragma, formado por tejido muscular cardíaco, desempeña un papel clave en la función cardíaca al dividir el ventrículo izquierdo en dos cámaras: la cámara anterior, que contiene sangre recién oxigenada, y la cámara posterior, que contiene sangre pre bombeada. cuerpo. . El grosor y la función del tabique telediastólico del ventrículo izquierdo se evalúa mediante ecocardiografía para detectar posibles anomalías como hipertrofia (engrosamiento excesivo) o disfunción de esta pared. Estas anomalías pueden indicar una enfermedad cardíaca, como presión arterial alta o enfermedad valvular. La medición y evaluación del intervalo telediastólico del ventrículo izquierdo es esencial para comprender la salud del corazón y guiar el diagnóstico y tratamiento de las enfermedades cardiovasculares.

La pared telediastólica posterior del ventrículo izquierdo es una medición ecocardiográfica importante que se refiere a la evaluación del estado completo del ventrículo izquierdo del corazón durante la diástole. Esta pared está compuesta principalmente de tejido de músculo cardíaco y desempeña un papel importante en la función cardíaca al facilitar la contracción y expulsión eficiente de sangre a la circulación sistémica. La medición y evaluación de la pared telediastólica posterior del ventrículo izquierdo es esencial para identificar posibles anomalías, como engrosamiento o dilatación excesivos, que pueden indicar hipertensión o enfermedad cardíaca como la miocardiopatía. Estos hallazgos son esenciales para comprender la salud del corazón y guiar el diagnóstico y tratamiento de las enfermedades cardiovasculares porque proporcionan información valiosa sobre la estructura y función del ventrículo izquierdo.

Masa del ventrículo izquierdo: es la medición principal obtenida por ecocardiografía. Esta medida indica el peso o masa muscular total del ventrículo izquierdo del corazón, que es la principal cámara encargada de bombear sangre oxigenada al resto del cuerpo. La masa del ventrículo izquierdo es un indicador importante de la salud cardiovascular y puede variar con la edad, el sexo y otros factores individuales.

La ecocardiografía calcula la masa del VI midiendo el espesor de la pared del VI y el diámetro de la cavidad ventricular. Esta medición es importante para identificar problemas cardíacos como la hipertrofia ventricular izquierda (engrosamiento excesivo de las paredes de la cámara), que a menudo se asocia con afecciones como hipertensión y miocardiopatía. El conocimiento del VI grande es esencial para el diagnóstico temprano y el tratamiento adecuado de la enfermedad cardíaca, lo que puede ayudar a prevenir complicaciones graves a largo plazo.

La fracción de eyección del ventrículo izquierdo (FEVI) es un parámetro importante evaluado mediante ecocardiografía, que mide la eficiencia con la que el ventrículo izquierdo del corazón bombea sangre. Expresada como porcentaje, la FEVI es la relación entre la sangre expulsada del ventrículo izquierdo durante cada contracción del corazón (sístole) y la cantidad total de sangre en el ventrículo izquierdo en el llenado máximo durante la fase de llenado completo (diástole). En otras palabras, la fracción de eyección del ventrículo izquierdo indica qué tan bien bombea sangre el corazón.

La FEVI es una medida clave de la función cardíaca y se utiliza para evaluar la salud del corazón. Una FEVI normal suele estar entre el 50% y el 70%. Un valor de FEVI bajo puede indicar un problema cardíaco, como insuficiencia cardíaca o miocardiopatía, mientras que un valor alto puede indicar un corazón hiperactivo. La FEVI desempeña un papel fundamental en el diagnóstico, la evaluación y el seguimiento de las enfermedades cardíacas y es fundamental para las decisiones de tratamiento y el tratamiento médico adecuado.

La velocidad máxima de la válvula mitral es un parámetro importante medido mediante ecocardiografía Doppler, una técnica ecocardiográfica especializada que evalúa el flujo sanguíneo a través del corazón y sus válvulas. Esta velocidad es la velocidad máxima a la que la sangre fluye a través de la válvula mitral durante la fase de llenado del ventrículo izquierdo, llamada diástole. La válvula mitral está situada entre la aurícula y el ventrículo izquierdos, y su función es permitir que la sangre fluya desde la aurícula al ventrículo cuando el corazón se relaja.

Medir la velocidad máxima de la válvula mitral es esencial para evaluar la función de la válvula mitral y detectar problemas potenciales como estenosis (estrechamiento) o regurgitación de la válvula mitral. Los valores anormales de esta frecuencia pueden indicar un flujo sanguíneo restringido o un flujo inverso en la aurícula izquierda. Estas mediciones ayudan a los médicos a diagnosticar y evaluar enfermedades valvulares y a tomar decisiones de tratamiento, como la reparación o el reemplazo de la válvula mitral.

La relación E/A: se refiere a la relación entre dos velocidades de flujo sanguíneo en el corazón, específicamente a través de la válvula mitral (VM) durante la fase de llenado del ventrículo izquierdo (diástole). La "E" representa la velocidad temprana del llenado diastólico, cuando la sangre fluye desde la aurícula izquierda al ventrículo izquierdo en respuesta a la relajación de las aurículas. La "A" representa la velocidad tardía del llenado diastólico, que ocurre cuando la aurícula se contrae para empujar la sangre adicional al ventrículo.

Esta relación E/A (E dividido por A) es un parámetro importante para evaluar la función cardíaca y detectar posibles trastornos. Una relación E/A normal suele estar en el rango de 0.8 a 2.5. Un valor bajo de E/A podría indicar un problema en el llenado temprano del ventrículo izquierdo,

como en la insuficiencia cardíaca con fracción de eyección preservada. Por otro lado, un valor alto de E/A podría sugerir una función de llenado anormal, como en la estenosis mitral. La relación E/A VM es esencial para ayudar a los médicos a comprender cómo fluye la sangre en el corazón y puede ser un indicador importante en el diagnóstico y tratamiento de enfermedades cardíacas y trastornos del llenado ventricular.

La relación E/E lateral es un parámetro importante en la evaluación de la ecocardiografía Doppler, una técnica ecocardiográfica especializada que mide el flujo sanguíneo y la velocidad de movimiento de las estructuras del corazón. Esta relación se refiere a la relación entre dos mediciones específicas: "E" representa el flujo sanguíneo máximo durante las etapas iniciales del llenado del ventrículo izquierdo y "E" representa la velocidad de la onda. La pared lateral del ventrículo durante la diástole.

La relación E/E lateral es una medida importante para estimar la presión auricular izquierda. Los valores elevados de E/E lateral pueden indicar una presión auricular izquierda elevada, lo que puede indicar una afección médica como insuficiencia cardíaca con fracción de eyección reducida o enfermedad de la válvula mitral. Esta medición es valiosa para los médicos en el diagnóstico y evaluación de enfermedades cardíacas y puede ayudar a guiar las decisiones de tratamiento y guiar la salud cardiovascular de los pacientes.

3.2.5 Analíticas de sangre

Los análisis de sangre, también conocidos como análisis de sangre o pruebas de laboratorio de sangre, son herramientas básicas de la medicina moderna que proporcionan información valiosa sobre la salud y el funcionamiento de una persona. Esto implica obtener una muestra de sangre de una persona para realizar más pruebas en un laboratorio clínico. Este análisis de sangre puede revelar una serie de datos importantes que ayudan a los médicos a diagnosticar la enfermedad, evaluar la eficacia del tratamiento y controlar el estado general de salud. [7]

El proceso de análisis de sangre incluye varias etapas importantes:

Extracción de muestra: El primer paso es extraer una pequeña cantidad de sangre del paciente. Esto generalmente se hace usando una aguja que se inserta en una vena, más comúnmente en el brazo. La cantidad de sangre necesaria depende del tipo de análisis y de la prueba específica a realizar. Procesamiento: Una vez obtenida la muestra, se coloca en tubos especiales que contienen anticoagulantes u otros aditivos para evitar que la sangre se coagule o se descomponga. Estos tubos se procesan en el laboratorio para separar los diferentes componentes de la sangre. Análisis: Una muestra de sangre se somete a diversas pruebas y métodos en un laboratorio para evaluar diversos aspectos de la salud. Estas pruebas pueden incluir química, hematología, inmunología, análisis genético y más. Información obtenida:

Los análisis de sangre proporcionan amplia información sobre la salud de una persona. Algunos de los datos más comunes e importantes que se pueden obtener incluyen:

Conteo sanguíneo completo (CBC): Evalúa los niveles de glóbulos rojos, glóbulos blancos y plaquetas en la sangre, lo que puede indicar anemia, infección o trastornos de la coagulación. Perfil bioquímico: mide los niveles en sangre de diversas sustancias químicas como glucosa, lípidos, enzimas hepáticas y electrolitos. Puede ayudar a diagnosticar afecciones como diabetes, enfermedades hepáticas y desequilibrios electrolíticos. Perfil lipídico: evalúa los niveles de

colesterol y triglicéridos en sangre, que son importantes para la evaluación del riesgo cardiovascular.

Función renal y hepática: las pruebas de función renal y hepática miden los niveles de creatinina, urea, bilirrubina y otras sustancias para evaluar qué tan bien están funcionando los riñones y el hígado. Prueba de coagulación: valora el tiempo que tarda la sangre en coagularse, lo cual es muy importante para diagnosticar trastornos de la coagulación y riesgo de trombosis.

Marcadores inmunológicos: detectan la presencia de anticuerpos y otros componentes inmunes para diagnosticar enfermedades autoinmunes y evaluar la respuesta inmune. Uso clínico:

Los análisis de sangre son esenciales para el diagnóstico y seguimiento de muchas enfermedades y condiciones de salud. Algunas de las aplicaciones más comunes incluyen:

Diagnóstico de enfermedades: los análisis de sangre pueden ayudar a diagnosticar afecciones como diabetes, anemia, enfermedades cardíacas y enfermedades de la tiroides. Seguimiento de la terapia: Pueden utilizarse para evaluar la eficacia del tratamiento médico y ajustar la dosis si es necesario.

Detección temprana: Ciertas enfermedades, como el cáncer, se pueden detectar tempranamente mediante análisis de sangre que buscan marcadores específicos. Evaluación de salud general: los análisis de sangre también se utilizan como parte de un examen físico de rutina para evaluar la salud general y detectar problemas potenciales antes de que se vuelvan graves.

3.2.5.1 Diagnóstico de enfermedades

El diagnóstico de enfermedades mediante análisis de sangre es un elemento básico de la medicina moderna. Estas pruebas proporcionan información valiosa que puede ayudar a los médicos a identificar la enfermedad, determinar su gravedad, elegir el tratamiento adecuado y controlar la respuesta del paciente al tratamiento. A continuación, se muestra información sobre cómo se utiliza la información obtenida de los análisis de sangre para diagnosticar diversas enfermedades:

1- Diabetes:

   Azúcar en sangre: El nivel alto de azúcar en sangre (hiperglucemia) puede indicar diabetes tipo 1 o tipo 2. Hemoglobina glicada (HbA1c): esta prueba mide el nivel promedio de azúcar en sangre durante los últimos 2 a 3 meses. Valores elevados indican diabetes mal controlada.

2- Enfermedades cardiovasculares:

   Perfil lipídico: evalúa el colesterol LDL (“malo”), el colesterol HDL (“bueno”) y los triglicéridos. Los niveles elevados de colesterol LDL y triglicéridos aumentan el riesgo cardiovascular. Marcadores de inflamación (como la proteína C reactiva): estos marcadores pueden indicar inflamación de los vasos sanguíneos asociada con enfermedades cardíacas.

3- Anemia:

   Recuento de glóbulos rojos y hemoglobina: los recuentos bajos indican anemia, que puede ser causada por deficiencia de hierro, deficiencia de vitamina B12, enfermedades crónicas, etc.

4- Enfermedad del hígado:

Pruebas de función hepática: los niveles elevados de enzimas hepáticas como ALT y AST pueden indicar daño hepático o una afección como hepatitis.

5- Enfermedades renales:

Niveles de creatinina y urea: los números elevados pueden indicar un problema renal, como insuficiencia renal.

6- Enfermedad de tiroides:

Los niveles anormales de hormonas tiroideas (T3, T4 y TSH) pueden indicar hipotiroidismo o hipertiroidismo.

7- Enfermedades autoinmunes:

Pruebas de autoanticuerpos: estas pruebas buscan la presencia de anticuerpos que atacan los propios tejidos del cuerpo, como en la artritis reumatoide o el lupus.

8- Enfermedades infecciosas:

Pruebas serológicas: detectan la presencia de anticuerpos específicos contra infecciones como el VIH, la hepatitis y muchas otras enfermedades infecciosas.

9- Cáncer:

Marcadores tumorales: algunos cánceres producen ciertas proteínas que pueden detectarse en la sangre, como el antígeno prostático específico (PSA) para el cáncer de próstata.

10- Trastornos de la coagulación:

Pruebas de coagulación: estas pruebas miden la capacidad de la sangre para coagularse normalmente, lo que es útil para diagnosticar enfermedades como la hemofilia

3.2.5.2 Monitorización de tratamientos

El seguimiento de la terapia mediante análisis de sangre es una parte importante de la atención médica porque puede evaluar la eficacia de tratamientos específicos y ajustarlos si es necesario. Estas prácticas son esenciales para garantizar que los pacientes reciban la atención adecuada y optimicen los resultados de salud. Aquí se explica detalladamente cómo seguir el tratamiento con análisis de sangre:

1. Evaluación inicial: Antes de comenzar el tratamiento, generalmente se realizan análisis de sangre iniciales para establecer una base de valores de referencia. Esto permite a los médicos comparar los resultados posteriores con los valores iniciales y determinar si la terapia está teniendo un efecto positivo.

2. Seleccione los parámetros a monitorear: Dependiendo de la enfermedad y el tratamiento relacionado, el médico elegirá parámetros específicos para monitorear. Por ejemplo, se pueden controlar los niveles de azúcar en sangre en pacientes con diabetes y los niveles de colesterol y triglicéridos en pacientes con enfermedades cardiovasculares.

3. Evaluación de la respuesta al tratamiento: Después de iniciar el tratamiento, se realizan análisis de sangre adicionales a intervalos regulares para evaluar la respuesta del paciente. Si el tratamiento es eficaz, los resultados de las pruebas deberían mostrar una mejora en los valores monitorizados. Por ejemplo, un paciente hipertenso puede esperar una caída en los niveles de presión arterial.

4. Adaptación del tratamiento: Su médico puede ajustar su tratamiento si los resultados de los análisis de sangre no muestran la mejora esperada o si hay efectos secundarios preocupantes. Esto puede incluir cambios en la dosis, la frecuencia de la dosificación o incluso la elección de tratamientos alternativos.

5. Ajuste del tratamiento: El seguimiento constante permite a los médicos adaptar el tratamiento a cada paciente. No todos los pacientes responden de la misma manera a un tratamiento concreto, por lo que se puede adaptar a las necesidades individuales.

6. Prevención de efectos secundarios y complicaciones: El seguimiento regular también puede detectar efectos secundarios relacionados con el tratamiento o complicaciones tempranas. Por ejemplo, si un paciente que toma medicamentos para la diabetes desarrolla hipoglucemia (bajo nivel de azúcar en la sangre), se puede ajustar la dosis para prevenir problemas más graves.

7. Mantenimiento a largo plazo:El seguimiento no se limita a las primeras etapas del tratamiento. Incluso después de obtener una reacción positiva, se pueden realizar análisis de sangre periódicos para garantizar que los resultados se mantengan en el rango deseado con el tiempo.

3.2.5.3 Variables extraídas de la analítica de sangre

En este apartado se nombran un amplio número de variables que pueden ser obtenidas a partir de una analítica y que han sido utilizadas en los procedimientos de Machine Learning y Deep Learning de este trabajo de fin de grado. Para el resumen rápido de un campo tan extenso como es el caso de las analíticas de sangre se ha optado por resumir la información que puede llegar a proporcionar cada una de estas variables. Por lo tanto, para cada variable se para que campo de la salud es interesante esta medida y una descripción general junto con lo que implica valores que se alejen de la normalidad. [8]

Glucosa: Medir la glucosa en muestras de sangre es esencial para evaluar los niveles de azúcar en sangre. Un nivel normal de azúcar en sangre en ayunas suele ser inferior a 100 mg/dL. La prueba es crucial en el diagnóstico y tratamiento de la diabetes, una enfermedad en la que los niveles elevados de azúcar en sangre pueden provocar complicaciones. También se utiliza para evaluar la respuesta al tratamiento antidiabético y la detección temprana de problemas metabólicos. Las mediciones de glucosa en sangre brindan información fundamental sobre la regulación metabólica y ayudan a los médicos a tomar decisiones informadas para prevenir complicaciones y mantener la salud general.

Creatinina: La determinación de creatinina en muestras de sangre es una medida importante de la función renal. La creatinina es un subproducto muscular que se filtra por los riñones y se excreta en la orina. Un nivel normal de creatinina en la sangre indica una función renal saludable, mientras que un nivel elevado puede indicar un problema de filtración renal, como insuficiencia renal. La prueba es esencial para diagnosticar y controlar la enfermedad renal, ajustar las dosis de los medicamentos y evaluar la salud general. Las mediciones de creatinina brindan información

importante sobre el estado de los riñones y ayudan a los médicos a tomar decisiones precisas sobre la atención del paciente.

Ácido úrico: La determinación del ácido úrico en muestras de sangre es fundamental para evaluar el metabolismo de las purinas (sustancias que se encuentran en los alimentos y tejidos corporales). Cuando se descomponen las purinas, se forma ácido úrico, que normalmente se excreta a través de los riñones. Los niveles elevados de ácido úrico en la sangre pueden provocar la formación de cristales en las articulaciones, lo que provoca gota, una enfermedad dolorosa. Esta prueba es fundamental para diagnosticar y controlar la gota y evaluar problemas renales. Las mediciones de ácido úrico proporcionan información metabólica importante y ayudan a los médicos a tomar decisiones informadas para prevenir complicaciones y mantener la salud de las articulaciones y los riñones.

Proteínas totales: La determinación de los niveles de proteínas totales en muestras de sangre es esencial para evaluar el estado general de las proteínas en el cuerpo. Las proteínas desempeñan un papel esencial en varias funciones biológicas, como la regulación del sistema inmunológico, el transporte de sustancias y la coagulación sanguínea. La proteína sanguínea total incluye albúmina y globulina. Una proporción anormal de estas proteínas puede indicar problemas con la función hepática, la nutrición o el sistema inmunológico. Esta prueba es esencial para diagnosticar y controlar enfermedades hepáticas, problemas nutricionales y trastornos del sistema inmunológico. Las mediciones de proteínas totales brindan información importante sobre la salud física y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Sodio: Medir el sodio en muestras de sangre es esencial para evaluar el equilibrio de electrolitos en el cuerpo. El sodio juega un papel importante en la regulación del equilibrio hídrico y en la transmisión de señales nerviosas y musculares. Los niveles adecuados de sodio son esenciales para mantener la presión arterial y la función celular. Tanto los niveles altos como los bajos de sodio en la sangre pueden indicar un desequilibrio grave, como deshidratación o problemas renales. Esta prueba es esencial para diagnosticar y controlar alteraciones electrolíticas y afecciones como la hiponatremia o la hipernatremia. Las mediciones de sodio brindan información importante sobre el estado de los líquidos y la función neuromuscular, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención al paciente.

Potasio: Medir el potasio en muestras de sangre es esencial para evaluar el equilibrio electrolítico y la función celular en el cuerpo. El potasio desempeña un papel clave en la regulación de la actividad eléctrica del corazón, la contracción muscular y el equilibrio de líquidos. Los niveles normales de potasio en sangre son esenciales para mantener la frecuencia cardíaca y la función neuromuscular. Tanto los niveles altos como los bajos de potasio en la sangre pueden indicar problemas graves, como problemas renales, deshidratación o problemas cardíacos. Esta prueba es esencial para diagnosticar y controlar trastornos electrolíticos, enfermedades renales y afecciones que afectan el corazón y los músculos. Las mediciones de potasio brindan información importante sobre la salud de los sistemas cardiovascular y neuromuscular, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención al paciente.

Calcio: Medir el calcio en muestras de sangre es esencial para evaluar la salud ósea, la función neuromuscular y la coagulación. El calcio es esencial para la formación de huesos y dientes, así como para la transmisión de señales nerviosas y la contracción muscular. Unos niveles adecuados de calcio en sangre son esenciales para mantener el equilibrio ácido-base y una coagulación sanguínea eficaz. Tanto los niveles altos como los bajos de calcio en sangre pueden indicar desequilibrios graves, como desequilibrios hormonales, enfermedades óseas o problemas renales. Esta prueba es esencial para diagnosticar y controlar trastornos del calcio como la hipocalcemia o la hipercalcemia, así como para evaluar la salud ósea y la función neuromuscular. Las

mediciones de calcio brindan información importante sobre la salud ósea y la función neurológica, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención al paciente.

Hierro: Medir el hierro en muestras de sangre es esencial para evaluar la salud de los glóbulos rojos y la capacidad del cuerpo para transportar oxígeno. El hierro es un componente importante de la hemoglobina, la proteína de los glóbulos rojos que transporta oxígeno desde los pulmones a los tejidos del cuerpo. Unos niveles adecuados de hierro en sangre son esenciales para prevenir la anemia y mantener una función celular óptima. Tanto los niveles altos como los bajos de hierro en sangre pueden indicar un problema de salud, como anemia por deficiencia de hierro o una enfermedad inflamatoria. Esta prueba es esencial para diagnosticar y controlar los trastornos del hierro y evaluar la salud de los glóbulos rojos. Las mediciones de hierro brindan información vital sobre la capacidad del cuerpo para transportar oxígeno y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Ferritina: Los niveles de ferritina en las muestras de sangre son importantes para evaluar las reservas de hierro en el cuerpo. La ferritina es una proteína que almacena y libera hierro cuando el cuerpo lo necesita. Un nivel normal de ferritina en la sangre indica un equilibrio adecuado de hierro, que es esencial para la función celular y la producción de glóbulos rojos. Tanto los niveles altos como los bajos de ferritina en sangre pueden indicar un desequilibrio como anemia por deficiencia de hierro o reducción del almacenamiento de hierro. Esta prueba es esencial para diagnosticar y controlar los problemas de hierro y evaluar la capacidad del cuerpo para mantener reservas adecuadas de hierro. Las mediciones de ferritina brindan información importante sobre el estado del almacenamiento de hierro y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Transferrina: Medir la transferrina en muestras de sangre es esencial para evaluar la capacidad del cuerpo para transportar hierro. La transferrina es una proteína que se une al hierro y lo transporta a través del torrente sanguíneo hasta los tejidos que lo necesitan, como la médula ósea y el hígado. Un nivel normal de transferrina en sangre indica una capacidad suficiente de transporte de hierro y favorece la formación de glóbulos rojos y otras funciones vitales. Tanto los niveles bajos como los altos de transferrina en la sangre pueden indicar desequilibrios como anemia o reducción del almacenamiento de hierro. Esta prueba es esencial para diagnosticar y monitorear problemas de hierro y evaluar la eficiencia del sistema de transporte de hierro del cuerpo. Las mediciones de transferrina brindan información esencial sobre la disponibilidad y distribución del hierro, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención al paciente.

Troponina: La detección de troponina en muestras de sangre es esencial para evaluar el daño del músculo cardíaco. La troponina es una proteína que ingresa al torrente sanguíneo cuando las células del corazón se dañan, como en un ataque cardíaco. Los niveles elevados de troponina en sangre indican daño miocárdico y contribuyen al diagnóstico precoz de insuficiencia cardíaca aguda. La prueba es muy importante en la evaluación de afecciones como el infarto de miocardio, permitiendo a los médicos determinar la gravedad y el tratamiento adecuado. Las mediciones de troponina brindan información importante sobre la salud del corazón y ayudan a los médicos a tomar decisiones rápidas y precisas sobre la atención del paciente.

Triglicéridos: La determinación de los niveles de triglicéridos en muestras de sangre es esencial para evaluar los niveles de grasa corporal. Los triglicéridos son un tipo de grasa que se encuentra en la sangre y se almacena en las células grasas. Los niveles normales de triglicéridos son importantes para el funcionamiento normal del cuerpo, pero los niveles elevados de triglicéridos pueden indicar un mayor riesgo de enfermedad cardiovascular, como ataque cardíaco y accidente cerebrovascular. La prueba es esencial para diagnosticar y monitorear la salud cardiovascular, y es especialmente importante para evaluar factores de riesgo como la obesidad y la diabetes. Las

mediciones de triglicéridos brindan información importante sobre el metabolismo de los lípidos y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Colesterol: La determinación de los niveles de colesterol en muestras de sangre es fundamental para evaluar los niveles de lípidos en el organismo. El colesterol es una grasa que juega un papel importante en la estructura celular y la producción de hormonas. Hay dos tipos principales de colesterol: el colesterol LDL ("malo") y el colesterol HDL ("bueno"). Los niveles elevados de colesterol LDL se asocian con un mayor riesgo de enfermedad cardiovascular porque se acumula en las arterias y provoca obstrucciones. Por el contrario, el colesterol HDL elevado se asocia con un menor riesgo de enfermedad cardíaca porque ayuda al cuerpo a eliminar el exceso de colesterol. La prueba es esencial para evaluar el riesgo cardiovascular y su interpretación, junto con otros factores, puede guiar las decisiones de los médicos para prevenir enfermedades cardíacas. Las mediciones de colesterol brindan información importante sobre la salud cardiovascular y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

NT-proBNP: La medición de NT-proBNP en muestras de sangre es esencial para evaluar la función cardíaca y detectar posibles problemas cardíacos. NT-proBNP es un péptido liberado por el corazón en respuesta a la presión y el estrés ventriculares, particularmente en la insuficiencia cardíaca. Los niveles elevados de NT-proBNP indican una posible disfunción cardíaca, ya que la liberación de NT-proBNP aumenta cuando el corazón tiene dificultades para bombear sangre de manera eficiente. Esta prueba es esencial para diagnosticar y monitorear la insuficiencia cardíaca, así como para evaluar la gravedad y progresión de la afección. La medición de NT-proBNP proporciona información fundamental sobre la salud del corazón y ayuda a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite una intervención temprana y un tratamiento adecuado de la enfermedad cardíaca.

GPT: La determinación de GPT (Alanina Aminotransferasa) en muestras de sangre es fundamental para evaluar la función hepática. GPT es una enzima que se encuentra en las células del hígado cuya liberación a la sangre puede indicar daño o enfermedad hepática. Los niveles elevados de GPT en la sangre pueden indicar afecciones como hepatitis, cirrosis o consumo excesivo de alcohol. Esta prueba es esencial para diagnosticar y controlar la salud del hígado, y su interpretación junto con otras pruebas puede ayudar a los médicos a determinar la causa y la gravedad del daño hepático. Las mediciones de GPT proporcionan información crítica sobre la salud del hígado, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite el diagnóstico temprano y el tratamiento adecuado de la enfermedad hepática.

GOT: La determinación de GOT (aspartato transaminasa) en muestras de sangre es fundamental para evaluar la salud del hígado y otros órganos. GOT es una enzima que se encuentra en varias partes del cuerpo, incluidos el hígado, el corazón y los músculos. Los niveles elevados de GOT en la sangre pueden indicar daño hepático como hepatitis o cirrosis, así como daño cardíaco o muscular. Sin embargo, la interpretación de los resultados de GOT debe combinarse con otras pruebas para lograr una comprensión integral. La prueba es esencial para el diagnóstico y seguimiento de la salud hepática y cardiovascular, y su uso también se extiende a la evaluación de diversas condiciones de salud. Las mediciones de GOT brindan información esencial sobre el funcionamiento de diversos órganos, lo que ayuda a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite un diagnóstico temprano y un manejo adecuado de las afecciones médicas.

TSH: La medición de TSH (hormona estimulante de la tiroides) en muestras de sangre es esencial para evaluar la función tiroidea y el equilibrio hormonal. La TSH, producida por la glándula pituitaria, regula la producción de hormonas tiroideas, que a su vez afectan el metabolismo y la

energía del cuerpo. Los niveles elevados de TSH en la sangre pueden indicar hipotiroidismo, que es la incapacidad de la glándula tiroides para producir suficiente cantidad de esta hormona. Por otro lado, un nivel bajo de TSH puede indicar hipertiroidismo, que es una producción excesiva de hormona tiroidea. Esta medición es necesaria para diagnosticar y monitorear problemas de tiroides, y su interpretación junto con otras pruebas puede ayudar a los médicos a determinar la función tiroidea y guiar el tratamiento adecuado. Las mediciones de TSH brindan información importante sobre el equilibrio hormonal y ayudan a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite un diagnóstico temprano y un tratamiento eficaz de la enfermedad de la tiroides.

T4 libre: La determinación de T4 libre (tiroxina libre) en muestras de sangre es fundamental para evaluar la función tiroidea y el equilibrio hormonal. La T4 libre es una hormona producida por la glándula tiroides que es muy importante en la regulación del metabolismo y la función general del cuerpo. Los niveles bajos de T4 libre pueden indicar hipotiroidismo, que es la incapacidad de la glándula tiroides para producir suficiente hormona. Por otro lado, los niveles elevados de T4 libre pueden indicar hipertiroidismo, que es una producción excesiva de hormona tiroidea. Esta prueba es esencial para diagnosticar y monitorear problemas de tiroides, y su interpretación junto con otras pruebas puede ayudar a los médicos a determinar la función tiroidea y guiar el tratamiento adecuado. Las mediciones gratuitas de T4 brindan información importante sobre el equilibrio hormonal y ayudan a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite un diagnóstico temprano y un tratamiento eficaz de la enfermedad de la tiroides.

T3 libre: La determinación de T3 libre (triyodotironina libre) en muestras de sangre es fundamental para evaluar la función tiroidea y el equilibrio hormonal. La T3 libre es una hormona producida por la glándula tiroides, que es crucial en el metabolismo, el crecimiento y el desarrollo del cuerpo. Los niveles bajos de T3 libre pueden indicar hipotiroidismo, que es la incapacidad de la glándula tiroides para producir suficiente hormona. Por otro lado, los niveles elevados de T3 libre pueden indicar hipertiroidismo, que es una producción excesiva de hormona tiroidea. Esta prueba es esencial para diagnosticar y monitorear problemas de tiroides, y su interpretación junto con otras pruebas puede ayudar a los médicos a determinar la función tiroidea y guiar el tratamiento adecuado. La medida de T3 libre brinda información importante sobre el equilibrio hormonal para ayudar a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite un diagnóstico temprano y un tratamiento eficaz de la enfermedad de la tiroides.

Hemoglobina: La determinación de los niveles de hemoglobina en muestras de sangre es esencial para evaluar la capacidad de la sangre para transportar oxígeno a los tejidos del cuerpo. La hemoglobina es una proteína de los glóbulos rojos que une el oxígeno en los pulmones y lo transporta a las células de todo el cuerpo. Los niveles normales de hemoglobina son esenciales para mantener una oxigenación adecuada y una función celular óptima. Tanto los niveles altos como los bajos de hemoglobina pueden indicar un problema de salud, como anemia o trastornos sanguíneos. Esta prueba es esencial para diagnosticar y controlar problemas sanguíneos y evaluar la capacidad de la sangre para transportar oxígeno. Las medidas de hemoglobina brindan información importante sobre la salud de la sangre y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Hematocrito: Medir el hematocrito en muestras de sangre es esencial para estimar la proporción de glóbulos rojos con respecto al volumen sanguíneo total. El hematocrito es el principal indicador que determina la capacidad de la sangre para transportar oxígeno y nutrientes a los tejidos del cuerpo. Un nivel normal de hematocrito es esencial para mantener la función celular óptima y el equilibrio de líquidos en el cuerpo. Tanto los niveles altos como los bajos de hematocrito pueden indicar problemas de salud como anemia o deshidratación. La prueba es esencial para diagnosticar y monitorear trastornos sanguíneos, evaluar la salud de la sangre y guiar el tratamiento de la

anemia. Las mediciones del hematocrito brindan información importante sobre la salud de la sangre y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Ancho de distribución eritrocitaria: El ancho de distribución de los glóbulos rojos (ADE) en una muestra de sangre es una medida de la variación en el tamaño de los glóbulos rojos. Esta medición proporciona información sobre la cantidad de uniformidad de glóbulos rojos en la sangre. Los niveles elevados de ADE pueden indicar cambios en el tamaño de los glóbulos rojos, a menudo asociados con afecciones como anemia, deficiencia de hierro o inflamación crónica. Sin embargo, los ADE deben interpretarse junto con otros valores sanguíneos y el historial médico del paciente para obtener una imagen completa de la salud del paciente. La medición del ancho de distribución de los glóbulos rojos ayuda a los médicos a identificar posibles trastornos circulatorios y guiar las decisiones clínicas en el diagnóstico y tratamiento de afecciones médicas.

Plaquetas: Medir las plaquetas en muestras de sangre es importante para evaluar la coagulación sanguínea y la salud del sistema circulatorio. Las plaquetas son pequeñas células sanguíneas que desempeñan un papel vital en la formación de coágulos sanguíneos que detienen el sangrado cuando se lesionan. Los niveles normales de plaquetas son esenciales para mantener un equilibrio entre la prevención del sangrado y la coagulación excesiva. Tanto los niveles altos como los bajos de plaquetas pueden indicar un problema médico, como un trastorno hemorrágico, una infección o un trastorno de la médula ósea. Esta prueba es importante para diagnosticar y monitorear trastornos sanguíneos y evaluar la capacidad del cuerpo para controlar el sangrado. Las mediciones de plaquetas aportan información importante sobre la salud circulatoria y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Leucocitos: Medir los glóbulos blancos en muestras de sangre es esencial para evaluar la salud del sistema inmunológico y la capacidad del cuerpo para combatir infecciones y enfermedades. Los glóbulos blancos, también conocidos como glóbulos blancos, son células sanguíneas que desempeñan un papel clave en la respuesta inmunitaria. Un nivel normal de glóbulos blancos es esencial para mantener una defensa adecuada contra los patógenos. Tanto los niveles bajos como los altos de glóbulos blancos pueden indicar un problema de salud, como una infección, un trastorno del sistema inmunológico o una inflamación. La prueba es muy importante para diagnosticar y monitorear problemas del sistema inmunológico, así como para evaluar la respuesta del cuerpo a amenazas externas. Las mediciones de glóbulos blancos proporcionan información importante sobre la salud del sistema inmunológico y ayudan a los médicos a tomar decisiones informadas sobre la atención del paciente.

Fibrinógeno: La determinación de fibrinógeno en muestras de sangre es importante para evaluar la capacidad de coagulación de la sangre. El fibrinógeno es una proteína producida en el hígado que desempeña un papel crucial en la formación de coágulos sanguíneos. Un nivel normal de fibrinógeno es esencial para mantener el equilibrio entre prevenir el sangrado y formar los coágulos sanguíneos necesarios para detener el sangrado. Tanto los niveles bajos como los altos de fibrinógeno pueden indicar un problema médico, como un trastorno hemorrágico, inflamación crónica o un mayor riesgo de coágulos sanguíneos. La prueba es esencial para diagnosticar y controlar condiciones médicas que afectan la coagulación sanguínea y evaluar el riesgo de complicaciones como la trombosis. Las mediciones de fibrinógeno brindan información importante sobre la salud circulatoria y ayudan a los médicos a tomar decisiones informadas sobre la atención al paciente.

Interleucinas: La medición de interleucinas en muestras de sangre es fundamental para evaluar la respuesta inmune y la presencia de inflamación en el organismo. Las interleucinas son proteínas producidas por células del sistema inmunológico que desempeñan un papel clave en la comunicación entre las células del sistema inmunológico. Los niveles elevados de ciertas interleucinas en la sangre pueden indicar una respuesta inmune activa o inflamación en el cuerpo.

Estas pruebas son fundamentales para el diagnóstico y seguimiento de enfermedades autoinmunes, alergias, infecciones y enfermedades inflamatorias crónicas. La medición de interleucinas proporciona información importante sobre la salud del sistema inmunológico y ayuda a los médicos a tomar decisiones informadas sobre la atención del paciente, lo que permite un diagnóstico temprano y un tratamiento adecuado de las afecciones médicas relacionadas con las respuestas inmunitarias y la inflamación.

3.2.6 Electrocardiograma

Un electrocardiograma, a menudo llamado ECG o EKG, es una prueba médica no invasiva que registra la actividad eléctrica del corazón con electrodos colocados en la piel del pecho, brazos y piernas. La grabación se presenta como un diagrama que muestra las señales eléctricas que controlan el ritmo y las contracciones del corazón.

Durante cada ciclo de despolarización y repolarización, el movimiento de los iones genera una corriente eléctrica que se propaga por todo el cuerpo; es como si el cuerpo actuara como conductor de volumen. Esto genera un potencial eléctrico que se transmite a través de soluciones electrolíticas como el líquido intersticial y el plasma y llega a la superficie de la piel, donde puede medirse mediante electrodos colocados en la superficie del cuerpo. El registro de la actividad eléctrica se llama electrocardiograma y representa la secuencia de despolarización y repolarización de las aurículas y ventrículos mediante ondas.

Las 12 derivaciones que componen el ECG son bipolares (que utilizan electrodos positivos y negativos) y unipolares (que combinan un electrodo positivo y otros electrodos como un electrodo negativo compuesto).

Se deben utilizar diez electrodos para medir un ECG de 12 derivaciones. Suelen consistir en un gel conductor incrustado en el centro de la cinta adhesiva; el gel también puede formar un adhesivo. Dado que el corazón es un órgano tridimensional, es necesario conocer todos los vectores representativos de la actividad cardíaca en todos los frentes (frontal y horizontal). Por eso, dependiendo de dónde estén situados los electrodos, el pie tendrá diferentes morfologías, complejos QRS.

Las derivaciones frontales (o derivaciones Goldberger) constan de 3 electrodos bipolares (I, II y III) y 3 electrodos unipolares (aVR, aVL y aVF), que son 6 derivaciones que implementan el sistema de seis ejes de Bailey (ver Figura 3) y 4) . Todos los cables horizontales (o cables Wilson) son unipolares y están etiquetados como V1 a V6.

3.2.6.1 Posicionamiento de los electrodos

Para realizar la medición se marcan 10 electrodos y se colocan en el cuerpo del paciente de la siguiente manera:

• RA: En el brazo derecho.

• LA: En la misma ubicación en la que se colocó RA, pero en el brazo izquierdo.

• RL: En la pierna derecha.

LL: En la misma ubicación en la que se colocó RL, pero en la pierna izquierda esta vez.

• V1: En el cuarto espacio intercostal (entre las costillas 4 y 5) justo a la derecha del esternón.

• V2: En el cuarto espacio intercostal (entre las costillas 4 y 5) justo a la izquierda del esternón.

• V3: Entre los conductores V2 y V4.

• V4: En el quinto espacio intercostal (entre las costillas 5 y 6) en la línea clavicular media (línea imaginaria que se extiende hacia abajo desde el punto medio de la clavícula).

V5: Mismo nivel con V4, pero en la línea axilar anterior (línea imaginaria que baja desde el punto intermedio entre el centro de la clavícula y el extremo lateral de la clavícula)

• V6: Horizontalmente incluso con V4 y V5 en la línea axilar media (línea imaginaria que se extiende desde el centro de la axila del paciente.)

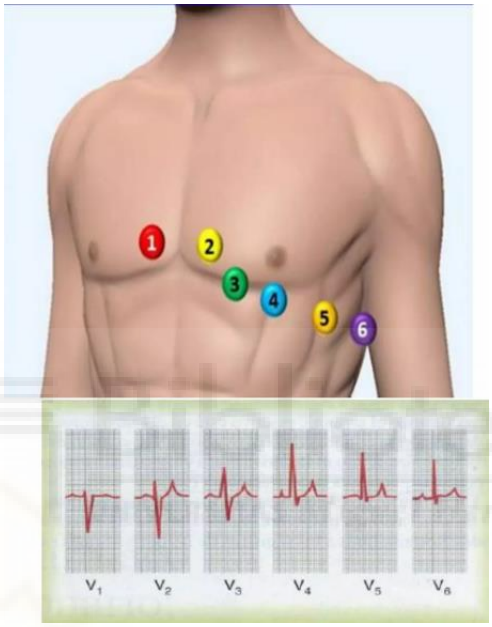

Ilustración 3 Posición de los electrodos

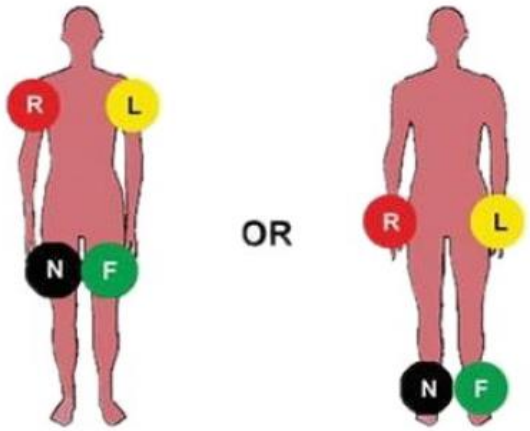

Ilustración 4 Posición de los electrodos

Como veremos a continuación, lo convencional es colocar cuatro electrodos en las extremidades y otros seis electrodos en el pecho (zona frontal).

Cada electrodo positivo se enfrenta al electrodo negativo y determina si la energía eléctrica se dirige hacia adentro (arriba) o hacia afuera (hacia abajo) desde su posición. Dependiendo del ángulo con el vector de energía eléctrica medio, la onda será más grande. Los ángulos de energía verticales producirán formas de onda con poca o ninguna polarización (isoeléctrica) o cantidades iguales de polarización positiva y negativa.

3.2.6.2 Obtención de los doce leads

De estos diez registros de electrodos, los cables V1 a V6 se obtuvieron directamente de los electrodos respectivos, mientras que los seis electrodos restantes se obtuvieron usando:

Cables de electrodos bipolares

Lead I: el electrodo positivo en el brazo izquierdo (LA) y el negativo negativo en el brazo derecho (RA). A medida que el vector medio se mueve desde la parte superior derecha a la inferior izquierda, algo de energía fluye hacia los electrodos LA, provocando una ligera desviación hacia arriba del complejo QRS.

$$I = LA - RA$$

$$(\ 1\ )$$

Lead II: Se refiere a la tensión de la pierna izquierda menos a la del electrodo en el brazo derecho (RA). El vector medio que se orienta hacia el electrodo LL, este lead generalmente tiene la onda P más vertical y la más prominente.

$$II = LL - RA$$

$$(\ 2\ )$$

Lead III: caracteriza el voltaje entre los electrodos de la pierna izquierda y del brazo izquierdo. Una vez más el complejo QRS es hacia arriba (mayor que el lead I), enfoque vectorial medio de lead III hacia abajo desde la derecha.

$$III = LL - LA$$

$$(\ 3\ )$$

El segundo tipo de leads de extremidades se denominan leads aumentados o unipolares y utilizan un solo electrodo de monitoreo positivo.

Estos cables se derivan de los mismos electrodos que los cables anteriores, pero miran el corazón desde un ángulo diferente. El electrodo negativo es la ubicación calculada eléctricamente del centro del corazón. Las derivaciones de las extremidades extendidas AVR, aVL y aVF junto con las derivaciones I, II y III forman la base de un sistema de referencia de seis ejes para calcular el eje eléctrico del corazón en el plano frontal. [9]

Lead aVR: el polo positivo está en el brazo derecho, el polo negativo es una combinación de electrodos LA y LL. El vector medio fluye hacia abajo y hacia la izquierda alejándose de aVR, y todas las desviaciones de forma de onda en esta derivación son negativas.

$$aVR = RA - \frac{1}{2}(LA + LL)$$

(4)

Lead AVL: el electrodo positivo se coloca en el brazo izquierdo y el negativo está formado por electrodos RA y LL. El vector medio que se acerca a este lead con un gran ángulo, el complejo QRS es el menos vertical entre los leads de las extremidades.

$$aVL = LA - \frac{1}{2}(RA + LL)$$

(5)

Lead AVF: tiene su electrodo positivo en la pierna izquierda y negativo producido por el brazo derecho e izquierdo. El ángulo de este lead con el vector medio está cerca del lead II, por lo tanto, aVF tiene QRS muy vertical y onda P alta también.

$$aVF = LL - \frac{1}{2}(RA + LA)$$

(6)

3.2.6.3 Las ondas del electrocardiograma

Cada ciclo de despolarización y repolarización auricular y ventricular coincide con el paso de impulsos eléctricos desde el nódulo sinoauricular hasta los extremos de las fibras de Purkinje, lo que provoca su contracción. Un electrocardiograma convierte esta actividad eléctrica en una imagen de diferentes ondas que viajan a través de las fibras. [9]

El ECG normal comprende las ondas P, Q, R, S, T y U introducidas por Einthoven en 1895 de la siguiente manera:

Onda P: Representa la propagación de la actividad eléctrica sobre las aurículas después de la despolarización inicial del nodo SA. La primera porción representa en gran parte la despolarización derecha, y el final corresponde esencialmente a la despolarización auricular izquierda. Las ondas P deben parecerse y no ser mayores de 0,3 mV.

Complejo QRS: Desde el inicio hasta el final del QRS y dura normalmente de 0,06 a 0,10s. El complejo QRS nos brinda información sobre la despolarización ventricular. Sus diferentes ondas (Q, R y S) tendrán tamaños en función de las características de los distintos vectores eléctricos generados por el músculo cardíaco y del trayecto recorrido a lo largo del sistema de conducción, o fuera de éste, a través del músculo cardíaco. La anchura normal del complejo QRS es < 0,12s, y esto corresponde a un complejo generado a nivel supraventricular y que

alcanza los ventrículos y los despolariza a través de un sistema de conducción (haz de His y sus ramas) que se encuentra íntegro (ver Ilustración 5).

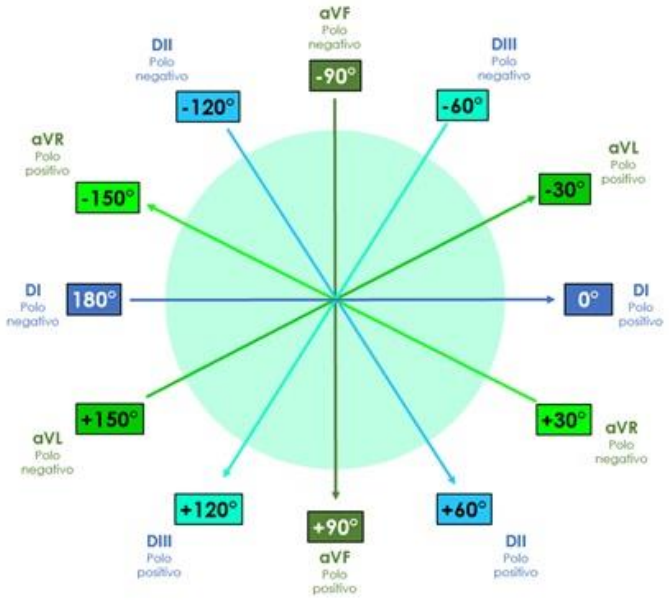

Ilustración 5 Sistema hexaxial de Bailey [10]

Onda Q: Visible en los leads ventriculares izquierdos (V5, V6, I y aVL), resultado de onda Q, despolarización de la región antero septal de izquierda a derecha. Por definición, corresponde a la primera elección hacia abajo después de la onda P. La onda Q Normal no debe exceder 0.04s en ancho y 0.2mV de profundidad.

Onda R: La primera deflexión vertical hacia arriba después de la onda P, representa la despolarización temprana de los ventrículos. Las variaciones consecutivas hacia arriba dentro del mismo complejo QRS están etiquetadas como R, R', etc.
Onda S: La primera deflexión descendente después de la onda R (en presencia de onda Q o no), caracteriza la despolarización tardía de los ventrículos. Al igual que la onda R, las desviaciones consecutivas hacia abajo se denominarán S, S', Etc.

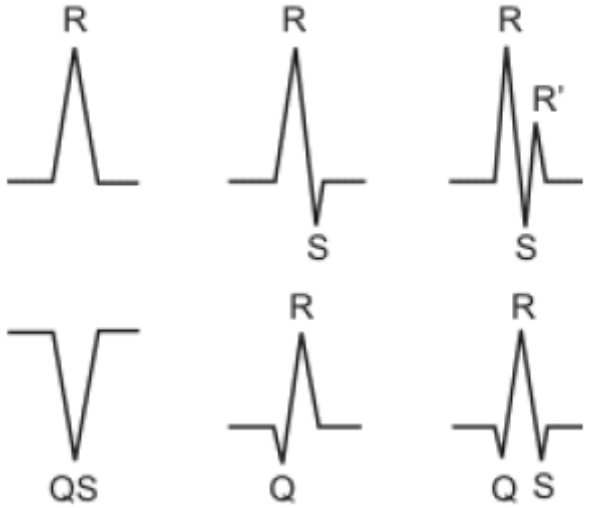

Ilustración 6 Complejos QRS más comunes

Onda T: Caracteriza el flujo de repolarización de los ventrículos y sigue el complejo QRS, siendo de mayor duración y amplitud inferior, separada por una línea isoeléctrica llamada segmento ST. Esta es la onda más lábil; su dirección suele ser la misma que el complejo QRS, excepto en el lead precordial derecho (V1, V2 y V3). La onda T normal es asimétrica con la primera mitad con una tendencia más lenta que la segunda mitad. En el ECG normal, la onda T siempre está de forma positiva en los conductores I, II, V3-6, y siempre invertida en lead aVR. Los otros leads son variables dependiendo de la dirección del QRS y la edad del paciente.

Onda U: Esta onda vertical sigue inmediatamente a la onda T, tiene la misma polaridad y es del 5% al 50% de esta última. Es más prominente en el lead V2 y V3, donde puede llegar a 0.2 mV. Al igual que la onda T, la onda U es asimétrica con la mitad ascendente con una tendencia más rápida que la mitad descendente (opuesto a la onda T).

Punto J: Este es un punto de referencia importante para medir la duración del complejo QRS. El punto J se encuentra al final de la despolarización ventricular y, en ocasiones, es difícil de encontrar. El complejo QRS es muy vertical, mientras que el segmento ST suele ser muy horizontal, y una forma de encontrar el punto J es buscar un cambio brusco de pendiente, ya que representa su intersección.

3.2.6.4 Los diferentes intervalos del ECG

Intervalo RR: Este intervalo representa la inversa de la frecuencia cardíaca y se mide entre los picos R de dos complejos QRS consecutivos.

Intervalo PR: El intervalo está delimitado por el inicio de la onda P y el comienzo del complejo QRS. Normalmente van desde 0.12s a 0.2s, representa el desplazamiento del impulso a través del nodo AVa las fibras de Purkinje, es decir, entre la activación del nodo SA y la despolarización

ventricular. Por lo tanto, este intervalo permite una buena estimación de la función del nodo AV. Si la onda Q está presente en el ECG, también se denomina comúnmente "intervalo PQ".

Segmento PR: El segmento caracteriza la parte isoeléctrica del intervalo PR desde el final de la onda T hasta el comienzo del QRS, un retraso de tiempo que permite que se lleve a cabo la sístole auricular, y el llenado ventricular. Coincide con la conducción eléctrica del nodo AV, a las ramas del haz de His y a las fibras de Purkinje.

Intervalo QT: El intervalo representa el tiempo desde el inicio del complejo QRS hasta el final de la onda T. Representa la duración de la sístole eléctrica ventricular. El intervalo QT incluye el intervalo QRS, el segmento ST y la onda T. El intervalo QT varía dependiendo de la frecuencia cardiaca, disminuye a frecuencias cardiacas rápidas y aumenta a frecuencias lentas. Por ello, para determinar si es normal o no, debemos realizar una adecuada corrección por la frecuencia (QT corregido o QTc).

Segmento ST: El segmento representa el tiempo entre el final del complejo QRS y la desviación inicial de la onda T. Durante este tiempo todo el ventrículo se despolariza.

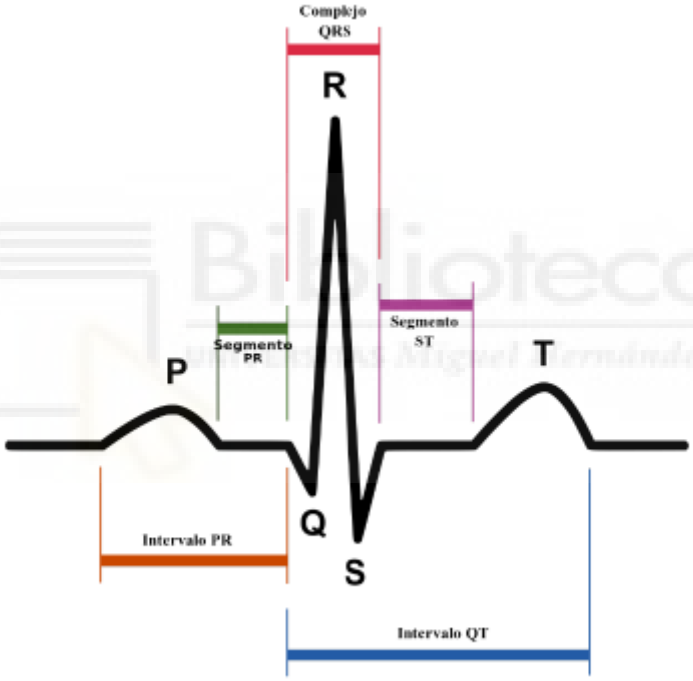

Ilustración 7 Segmento QRS

3.2.7 Descripción de la miocardiopatía hipertrófica (HCM)

La miocardiopatía hipertrófica (HCM) es una enfermedad del corazón que afecta principalmente al músculo cardíaco, conocido como el miocardio. Se trata de una afección hereditaria, lo que significa que se transmite de padres a hijos a través de mutaciones genéticas específicas. La HCM se caracteriza por el engrosamiento anormal del músculo cardíaco, especialmente en el ventrículo izquierdo, la cámara principal del corazón responsable de bombear la sangre oxigenada al cuerpo (ver Ilustración 8).

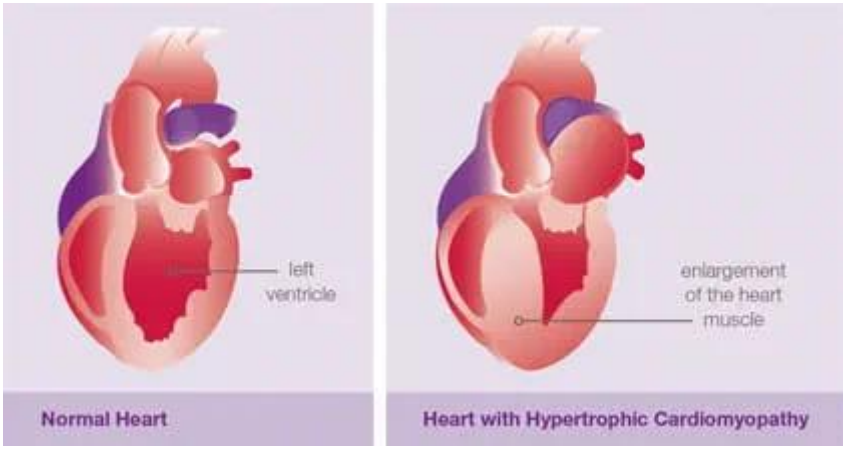

Ilustración 8 Diferencia entre un corazón normal y un corazón afectado de HCM

La miocardiopatía hipertrófica (HCM) es la causa más frecuente de muerte súbita cardíaca (MSC) en adultos jóvenes. La HCM se caracteriza por heterogeneidades en las expresiones morfológicas y los cursos clínicos. Además, es el trastorno cardiovascular hereditario autosómico dominante más común. En hasta el 60% de los pacientes con HCM se han detectado más de 1.400 mutaciones en genes que codifican proteínas sarcómeras. Las pruebas genéticas tienen un impacto limitado en las estrategias de tratamiento para pacientes individuales, pero un resultado positivo de la prueba genética podría confirmar la etiología de la enfermedad y permitir la detección genética en cascada de sus familiares. [11]

Aunque no se han establecido correlaciones precisas entre el fenotipo y el genotipo en pacientes con HCM, se consideró que los pacientes con mutaciones mostraban diferencias significativas tanto en las características clínicas como en las imágenes en comparación con aquellos sin mutaciones, como antecedentes familiares de HCM. Estudios anteriores han propuesto varias estrategias para predecir genotipos positivos en HCM basándose en el modelado de regresión lineal de variables clínicas y de imagen. Sin embargo, estas puntuaciones no fueron completamente validadas en la práctica clínica y no pudieron reflejar directamente la complejidad dinámica y fisiológica del miocardio.[11]

3.2.7.1 Causas y genética de la HCM

Mutación genética: la principal causa de la HCM es la mutación genética. Estas mutaciones afectan el gen que codifica las proteínas del músculo cardíaco, provocando un crecimiento anormal y deficiente de las fibras musculares en el corazón. Los dos genes principales implicados son: [12]

- **Gen MYH7**: este gen codifica la proteína cardíaca betamiosina, cuyas mutaciones son una causa importante de casos de MCH. Las mutaciones en MYH7 tienden a asociarse con enfermedades más graves.
- **Gen MYBPC3**: este gen codifica la proteína C de unión a miosina cardíaca y las mutaciones en este gen también son comunes en la MCH. Las mutaciones de MYBPC3 pueden causar una variedad de manifestaciones clínicas, desde casos leves hasta casos más graves. La MCH generalmente sigue un patrón de herencia autosómico dominante,

lo que significa que un individuo solo necesita una copia del gen mutado de uno de los padres para desarrollar la enfermedad. Esto aumenta el riesgo de que los hijos de personas afectadas hereden la mutación y, por lo tanto, corran el riesgo de desarrollar MCH.

Heterogeneidad genética: la MCH es una enfermedad genéticamente heterogénea, lo que significa que se han identificado varios genes que pueden contribuir a su desarrollo. Además de MYH7 y MYBPC3, también se han encontrado mutaciones en otros genes como TNNT2, TNNI3, TPM1, etc., que también pueden estar asociados con la enfermedad.

Penetrancia variable: La penetrancia se refiere a la probabilidad de que una persona con una mutación genética desarrolle la enfermedad. Con la MCH, la penetrancia es variable, lo que significa que algunas personas con la mutación genética pueden permanecer asintomáticas o tener síntomas leves, mientras que otras pueden desarrollar síntomas graves.

Expresividad variable: la expresividad se refiere a la variabilidad en la gravedad y la presentación clínica de los síntomas de la MCH. Incluso en familias con la misma mutación genética, los síntomas pueden variar ampliamente, lo que hace que el impacto clínico de la enfermedad sea difícil de predecir.

Asesoramiento genético: dada la naturaleza hereditaria de la MCH, el asesoramiento genético es fundamental. Las pruebas genéticas pueden identificar mutaciones específicas en un individuo y ayudar a determinar el riesgo de transmisión de enfermedades a generaciones futuras. Esto puede ser importante para las decisiones de planificación familiar.

3.2.7.2 Diagnóstico de la enfermedad

El diagnóstico de la HCM es un proceso integral que implica una serie de pruebas y evaluaciones clínicas para confirmar la presencia de esta enfermedad cardíaca hereditaria. A continuación, se detallan los aspectos clave del diagnóstico: [13] [14]

1. Historia Clínica y Examen Físico:

El proceso de diagnóstico comienza con una historia clínica completa. El médico recopila información detallada sobre los síntomas del paciente, antecedentes médicos personales y familiares, y cualquier factor de riesgo cardiovascular.

Durante el examen físico, el médico escucha el corazón con un estetoscopio para detectar posibles soplos cardíacos, ritmos cardíacos anormales, signos de insuficiencia cardíaca y otras anomalías.

2. Electrocardiograma (ECG o EKG):

El ECG registra la actividad eléctrica del corazón y puede revelar patrones característicos en pacientes con HCM. Algunos hallazgos comunes incluyen cambios en el voltaje eléctrico y alteraciones en la repolarización ventricular.

Es importante destacar que el ECG por sí solo no es suficiente para el diagnóstico definitivo de la HCM, pero puede proporcionar indicios importantes.

3. Ecocardiografía (Ecocardiograma):

La ecocardiografía es una herramienta fundamental para el diagnóstico de la HCM. Utiliza ultrasonidos para crear imágenes en tiempo real del corazón, lo que permite medir el grosor del músculo cardíaco, evaluar la función del corazón y detectar cualquier obstrucción del flujo de salida.

En pacientes con HCM, la ecocardiografía suele revelar un engrosamiento anormal del ventrículo izquierdo y, en algunos casos, anomalías en las válvulas cardíacas.

4. Resonancia Magnética Cardíaca (RMC):

En situaciones donde la ecocardiografía no proporciona información suficiente o se necesita una evaluación más detallada, se puede realizar una RMC. Esta técnica de imagen permite obtener imágenes tridimensionales del corazón y detectar cicatrices o áreas de fibrosis en el miocardio.

5. Pruebas de Esfuerzo:

Las pruebas de esfuerzo, como la ergometría o la prueba de esfuerzo con imágenes (ejemplo: ecocardiografía de estrés), pueden ayudar a evaluar la respuesta del corazón a la actividad física. Esto es especialmente útil para determinar si el ejercicio provoca síntomas o cambios en el patrón de la HCM.

6. Pruebas Genéticas:

Las pruebas genéticas pueden desempeñar un papel importante en el diagnóstico, especialmente en personas con antecedentes familiares de HCM. Estas pruebas pueden identificar mutaciones en genes relacionados con la HCM y ayudar a confirmar el diagnóstico.

Es importante destacar que no todas las personas con HCM tienen mutaciones genéticas identificables, por lo que la ausencia de una mutación no excluye el diagnóstico.

7. Evaluación Familiar:

Dado que la HCM es una enfermedad hereditaria, es fundamental evaluar a los familiares de las personas diagnosticadas. Se pueden identificar otros miembros de la familia en riesgo y realizar evaluaciones cardíacas regulares, incluidas pruebas genéticas si es necesario.

8. Consideración de las Directrices Clínicas:

Los médicos se basan en las directrices clínicas y los criterios diagnósticos específicos para confirmar el diagnóstico de HCM. Estos criterios pueden variar según la organización médica y se actualizan periódicamente.

4 Materiales y métodos

4.1 La base de datos

En el marco de nuestro Trabajo de Fin de Grado (TFG), hemos tenido el privilegio de acceder y utilizar una valiosa base de datos proporcionada por el Hospital Universitario Virgen de la Arrixaca de Murcia. Esta base de datos en particular se ha convertido en una herramienta fundamental para nuestra investigación, ya que contiene registros médicos relevantes relacionados con la enfermedad de la miocardiopatía hipertrófica (HCM, por sus siglas en inglés).

La base de datos del Hospital de Murcia nos ha proporcionado acceso a una variedad de registros médicos esenciales para nuestro estudio de la HCM. En particular, hemos podido analizar electrocardiogramas (ECG), analíticas de laboratorio y ecocardiogramas (ECO) relacionados con pacientes diagnosticados con miocardiopatía hipertrófica.

Es importante destacar que hemos seguido todas las regulaciones éticas y de privacidad para el manejo y utilización de la base de datos del hospital. Las identidades de cada paciente han sido anonimizadas de dos maneras diferentes antes de comenzar con el análisis para garantizar la privacidad de cada uno de los pacientes.

La base de datos original proporcionada por el hospital de Murcia contaba con 4 tablas de variables que contenían los siguientes datos: ecocardiogramas, analíticas de sangre, electrocardiogramas y una cuarta tabla que aportaba información sobre una clasificación de cada paciente en función de su relación con la enfermedad diagnosticados previamente por los médicos del hospital de Murcia. Los grupos en los que se clasifican son los siguientes:

Grupo 0: A este grupo de pacientes se les ha realizado diversos controles y los resultados han sido normales, además presentan genética negativa.

Grupo 1: En este grupo se encuentran los pacientes a los cuales se les han hecho las pruebas pertinentes y no manifiestan síntomas de la enfermedad. Sin embargo las pruebas genéticas son positivas por lo que existe la posibilidad de que la enfermedad se desarrolle en una época futura.

Grupo 2: A los pacientes de este grupo se les han realizado las pruebas para detectar HCM pero los resultados no han terminado de ser resolutivos y además presentan genética positiva.

Grupo 3: En este grupo los pacientes han dado positivo en las pruebas por lo que sin ninguna duda padecen de HCM y además tienen genética positiva.

4.1.1 Valores vacíos y relleno de huecos

En nuestra base de datos, las variables se consideran Missing Completely At Random (MCAR), lo que significa que los valores faltantes se distribuyen de forma aleatoria y no están relacionados con los valores observados o cualquier otra característica del conjunto de datos. Esto nos permite aplicar técnicas de relleno de huecos aleatorias o basadas en promedios sin preocuparnos por sesgos o patrones específicos en los datos faltantes. Sin embargo, es importante considerar cuidadosamente las estrategias de relleno y evaluar el impacto en los análisis posteriores.

El tratamiento de relleno de huecos para variables Missing Completely At Random (MCAR) implica llenar los valores faltantes de manera aleatoria sin que estén relacionados con los valores observados o cualquier otra característica del conjunto de datos. Esto se debe a que los valores faltantes se consideran aleatorios y no están influenciados por ninguna variable o patrón específico en los datos.

Los datos se derivan de una base de datos médica que contiene información clínica y médica de pacientes con y sin la Enfermedad de Cardiomiopatía Hipertrófica (HCM). Dada la importancia crítica de la precisión y la integridad de los datos en el ámbito médico, cualquier imputación de datos debe ser cuidadosamente considerada y justificada. La imputación incorrecta podría llevar a diagnósticos erróneos o decisiones médicas inadecuadas.

Hay varias estrategias que puedes seguir para realizar el relleno de huecos en variables MCAR. Algunas de las técnicas comunes incluyen:

Relleno con valores aleatorios: Puedes generar valores aleatorios a partir de una distribución similar a la de los valores observados y utilizarlos para llenar los huecos. Esto se puede hacer utilizando funciones de generación de números aleatorios en el lenguaje de programación que estés utilizando.

Relleno con valores promedio: Puedes calcular el promedio de los valores observados en la variable y utilizar ese valor para rellenar los huecos. Esto proporciona una estimación neutral basada en los valores existentes.

Relleno con valores de interpolación: Si los datos tienen una estructura temporal o espacial, puedes utilizar técnicas de interpolación para estimar los valores faltantes. Algunos métodos de interpolación comunes incluyen la interpolación lineal, spline cúbico o el método de vecinos más cercanos.

Estas técnicas se han utilizado en los algoritmos para comprobar que producirían mejores resultados que dejando los valores vacíos. Sin embargo, no produjeron un buen resultado. La justificación es que rellenar los huecos con valores aleatorios puede alterar las características que define al paciente de su grupo en concreto. Con el posterior análisis de superposición de variables está demostrado que un relleno aleatorio no sería adecuado para nuestra base de datos.

Debido a que los datos no tienen ningún tipo de estructura espacial o temporal no hemos optado por técnicas con valores de interpolación. La interpolación se basa en la existencia de patrones o tendencias, lo que no es apropiado cuando se asume que la falta de datos es completamente aleatoria.

En resumen, a los efectos de este trabajo, hemos comprobado que estos métodos no contribuyen en medida alguna a la mejora de los resultados, por lo que los métodos finalmente aplicados han incorporado algoritmos que permitan trabajar con valores Nan (not a number), evitando cualquier estrategia de rellenado de huecos.

4.1.2 Preprocesado de la base de datos

Como ya se ha comentado anteriormente partimos de 4 tablas de variables en nuestra base de datos. Todos los métodos de Machine Learning y Deep Learning actuales precisan de una matriz numérica como valor de “input”. Por ese motivo se han agrupado las 4 tablas en una sola de manera que todos los datos para cada paciente se encuentren en una única fila.

Como punto de partida para agrupar todos los datos hemos elegido los electrocardiogramas (en total para 1878 pacientes). A partir de aquí elegiremos la analítica y ecocardiograma más cercanas a la fecha en la que se realizó cada electrocardiograma.

Es importante mencionar que hemos puesto un límite al intervalo entre el cual se realizó un electrocardiograma para cada analítica y ECO correspondiente con el fin de proporcionar un conjunto muestral con datos coherentes en cuanto a la correspondencia temporal de las muestras de las distintas tablas origen. Por ejemplo, no sería correcto asociar una analítica o una ECO que se hizo 10 años después al electrocardiograma que estamos tomando como referencia. El tiempo para el cual se ha decidido que un electrocardiograma con otras medidas es coherente y se ha definido en un año. Este valor ha sido sugerido por el equipo clínico participante en el proyecto.

El pseudocódigo seguido con todo el preprocesado se muestra en la Ilustración 9 y se puede dividir en 5 pasos.

```
Preprocesado de la Base de datos

Cargamos la base de datos

1.- Búsqueda de índices

    # Funciones de búsqueda de índices para los cuales hay una o varias ECOs o analíticas dentro de el margen de 365 días
    # Para la id del paciente de cada ECG se buscan coincidencias entre analíticas y ECOs. Apuntamos el índice si hay coincidencia
    # La función también ordena los índices en función de su cercanía temporal con cada electrocardiograma.
    Aplicar BuscarECO
    Aplicar BuscarAnalítica

    # En el caso de que en el margen de un año no haya ni eco ni analítica para cierto electrocardiograma
    Creación de una fila vacía de ECO
    Creación de una fila vacía de analítica

2.- Creación de la matriz de datos

    Se comprueba para cada electrocardiograma si ha encontrado una eco o analítica, apuntando el número de casos encontrados
    ¿Encuentra ECO/Analitica?
        Si -> Se rellena con los datos correspondientes
        No -> Se rellena con la fila ECO/Analitica vacía

3.- Pulido de la base de datos

    Variables no numéricas -> Numéricas
    Traducimos las fechas a num_dias
    Borrar pacientes sin Analitica ni ECO

4.- Relleno de datos vacíos

    Relleno de huecos a partir de otras Analíticas/ECOs
    Borrar variables vacías

5.- Añadir a la base de datos la tabla de información

    Añadir información grupo por electrocardiograma
    OneHotEncoding variable Grupo0123
```

Ilustración 9 Preprocesado base de datos

1.- **Búsqueda de índices**: En este paso se utiliza un bucle se encarga de buscar para cada identidad de cada ECG los ECOs cuya identidad coinciden, además los filtra según la variable número de días (en este caso 365 para filtrar un año), de manera que si la fecha de cierta ECO se hizo con una diferencia mayor a la de esta variable entonces será descartada. Una vez obtenidos todos los índices de los ECOs que cumplen estas condiciones se ordenan y guardamos los índices ordenados.

Puede existir la posibilidad de que al buscar una ECO o una analítica no se encuentre ninguna. En caso de que esto ocurra se añade una fila vacía de ECO o analítica debido a que todas las filas deben tener el mismo número de columnas. Por otro lado, igualmente se ha creado una columna que indica el número de analíticas/ECOs que se han encontrado.

2.- **Creación de la matriz de datos**: Como tenemos con el paso anterior el número de índices y además ordenados podemos rellenar para cada ECG con la analítica/ECO del primer índice correspondiente.

Creamos dos bucles, uno para los ECOs y otro para las analíticas. Extrayendo el primer índice se obtiene la analítica/ECO más cercana y se añade a la fila del ECG correspondiente.

En la Ilustración 10 se muestran un histograma con el número de analíticas y ECOs encontradas en total. Observar que en muchos casos no encontramos ECO (600 casos) o analítica (500 casos) mientras que para otro paciente se han llegado a encontrar hasta 98 analíticas.

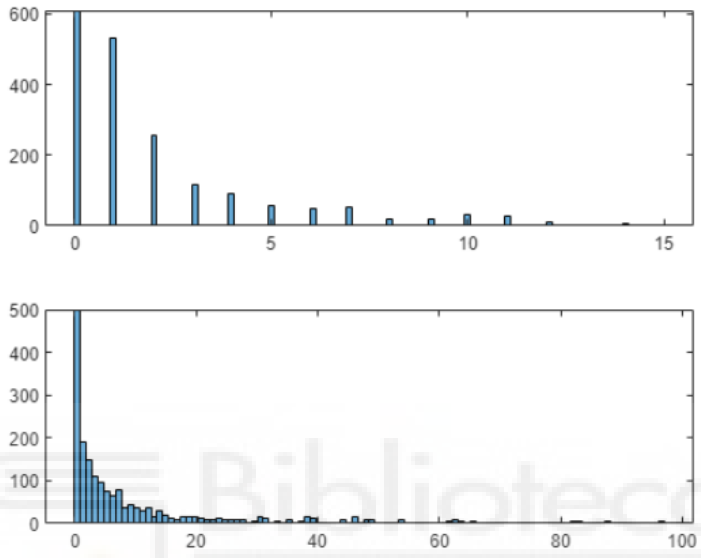

Ilustración 10 Ecos encontrados (arriba). Analíticas encontradas (abajo)

3.- Ajuste de la base de datos: En la base de datos original existían muchos datos no numéricos, como es el caso de las fechas. Debido a que las fechas son un tipo de dato distinto se deben pasar a numérico. Para este trabajo se ha transformado las fechas asociando cada una de ellas al número de días trascurridos desde el uno de enero de 1900.

Por otra parte, hay bastantes casos de pacientes que no tienen ni una analítica ni ninguna ECO asignada. En estos casos debido al tipo de datos con los que estamos tratando (MCAR) hemos optado por excluirlos del espacio muestral de nuestro análisis.

4.- Relleno de datos vacíos: El número de huecos en total de la base de datos en este punto es de 50.527 huecos. Este valor es muy alto y se puede reducir con el siguiente método: Como se puede observar en los histogramas anteriores por norma general el número de analíticas y ECOs encontrados es mayor que 1, por lo que se pueden utilizar esas ECOs y analíticas para rellenar los huecos vacíos en caso de que la primera ECO o analítica no hubiese rellenado algo que otras ECOs o analíticas si tengan. Se les da prioridad a los datos cuanto más cercanos son al ECG, para ello tenemos todos los índices de las analíticas y ecos encontradas para cada ECG ordenados.

Utilizando dos bucles, uno para analíticas y otro para ECOs se obtiene en este punto 38.072 huecos vacíos. Por otro lado, y a pesar de todo, hay variables que están completamente vacías en todos los casos y evidentemente no van a aportar ningún resultado positivo. Estas variables han sido excluidas de este estudio. Tras su eliminación, nuestra base de datos presenta 27.073 huecos.

5.- Añadir a la base de datos la tabla información:

Con esta tabla se relaciona cada identidad de paciente con su respectivo electrocardiograma de manera que donde se aporta información de su nivel de afectación estratificado (grupo 0 hasta el grupo 3). A esta variable se le ha aplicado "One hot encoding" para tener esta variable en formato binario.

La Ilustración 11 muestra el número de pacientes en cada grupo. La abismal diferencia entre pacientes afectados y el resto podría llevar a dificultades de convergencia de los modelos a ser implementados. Este hecho, y sus alternativas de gestión, será comentado con posterioridad

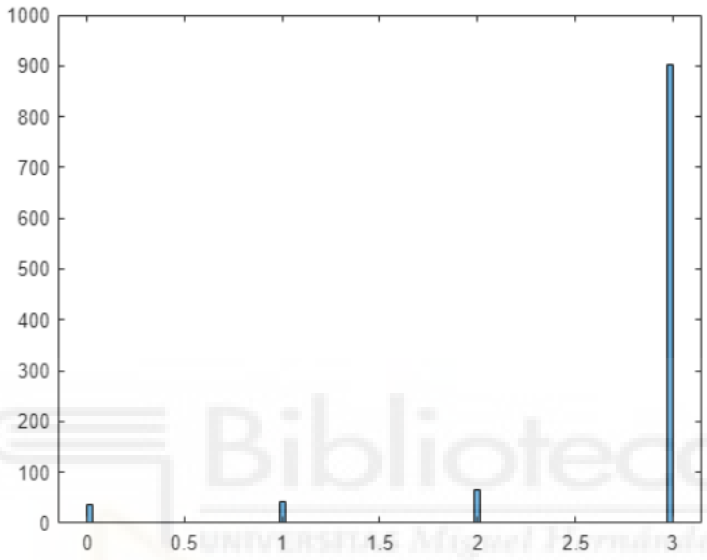

Ilustración 11 Número de pacientes por grupo

4.2 Los métodos de ML y DL que vamos a usar

4.2.1 Principal Component Analysis

El análisis de componentes principales (PCA) es una técnica utilizada para reducir la dimensionalidad de dichos conjuntos de datos, mejorando la interpretación y reduciendo la pérdida de información. Esto se hace creando nuevas variables no correlacionadas ordenadas en función de su varianza. Encontrar estas nuevas variables, los componentes principales, implica resolver el problema de valores propios/vectores propios, y las nuevas variables están definidas por el conjunto de datos en cuestión, no a priori, lo que convierte a PCA en una técnica de análisis de datos adaptativa. También es adaptable en el sentido de que se desarrollan variantes tecnológicas para adaptarse a muchos tipos y estructuras de datos diferentes.

Los grandes conjuntos de datos son cada vez más comunes en muchas disciplinas. La interpretación de tales conjuntos de datos requiere métodos que reduzcan drásticamente su dimensionalidad de una manera interpretable para que se conserve la mayor parte de la información contenida en los datos. Se han desarrollado muchos métodos para este propósito,

pero el análisis de componentes principales (PCA) es uno de los más antiguos y utilizados. Su idea es simple: reducir la dimensionalidad del conjunto de datos conservando la mayor "variabilidad" (es decir, estadísticas) como sea posible.

Es decir, "preservar tanta variabilidad como sea posible" significa encontrar nuevas variables que sean funciones lineales de esas variables iniciales del conjunto de datos original, ordenadas por varianza y que no estén correlacionadas. Para encontrar estas nuevas variables, conocidas como componentes principales (PC), se debe resolver el problema de valores propios/vectores propios. La literatura más antigua sobre PCA se escribió en 1901, pero no fue hasta que las computadoras electrónicas estuvieron ampliamente disponibles décadas después que se volvió computacionalmente factible usarlo en conjuntos de datos muy pequeños. Desde entonces, el uso ha aumentado y se han desarrollado un gran número de variantes en muy diferentes disciplinas.

Aunque a menudo se supone una distribución normal multivariante (gaussiana) de un conjunto de datos con fines de inferencia, PCA como herramienta descriptiva no requiere suposiciones de distribución y, por lo tanto, es en gran medida un método heurístico adaptativo que se puede aplicar a datos numéricos de varios tamaños. De hecho, se han desarrollado muchas adaptaciones del método básico para diferentes tipos y estructuras de datos. Algunos métodos proporcionan versiones simplificadas de PC para facilitar la interpretación. La explosión de conjuntos de datos muy grandes en áreas como el análisis de imágenes o el análisis de datos web ha llevado a avances metodológicos significativos en el análisis de datos, a menudo como resultado de PCA [15].

4.2.1.1 Algoritmo de PCA

El algoritmo de PCA se compone de 5 pasos basados en las ideas matemáticas de varianza y covarianza además de vectores y valores propios. Los pasos en orden son los siguientes: Normalización de los datos, cálculo de la matriz de covarianza, cálculo de los vectores y valores propios, selección de componentes principales y proyección de los datos.

El propósito de la normalización de los datos es estandarizar el rango con el fin de que cada una de las variables iniciales continuas contribuya igualmente al análisis [16].

En particular, es muy importante realizar la estandarización antes de PCA porque PCA es muy sensible a la varianza de las variables originales. Es decir, si hay una gran diferencia entre los intervalos de las variables originales, la variable con el intervalo mayor dominará a la variable con el intervalo menor (por ejemplo, la variable con el intervalo entre 0 y 100 dominará a la variable con el intervalo entre 0 y 1), lo que dará lugar a resultados sesgados. Entonces, convertir los datos a una escala comparable resuelve este problema.

Matemáticamente, esto se puede hacer restando la media de cada variable y dividiendo por la desviación estándar.

$$z = \frac{valor - media}{desviación\ estándar} \qquad (7)$$

Una vez la estandarización está hecha todas las variables serán transformadas a la misma escala.

En cuanto al cálculo de la matriz de covarianza, el propósito de este paso es entender cómo las variables en el conjunto de datos de entrada difieren de su media, o en otras palabras, ver si existe alguna relación entre ellas. Porque a veces las variables están altamente correlacionadas de una

manera que contiene información redundante. Por lo tanto, para identificar estas correlaciones, calculamos la matriz de covarianza.

La matriz de covarianza es una matriz p × p simétrica (donde p es el número de dimensiones) cuyas entradas son las covarianzas asociadas a todos los posibles pares de variables originales. Para un conjunto de datos tridimensionales con 3 variables x, y, z, la matriz de covarianza es una matriz de datos de 3 × 3, por ejemplo:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

(8)

Como la covarianza de una variable consigo misma es su varianza (Cov(a,a)=Var(a)), en la diagonal principal (de arriba a la izquierda a abajo a la derecha) tenemos la varianza de cada variable original. Y dado que las covarianzas son conmutativas (Cov(a,b)=Cov(b,a)), entonces las entradas de la matriz de covarianza son simétricas con respecto a la diagonal principal, lo que significa que las partes superior e inferior del triángulo son iguales.

El tercer paso es el cálculo de los valores y vectores propios, los vectores y valores propios son los conceptos de álgebra lineal que se necesitan calcular a partir de la matriz de covarianza para determinar los componentes principales de los datos.

El número de valores y vectores propios es igual a la dimensión de los datos. Por ejemplo, un conjunto de datos 3D tiene 3 variables, por lo que hay 3 vectores propios y 3 valores propios correspondientes. Los vectores propios de la matriz de covarianza son en realidad las direcciones de los ejes con más varianza (más información), a los que llamamos componentes principales. Mientras que los valores propios son simplemente coeficientes agregados a los vectores propios, que indican la cantidad de varianza en cada componente principal. Al clasificar sus vectores propios en orden de sus valores propios, de mayor a menor, obtiene los componentes principales en orden de importancia.

Calcular los vectores propios y ordenarlos en orden descendente de valores propios nos permite encontrar los componentes más importantes en orden de importancia. En este paso, se elige si se quiere mantener todos estos componentes o descartar los más pequeños (con valores propios bajos) y crear una matriz de vectores con el resto, que llamamos vectores propios.

Entonces, un vector propio es solo una matriz cuyas columnas son los vectores propios de los componentes que se deciden mantener. Esto lo convierte en el primer paso en la reducción de la dimensionalidad, porque si se eligen mantener solo p vectores propios (componentes) fuera de n, el conjunto de datos final tendrá solo p dimensiones.

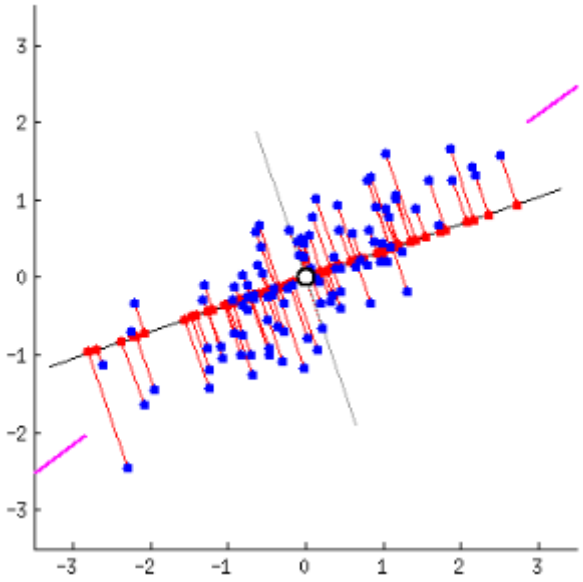

Ilustración 12 Reducción de dimensionalidad mediante PCA

En el paso anterior, no se realiza ningún cambio en los datos además de la estandarización, solo selecciona los componentes principales y crea los vectores propios, pero el conjunto de datos de entrada siempre mantiene los ejes originales (es decir, las variables originales).

En el paso final, el objetivo es reorientar los datos de los ejes originales a los ejes representados por los componentes principales utilizando los vectores propios formados por los vectores propios de la matriz de covarianza (de ahí el nombre de análisis de componentes principales). Esto se puede hacer multiplicando la transposición del conjunto de datos original por la transposición de los vectores propios.

4.2.2 Autoencoders

Los Autoencoder (AE) son redes neuronales diseñadas para obtener a la salida los mismos elementos que dispone a su entrada, mientras en este proceso el conjunto de variables se reduce significativamente en su espacio intermedio (espacio latente). Funcionan comprimiendo la entrada en una representación latente del espacio y luego reconstruyendo la salida de esa representación. Este tipo de red consta de 3 partes, un codificador, el espacio latente, y un decodificador.

El codificador es la parte de la red que comprime la entrada en una representación de espacio latente. Esto se puede representar mediante la función de codificación h=f(x). El propósito de la parte del decodificador es reconstruir la entrada a partir de la representación del espacio latente.

Esto se puede representar mediante la función de decodificación r=g(h). [4]

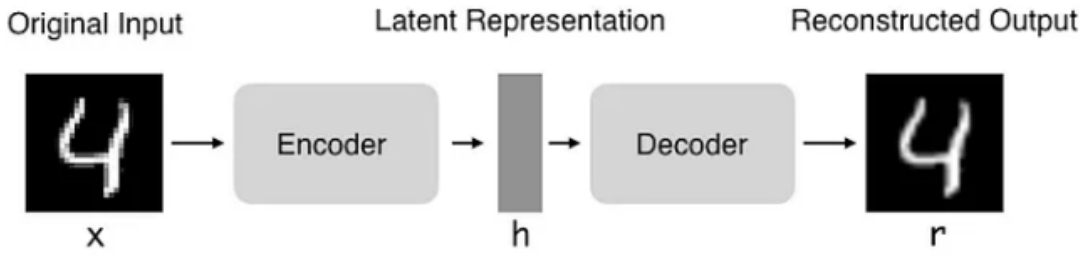

Ilustración 13 Arquitectura de un autoencoder

Para comprender mejor los códigos automáticos, debemos referirnos a su arquitectura típica. En la Ilustración 13, podemos visualizar el primer componente principal del código automático es tres: códigos, representación funcional potencial y decodificador. El COD y el decodificador son solo funciones, y los nombres de funciones potenciales generalmente se refieren al número real . Con todo, esperamos que los códigos automáticos puedan reconstruir bien la entrada. Sin embargo, al mismo tiempo, se debe crear una representación latente útil y significativa (la salida de la parte del codificador en la Figura 1). Por ejemplo, las posibles características de los dígitos escritos a mano podrían ser la cantidad de líneas necesarias para escribir cada dígito o el ángulo de cada línea y cómo se unen. Aprender a escribir números no requiere necesariamente aprender el valor de escala de grises de cada píxel en la imagen de entrada. Por supuesto, las personas no aprenden a escribir llenando píxeles con valores de escala de grises. Cuando aprendemos, adquirimos la información básica que nos permite resolver problemas (como escribir números). Esta representación latente (cómo se escribe cada número) es útil para varias tareas (como extraer características que luego se pueden usar para clasificar o agrupar) o simplemente para comprender las propiedades básicas de un conjunto de datos. [5]

4.2.2.1 Función de pérdidas

Como con cualquier modelo de red neuronal, necesitamos una función de pérdida para minimizarlo. Estas funciones de pérdida deberían medir cuánta diferencia hay entre la entrada xi y la salida ~xi.

Se utiliza una función de pérdida para ajustar los pesos de la red neuronal para que la reconstrucción sea lo más cercana posible a los datos de entrada. En general, el objetivo es minimizar la función de pérdida para que la red pueda aprender a comprimir y reconstruir los datos de manera eficiente.

Es importante señalar que la elección de la función de pérdida en el autocodificador puede afectar significativamente el rendimiento y la calidad de la reconstrucción. Por ejemplo, la función de pérdida de distancia euclidiana puede ser apropiada para datos continuos y lineales, mientras que la función de pérdida de desviación KL puede ser más apropiada para datos categóricos o discretos.

En general, la elección de la función de pérdida depende del tipo de datos utilizados y de los objetivos específicos del modelo. Elegir la función de pérdida adecuada es importante para garantizar un buen rendimiento del modelo. A continuación, se muestran algunos ejemplos utilizados como funciones de pérdidas en los autoencoders:

Error cuadrático medio (MSE): Es la función de pérdida más común en los codificadores automáticos y se utiliza para datos continuos. MSE mide la distancia euclidiana entre los datos de entrada y la reconstrucción producida por el modelo. Esta función de pérdida penaliza los grandes errores y se define como el cuadrado medio de la diferencia entre los datos originales y los reconstruidos.

$$L_{\text{MSE}} = \text{MSE} = \frac{1}{M}\sum_{i=1}^{M} |\mathbf{x}_i - \tilde{\mathbf{x}}_i|^2$$

(9)

Error absoluto medio (MAE): Es similar a MSE, pero en lugar de calcular el cuadrado medio de la diferencia, calcula la media de la diferencia absoluta. Esta función de pérdida se usa a menudo para datos atípicos porque es menos sensible a errores grandes.

$$\mathbf{EAM} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

(10)

Entropía cruzada binaria (BCE): esta es una función de pérdida comúnmente utilizada en problemas de clasificación binaria que intentan predecir la probabilidad de que una muestra pertenezca a una de dos clases posibles.

BCE se utiliza cuando la variable de destino tiene un valor binario (0 o 1). Esta función de pérdida mide la diferencia entre la distribución de probabilidad de los valores reales y la distribución de probabilidad predicha por el modelo. Es decir, BCE mide la diferencia entre dos distribuciones de probabilidad y penaliza la diferencia en la distribución de probabilidad entre los datos originales y las predicciones del modelo.

$$L_{\text{CE}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{n}[x_{j,i}\log\tilde{x}_{j,i} + (1 - x_{j,i})\log(1 - \tilde{x}_{j,i})]$$

(11)

4.2.2.2 Error de reconstrucción

El error de reconstrucción (RE) es una métrica que indica la capacidad (o capacidad) de un codificador automático para reconstruir las observaciones de entrada xi. El RE más típico es MSE.

$$\text{RE} \equiv \text{MSE} = \frac{1}{M}\sum_{i=1}^{M} |\mathbf{x}_i - \tilde{\mathbf{x}}_i|^2$$

(12)

Esto se puede calcular fácilmente. RE se usa a menudo cuando se realiza la detección de anomalías con codificadores automáticos. Los errores de restauración tienen una explicación sencilla e intuitiva. Si RE es significativo, el codificador automático no puede reconstruir bien la entrada, mientras que, si es pequeño, la reconstrucción tiene éxito. La *Ilustración 14* muestra un ejemplo de error de reconstrucción de magnitud cuando el codificador automático intenta reconstruir una imagen.

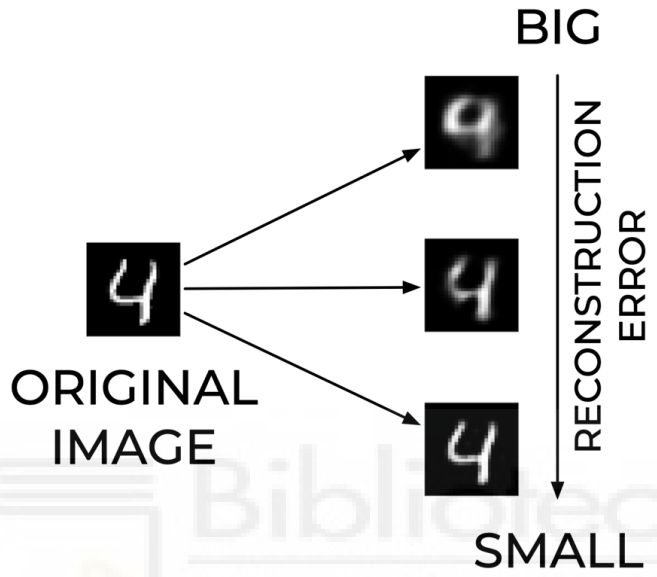

Ilustración 14 Un ejemplo de error de reconstrucción grande y pequeño cuando un codificador automático intenta reconstruir una imagen. [17]

4.2.2.3 Reducción de dimensionalidad

La reducción de la dimensionalidad en los codificadores automáticos se basa en el hecho de que el codificador transforma los datos de alta dimensión en una representación de baja dimensión. Por lo general, las funciones de codificación y decodificación se modelan mediante redes neuronales, mientras que la reducción de la dimensionalidad se logra mediante arquitecturas que producen imágenes más pequeñas en el centro de la red.

La formulación utilizada para la reducción de dimensionalidad en autocodificadores se basa en la función de pérdida utilizada para entrenar la red.

Con el método del cuello de botella, los rasgos latentes tendrán una dimensión más pequeña que las dimensiones de las observaciones de entrada. La parte del codificador (si está entrenada) naturalmente (por diseño) realiza una reducción de dimensionalidad que produce números reales. Las características latentes se pueden utilizar para diversas tareas, como la clasificación o la agrupación. Puede ser interesante señalar algunas ventajas del uso de codificadores automáticos para la reducción de la dimensionalidad sobre los métodos PCA más clásicos. Un codificador automático tiene una gran ventaja desde el punto de vista computacional: puede manejar grandes cantidades de datos de manera eficiente porque su entrenamiento se puede realizar en mini lotes, mientras que PCA, uno de los algoritmos de reducción de dimensionalidad más utilizados,

requiere que se realice sus cálculos en todo el conjunto de datos. PCA es un algoritmo que proyecta un conjunto de datos en los vectores propios de su matriz de covarianza, produciendo una transformación lineal de las características. Los codificadores automáticos son más flexibles y tienen en cuenta las transformaciones no lineales de funciones. En muchos casos, esto es computacionalmente inviable y el algoritmo no escala a medida que crece el conjunto de datos. Esto puede parecer trivial, pero en muchas aplicaciones prácticas, la cantidad de datos y la cantidad de funciones son tan grandes que PCA no es práctico desde el punto de vista computacional [17].

4.2.3 tSNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) es una técnica de reducción de dimensionalidad no lineal que se utiliza para visualizar grandes conjuntos de datos en dos o tres dimensiones. A diferencia de otras técnicas de reducción de dimensionalidad, t-SNE está diseñada para manejar datos no lineales y preservar mejor las relaciones locales entre los puntos de datos.

La técnica de t-SNE se basa en la idea de que los puntos de datos similares en el espacio de alta dimensión deben estar cercanos en el espacio de baja dimensión. t-SNE utiliza una distribución de probabilidad para medir las similitudes entre los puntos de datos en el espacio de alta dimensión y una distribución de probabilidad gaussiana para medir las similitudes entre los puntos de datos en el espacio de baja dimensión. El algoritmo busca minimizar la diferencia entre estas dos distribuciones de probabilidad, lo que resulta en una visualización que preserva mejor las relaciones locales entre los puntos de datos.

Una de las ventajas de t-SNE es que puede manejar grandes conjuntos de datos y preservar las estructuras complejas y no lineales de los datos. Sin embargo, t-SNE puede ser computacionalmente costoso y requiere cierta configuración para obtener los mejores resultados.

t-SNE se utiliza comúnmente en aplicaciones de análisis de datos y aprendizaje automático, como la visualización de datos de imágenes, la agrupación de genes en datos de expresión génica, la clasificación de documentos en datos de texto, entre otros.

En resumen, t-SNE es una técnica de reducción de dimensionalidad no lineal que se utiliza para visualizar grandes conjuntos de datos en dos o tres dimensiones. Utiliza una distribución de probabilidad para medir las similitudes entre los puntos de datos en el espacio de alta dimensión y una distribución de probabilidad gaussiana para medir las similitudes entre los puntos de datos en el espacio de baja dimensión. El algoritmo busca minimizar la diferencia entre estas dos distribuciones de probabilidad, lo que resulta en una visualización que preserva mejor las relaciones locales entre los puntos de datos.

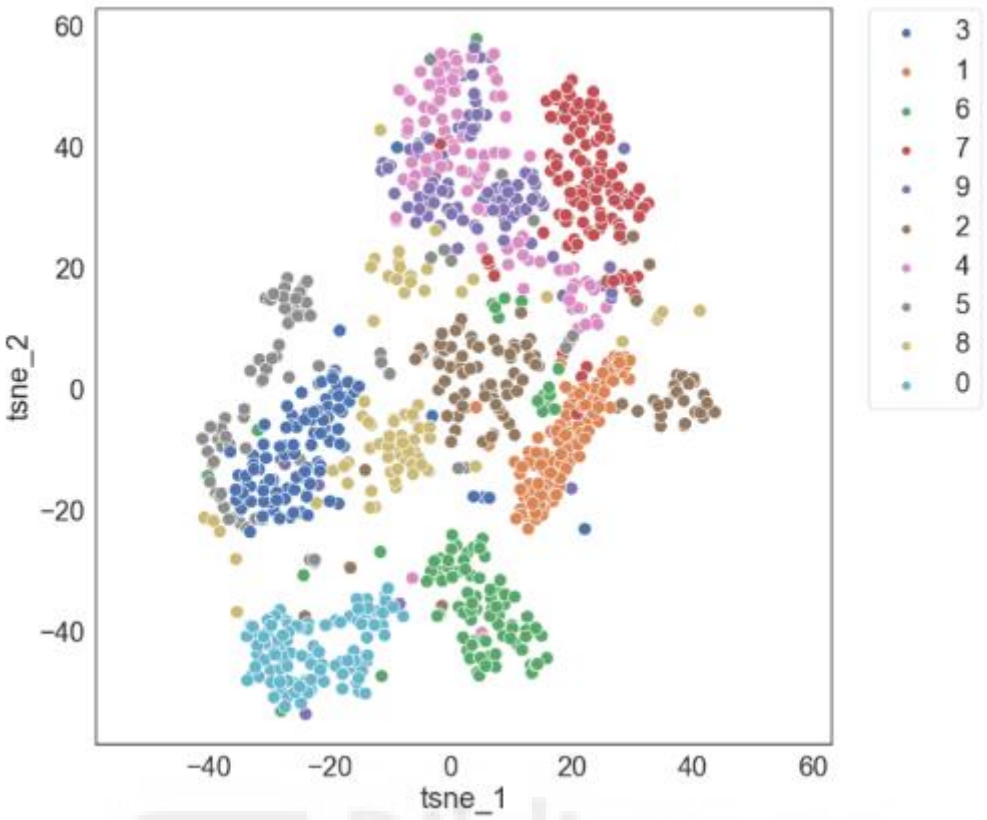

Ilustración 15 Ejemplo aplicación 2D tSNE

4.2.3.1 Algoritmo de tSNE

El método tSNE interpreta la distancia total entre puntos de datos en un espacio de alta dimensión como una distribución de probabilidad conjunta simétrica P. También calcula una distribución de probabilidad conjunta Q que caracteriza la similitud en un espacio de baja dimensión. El objetivo es obtener una representación llamada incrustación en un espacio de baja dimensión donde Q representa exactamente a P. Esto se logra optimizando una posición en el espacio de baja dimensión para minimizar la función de costo C dada por la divergencia de Kullbeck-Leibler (KL) entre las distribuciones de probabilidad conjunta P y Q [18] :

$$C(P,Q) = KL(P||Q) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} p_{ij} \ln\left(\frac{p_{ij}}{q_{ij}}\right)$$

(13)

Dados dos puntos de datos x_i y x_j en el conjunto de datos X = {fx1...x_N}, la probabilidad p_{ij} modela la similitud entre estos puntos en un espacio de alta dimensión. En este sentido, para cada punto cuya varianza σ_i se define como una función de densidad local en un espacio de alta dimensión, se elige un núcleo gaussiano P_i, luego p_{ij} se describe de la siguiente manera [18]:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N},$$

Donde $$p_{j|i} = \frac{\exp(-(||\mathbf{x}_i - \mathbf{x}_j||^2)/(2\sigma_i^2))}{\sum_{k \neq i}^{N} \exp(-(||\mathbf{x}_i - \mathbf{x}_k||^2)/(2\sigma_i^2))}.$$

(14)

La razón para dividir por la suma de todos los demás puntos con centro en x_i es que es posible que tengamos que tratar con grupos de diferentes densidades. Para explicar esto, volvamos al ejemplo de la Ilustración 16. Como puede ver, el racimo naranja es menos denso que el racimo azul. Entonces, si solo calculamos la similitud entre cualquier otro punto usando un Gaussiano, veremos que la similitud entre el punto naranja y el punto azul es baja. En nuestro resultado final, no nos importa que algunos grupos tengan densidades diferentes, solo queremos tratarlos como grupos, así que hacemos esta normalización. [19]

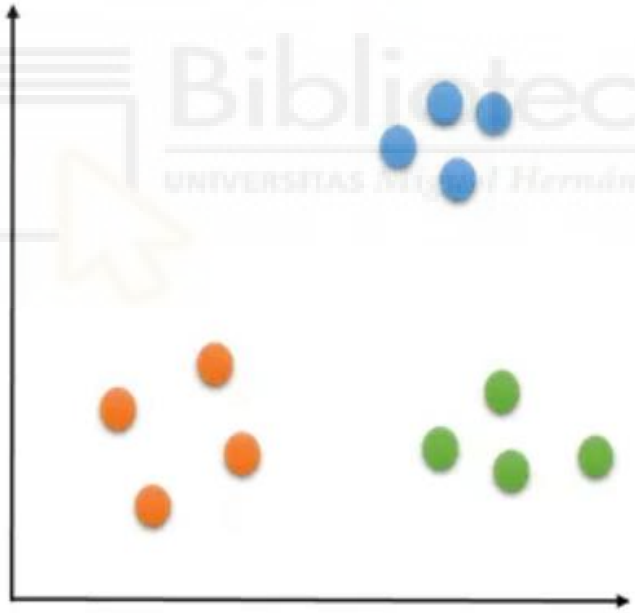

Ilustración 16 Puntos de datos originales en el espacio de alta dimensión [19]

Las medidas p_{jji} se pueden considerar como mediciones relativas de similitud basada en la vecindad local de los puntos de datos xi. El valor de perplejidad μ es un parámetro definido por el usuario que describe el número de vecinos válidos considerados para cada punto de datos. Elige valores de σ_i tales que para μ fija y cada i [18]:

$$\mu = 2^{-\sum_j^N p_{j|i} \log_2 p_{j|i}}$$

(15)

Calcula la distribución de probabilidad conjunta en el espacio de baja dimensión Q utilizando la distribución t-Student con un grado de libertad, donde se optimizarán las ubicaciones de los puntos de datos. Dados dos puntos de baja dimensión y_i e y_j, la probabilidad q_{ij} que describe su similitud viene dada por [18]:

$$q_{ij} = \left((1 + ||\mathbf{y}_i - \mathbf{y}_j||^2) Z\right)^{-1},$$

$$\text{with} \quad Z = \sum_{k=1}^{N} \sum_{l \neq k}^{N} (1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}$$

(16)

El gradiente de divergencia de Kullbeck-Leibler entre P y Q se utiliza para minimizar C Ecuación (13). Representa el cambio de posición de un punto de baja dimensión en cada paso del descenso del gradiente, dado por la siguiente fórmula [18]:

$$\frac{\delta C}{\delta \mathbf{y}_i} = 4 \sum_{i=1}^{N} (F_i^{\text{attr}} - F_i^{\text{rep}})$$

$$= 4 \sum_{i=1}^{N} (\sum_{j \neq i}^{N} p_{ij} q_{ij} Z(\mathbf{y}_i - \mathbf{y}_j) - \sum_{j \neq i}^{N} q_{ij}^2 Z(\mathbf{y}_i - \mathbf{y}_j)$$

(17)

4.2.4 UMAP

4.2.4.1 Fundamentos teóricos de UMAP

UMAP (Uniform Manifold Aproximation and Projection) es un método de reducción de dimensionalidad no lineal para la visualización y el análisis de grandes conjuntos de datos. A diferencia de otros métodos de reducción de dimensionalidad como PCA y t-SNE, UMAP está diseñado para manejar datos complejos de alta dimensión y puede preservar mejor la estructura no lineal y las relaciones locales entre los puntos de datos. UMAP se basa en la idea de que los

datos se distribuyen uniformemente en una variedad, una estructura matemática que describe cómo se conectan los puntos de datos entre sí. El algoritmo utiliza técnicas topológicas y geométricas para aproximar la variedad subyacente y proyectar los datos en un espacio dimensional más bajo mientras conserva la estructura y las relaciones locales. Una de las principales ventajas de UMAP es su velocidad y escalabilidad. A diferencia de otros métodos de reducción de dimensionalidad no lineal, UMAP se puede aplicar de manera efectiva a grandes conjuntos de datos y puede manejar fácilmente datos de alta dimensión [8].

UMAP se usa ampliamente en aplicaciones de análisis de datos y aprendizaje automático, como visualización de datos de imágenes, agrupación de genes en datos de expresión génica, clasificación de documentos en datos de texto, etc.

En resumen, UMAP es un método de reducción de dimensionalidad no lineal para la visualización y análisis de grandes conjuntos de datos. Utiliza técnicas topológicas y geométricas para aproximar la variedad subyacente y proyectar los datos en un espacio dimensional más bajo mientras conserva la estructura y las relaciones locales.

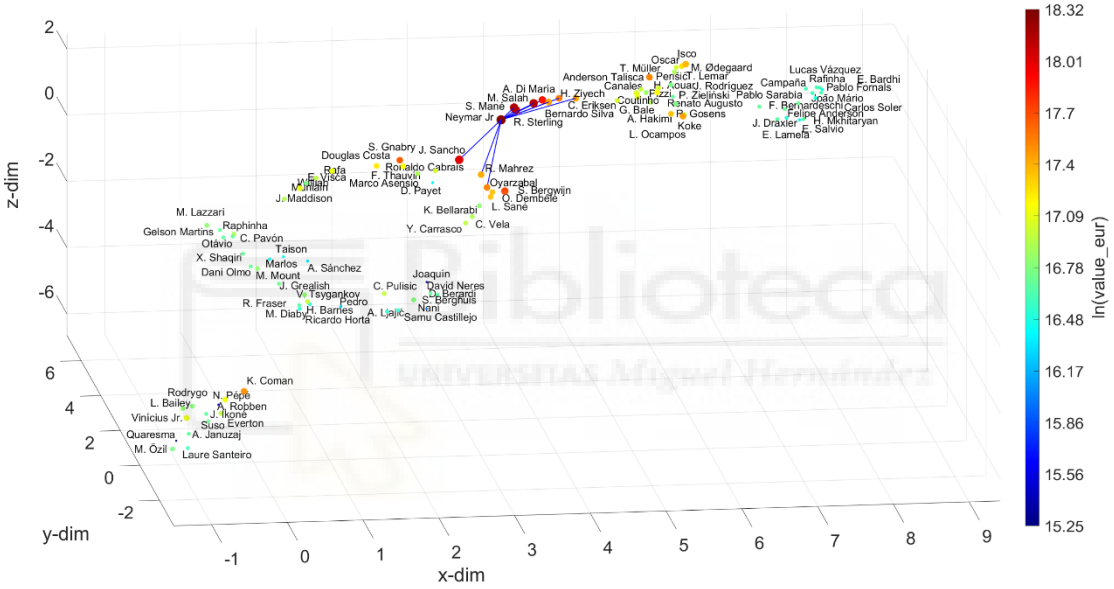

Ilustración 17 Ejemplo aplicación UMAP

4.2.4.2 Aproximación de la variedad subyacente

En un nivel alto, UMAP utiliza aproximaciones de variedades locales y une sus representaciones de conjuntos simplistas "fuzzy" locales para construir una representación topológica de los datos de alta dimensión. Dada una representación de datos de baja dimensión, se puede utilizar un procedimiento similar para construir una representación topológica equivalente. Luego, UMAP optimiza la disposición de las representaciones de datos en un espacio de baja dimensión para minimizar la entropía cruzada entre dos representaciones topológicas.

La construcción de representaciones topológicas "fuzzy" se puede dividir en dos problemas: aproximar una variedad en la que se supone que se encuentran los datos; y construir una representación de conjunto simplista "fuzzy" de la variedad aproximada. Primero se analiza el

método de aproximación de la variedad para los datos de origen. A continuación, se discute cómo construir una estructura de conjunto simplicial "fuzzy" a partir de la aproximación múltiple.

El algoritmo de representación de UMAP busca una representación de una basa de datos dada (D) en $\mathbb{R}^N$ un espacio de $\mathbb{R}^M$ dimensión inferior. Pensamos en los puntos de datos como extraídos de una variedad riemanniana M y luego mapeados mediante alguna incrustación $\varphi : M \rightarrow \mathbb{R}^N$. Consulte la Ilustración 18 para ver un esclarecimiento de la configuración.

Un pensamiento que podríamos tener es reconstruir M, luego encontrar un buen mapa de M en $\mathbb{R}^M$. Para hacer esto, asumimos que D se extrae uniformemente de M, como en el ejemplo de la Ilustración 18. (Tenga en cuenta que partes de M pueden estirarse o comprimirse bajo la incrustación en $\mathbb{R}^N$, por lo que esto no implica que los datos se extraigan uniformemente). distribuidos en $\mathbb{R}^N$.) Esta suposición significa que D se aproxima bien a M. También supondremos que M está conectado localmente y que hay suficientes puntos en D para que ningún punto en D esté aislado en su propio componente conectado. Estos supuestos de conectividad implican que cada punto en D está conectado en M con su vecino más cercano en D → $\mathbb{R}^N$.

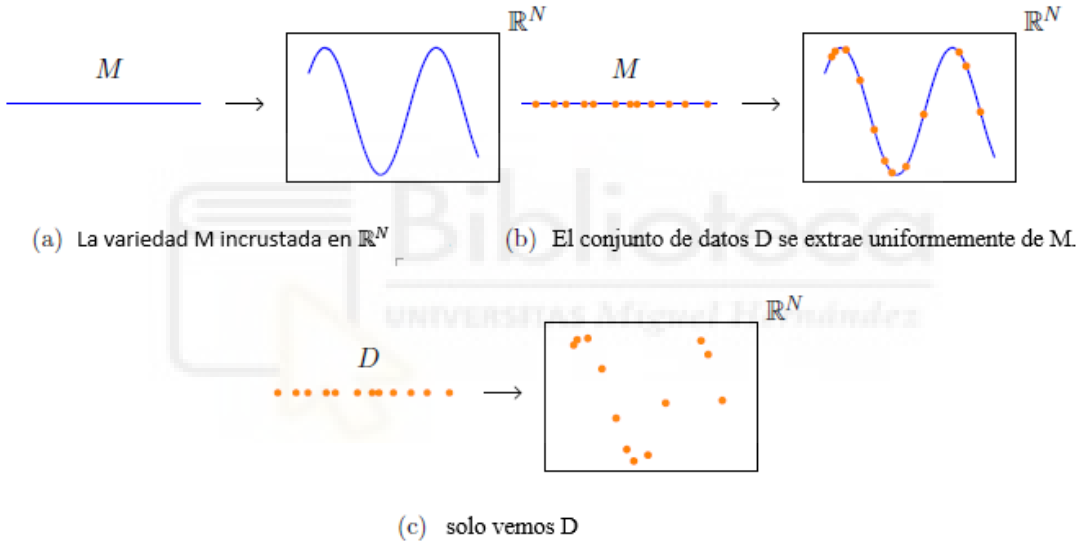

Ilustración 18 El conjunto de datos se extrae de una variedad incrustada en $\mathbb{R}^N$ [20]

Lema [21]: Sea (M,g) una variedad riemanniana incrustada en $\mathbb{R}^N$. Sea p ∈ M un punto. Suponga que g es localmente una constante. Sea B una bola en M, que contiene p, cuyo volumen es $\frac{\pi^{\frac{n}{2}}}{\Gamma(n/2+1)}$ con respecto a la métrica en M. Entonces la distancia del camino más corto en M desde "p" a un punto q ∈ B es $\frac{1}{r} d\mathbb{R}^n(p,q)$, donde r es el radio de B en $\mathbb{R}^n$ y $d\mathbb{R}^n(p,q)$ es la distancia desde p hasta q en $\mathbb{R}^n$.

Lo que esto hace por nosotros es que podemos aproximar la distancia alrededor de M cerca de nuestros puntos de datos escalando la distancia $\mathbb{R}^N$. Lo hacemos de esta manera. Dado que asumimos que nuestros datos están uniformemente distribuidos en M, cualquier esfera M de volumen fijo R debe contener el mismo número de puntos de datos. Trabajando hacia atrás, sea $N_k(x)$ la esfera $\mathbb{R}^N$que rodea el punto de datos x, que contiene sus k vecinos más cercanos $\mathbb{R}^N$ (con respecto a la distancia $\mathbb{R}^N$). Luego, para cualquier punto de datos xi, considere la vecindad de M correspondiente a $N_k(x_i)$ $\mathbb{R}^N$ (es decir, considere $\varphi -1$ ($N_k(x_i)$)). Esta esfera debe tener el

mismo volumen ya que hacemos el mismo procedimiento para cualquier otro punto de datos x_j. Por el lema, para k suficientemente pequeño, podemos aproximar la distancia de M desde x_i a uno de sus k vecinos más cercanos x_j como sigue. Se establece k como un hiperparámetro y se escribe $\{x_{i1}, \ldots, x_{ik}\}$ como el k-vecino más cercano de xi. Entonces podemos obtener del lema que la distancia de x_i a x_j en M es aproximadamente $\frac{1}{r_i} d\mathbb{R}^n(x_i, x_j)$, donde r_i es la distancia al k-vecino más cercano de xi. Para suavizar este valor y reducir el efecto de k-vecinos más cercanos que están lejos mientras que (k-1) vecinos están agrupados alrededor de x_i, tomamos r_i de tal manera que:

$$\sum_{j=1}^{k} e^{-\frac{\left|x_i - x_{i_j}\right|}{r_i}} = \log_2 k$$

(18)

Tenga en cuenta que, aunque se utilizó la métrica $\mathbb{R}^N$ para desarrollar la teoría, todavía se aplica a cualquier otra métrica. UMAP se puede usar con medidas de distancia personalizadas para que pueda manejar medidas de distancia entre datos categóricos y otros puntos de datos.

El primer paso es proporcionar algunos bloques de construcción de composición simples llamados *simple*. Geométricamente, el símplex es una forma muy sencilla de construir objetos de k dimensiones. Se llama símplex k-dimensional.

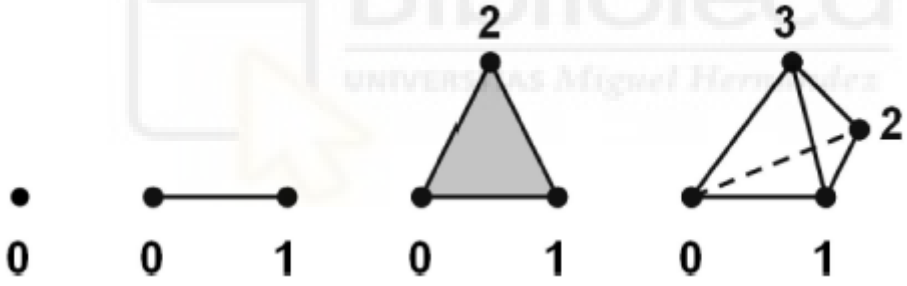

Ilustración 19 Ejemplos de n-simples para $0 \leq n \leq 3$ [22]

4.2.4.3 Análisis de datos topológicos y complejos simplistas

Un complejo simple es un método para construir espacios topológicos a partir de compuestos simples. Esto hace posible reducir la complejidad del procesamiento de geometrías continuas de espacios topológicos a tareas de conteo y combinación relativamente simples. Este enfoque del aprendizaje de la geometría y la topología será la base de nuestro enfoque del análisis de datos topológicos en general y de los métodos de reducción de la dimensionalidad en particular.

El primer paso es proporcionar algunos bloques de construcción de composición simples llamados *simple*. Geométricamente, el simplex es una forma muy sencilla de construir objetos de k dimensiones. Un simplex k-dimensional se llama k-simplex y se forma tomando la envolvente convexa de k+1 puntos independientes. Así, un 0-simple es un punto, un 1-simple es un segmento de línea (entre dos ceros simplex), un 2-simple es un triángulo (con tres 1-simples como "caras")

y un 3-simple es un tetraedro (con cuatro 2-simples como "caras"). Una construcción tan simple permite una fácil generalización a dimensiones arbitrarias.

Los simples pueden proporcionar bloques de construcción, pero para construir espacios topológicos interesantes necesitamos poder unir dichos bloques de construcción. Esto se puede hacer construyendo un *complejo simplicial*. Aparentemente, un complejo simplicial es un conjunto de simples pegados a lo largo de las caras. Más explícitamente, un complejo simplicial K es un conjunto de simples tales que cualquier cara de cualquier simple en K también está en K (asegurando que todas las caras existan), y la intersección de dos simples en K es una cara de ambos simples. Se puede construir una gran clase de espacios topológicos de esta manera, simplemente pegando simples de varias dimensiones a lo largo de sus caras.

Simplex puede proporcionar los bloques de construcción, pero para crear espacios topológicos interesantes, necesitamos poder combinar estos bloques. Esto se puede hacer creando un *complejo simple*. Claramente, un complejo simple es un conjunto de simples pegados en una cara. Más específicamente, un complejo simple K es un conjunto de simplificaciones tales que cualquier escala simplificada de K también es K (asegurando que todas las caras existan), y la intersección de dos simplificaciones en K es dos escalas del simplex A. De esta forma, se puede crear una gran clase de espacios topológicos simplemente colocando cuerpos simples de varios tamaños a lo largo de sus superficies.

4.2.4.4 Algoritmo de UMAP

Juntando todas estas piezas podemos construir el algoritmo UMAP. La primera fase consiste en construir una representación topológica “fuzzy”. La segunda fase es simplemente optimizar la representación de baja dimensión para tener una representación topológica “fuzzy” lo más cercana posible medida por entropía cruzada.

Al construir la representación topológica “fuzzy” inicial, podemos tomar algunos atajos. En la práctica, dado que las fuerzas de pertenencia de los conjuntos “fuzzy” decaen hasta ser extremadamente pequeñas, solo necesitamos calcularlas para los vecinos más cercanos de cada punto. En última instancia, eso significa que necesitamos una forma de calcular rápidamente (aproximar) los vecinos más cercanos de manera eficiente, incluso en espacios de gran dimensión. Podemos hacer esto aprovechando el algoritmo de descenso del vecino más cercano. Los cálculos restantes ahora solo se ocupan de los vecinos locales de cada punto y, por lo tanto, son muy eficientes.

Al optimizar incrustaciones de baja dimensión, nuevamente podemos usar algunos atajos. Podemos utilizar el descenso de gradiente estocástico para el proceso de optimización. Para aliviar el problema del descenso del gradiente, sería beneficioso que la función objetivo final fuera diferenciable. Podemos resolver este problema utilizando una aproximación suave de la función de pertenencia real utilizando una representación de baja dimensión elegida de una familia general adecuada. En la práctica, UMAP utiliza una familia de curvas de la forma:

$$\frac{1}{1 + ax^{2b}}$$

De la misma manera, no hay que lidiar con todos los bordes posibles, por lo que se puede usar el truco de muestreo negativo, para simplemente muestrear ejemplos negativos según sea necesario. Finalmente, dado que el laplaciano de la representación topológica es una buena aproximación del operador de la variedad de Laplace-Beltrami, es posible usar técnicas de incrustación espectral para inicializar la representación en bajas dimensiones en un buen estado.

4.2.4.5 *Implementación del algoritmo*

En resumen, el algoritmo UMAP es relativamente sencillo (ver Algoritmo 1). Al realizar una unión difusa sobre conjuntos simpliciales difusos locales hemos encontrado que es más efectivo trabajar con la t-conorma probabilística (como cabría esperar si se trataran las fortalezas de los miembros como una probabilidad de que el símplex existe). e funciones individuales para la construcción de la difusa local conjuntos simpliciales, determinando la incrustación espectral y optimizando la incrustación con respecto a la entropía cruzada de conjuntos difusos, se describen en más detalle a continuación.

las entradas al algoritmo 1 son: X, el conjunto de datos cuya dimensión se reducirá; n, el tamaño del vecindario que se usará para la aproximación métrica local; d, la dimensión del espacio reducido objetivo; min-dist, un parámetro algorítmico que controla el diseño; y n-épocas, controlando la cantidad de trabajo de optimización a realizar.

Algorithm 1 UMAP algorithm

function UMAP(X, n, d, min-dist, n-epochs)

 # Construct the relevant weighted graph
for all $x \in X$ **do**
fs-set$[x] \leftarrow$ LocalFuzzySimplicialSet(X, x, n)
top-rep $\leftarrow \bigcup_{x \in X}$ fs-set$[x]$ *# We recommend the probabilistic t-conorm*

 # Perform optimization of the graph layout
$Y \leftarrow$ SpectralEmbedding(top-rep, d)
$Y \leftarrow$ OptimizeEmbedding(top-rep, Y, min-dist, n-epochs)
return Y

Ilustración 20 Algoritmo 1 de UMAP [21]

El algoritmo 2 describe la construcción de conjuntos simpliciales difusos locales. Para representar conjuntos simpliciales borrosos, trabajamos con las imágenes de conjuntos borrosos, que denotamos como fs-set0 y fs-set1. También se puede trabajar con simples de orden superior, pero la implementación actual no lo hace. Podemos construir el conjunto simplicial borroso local a un punto dado x encontrando los n vecinos más cercanos, generando la distancia normalizada adecuada en la variedad y luego convirtiendo el espacio métrico finito en un conjunto simplicial a través del funtor FinSing, que se traduce en exponencial de la distancia negativa en este caso.

Algorithm 2 Constructing a local fuzzy simplicial set

function LocalFuzzySimplicialSet(X, x, n)

knn, knn-dists $\leftarrow$ ApproxNearestNeighbors(X, x, n)

$\rho \leftarrow$ knn-dists[1] # *Distance to nearest neighbor*

$\sigma \leftarrow$ SmoothKNNDist(knn-dists, n, ρ) # *Smooth approximator to knn-distance*

fs-set$_0 \leftarrow X$

fs-set$_1 \leftarrow \{([x, y], 0) \mid y \in X\}$

for all $y \in$ knn **do**

$d_{x,y} \leftarrow \max\{0, \text{dist}(x, y) - \rho\}/\sigma$

fs-set$_1 \leftarrow$ fs-set$_1 \cup ([x, y], \exp(-d_{x,y}))$

return fs-set

Ilustración 21 Algoritmo 2 de UMAP [21]

En lugar de utilizar directamente la distancia al enésimo vecino más cercano como la normalización, utilizamos una versión suavizada de la distancia kernel con el fin de arreglar el conjunto de elementos "fuzzy" simples de primer orden. Esta función utilizada en el algoritmo 2 está implementada en el algoritmo 3

Algorithm 3 Compute the normalizing factor for distances σ

function SmoothKNNDist(knn-dists, n, ρ)

Binary search for σ such that $\sum_{i=1}^{n} \exp(-(\text{knn-dists}_i - \rho)/\sigma) = \log_2(n)$

return σ

Ilustración 22 Algoritmo 3 de UMAP [21]

La incrustación espectral se realiza tratando el esqueleto de la representación topológica difusa global 1 como un gráfico ponderado y utilizando métodos espectrales estándar en Laplace normalizado simétrico. Este proceso está descrito en el Algoritmo 4.

Algorithm 4 Spectral embedding for initialization

function SPECTRALEMBEDDING(top-rep, d)

$A \leftarrow$ 1-skeleton of top-rep expressed as a weighted adjacency matrix

$D \leftarrow$ degree matrix for the graph A

$L \leftarrow D^{1/2}(D - A)D^{1/2}$

evec $\leftarrow$ Eigenvectors of L (sorted)

$Y \leftarrow$ evec$[1..d + 1]$ *# 0-base indexing assumed*

return Y

Ilustración 23 Algoritmo 4 de UMAP [21]

Como último paso se lleva a cabo un proceso de optimización. El proceso de optimización de UMAP implica calcular la entropía cruzada, el gradiente y actualizar las variables para ajustar la posición de los puntos en el espacio de baja dimensión. Este proceso se repite varias veces hasta que se encuentra una proyección adecuada de los datos en el espacio de baja dimensión.

La entropía cruzada se usa para medir la similitud entre los vecinos en el espacio de alta dimensión y los vecinos en el espacio de baja dimensión.

Algorithm 5 Optimizing the embedding

function OPTIMIZEEMBEDDING(top-rep, Y, min-dist, n-epochs)

$\alpha \leftarrow 1.0$

Fit Φ from Ψ defined by min-dist

for $e \leftarrow 1, \ldots,$ n-epochs **do**

for all $([a, b], p) \in$ top-rep$_1$ **do**

if RANDOM( ) $\leq p$ **then** *# Sample simplex with probability p*

$y_a \leftarrow y_a + \alpha \cdot \nabla(\log(\Phi))(y_a, y_b)$

for $i \leftarrow 1, \ldots,$ n-neg-samples **do**

$c \leftarrow$ random sample from Y

$y_a \leftarrow y_a + \alpha \cdot \nabla(\log(1 - \Phi))(y_a, y_c)$

$\alpha \leftarrow 1.0 - e/$n-epochs

return Y

Ilustración 24 Algoritmo 5 de UMAP [21]

4.2.5 SVM (Support vector machine)

La máquina de vectores de soporte es una técnica de clasificación lineal binaria en Machine Learning, que separa las clases con la brecha más grande (llamada margen óptimo) entre las

instancias de la línea de borde (llamadas Vectores de soporte). Por eso se le conoce como clasificador de margen óptimo.

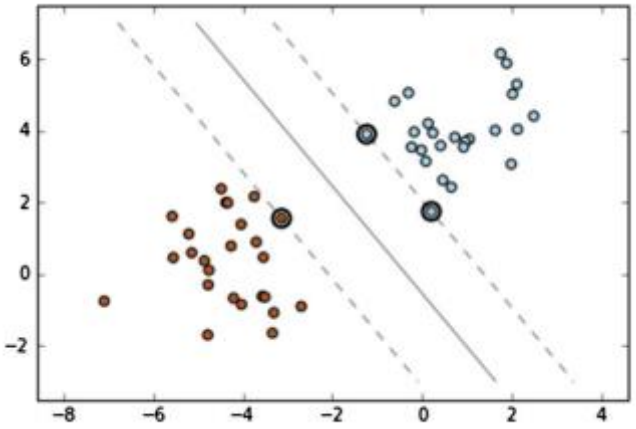

Ilustración 25 Separación de muestras con un hiperplano

La Ilustración 25, representa la vista geométrica de SVM. Se puede dibujar un número infinito de clasificadores para los datos dados, pero SVM encuentra el clasificador con la mayor brecha entre los vectores de soporte. Los círculos representan los vectores de soporte. Para problemas de datos separables no linealmente, SVM se ha ampliado utilizando Kernels. Los Kernels son funciones matemáticas que transforman los datos de un espacio dado (conocido como espacio de entrada) en un nuevo espacio de alta dimensión (conocido como espacio de características) donde los datos se pueden separar con una superficie lineal (llamada hiperplano). La Ilustración 26 representa la vista geométrica de Kernels. Matemáticamente, un Kernel es una función que toma dos argumentos, aplica una asignación en los argumentos y luego devuelve el valor de su producto escalar. [23]

SVM lineal es muy eficiente en aplicaciones de datos de alta dimensión. Si bien su precisión en ºaplicaciones. Por ejemplo, las aplicaciones de clasificación de documentos tienen un espacio de entrada de gran dimensión y no es necesario agregar más funciones al espacio de entrada porque no hace mucha diferencia en el rendimiento. Por lo tanto, la precisión de la prueba de SVM lineal es similar a la de SVM no lineal, pero al mismo tiempo, el entrenamiento de SVM lineal es mucho más rápido que el de SVM no lineal debido a la diferencia en sus complejidades computacionales. SVM lineal involucra formulaciones de problemas, solucionadores para resolver el problema y estrategias de optimización para hacer que los solucionadores sean eficientes.

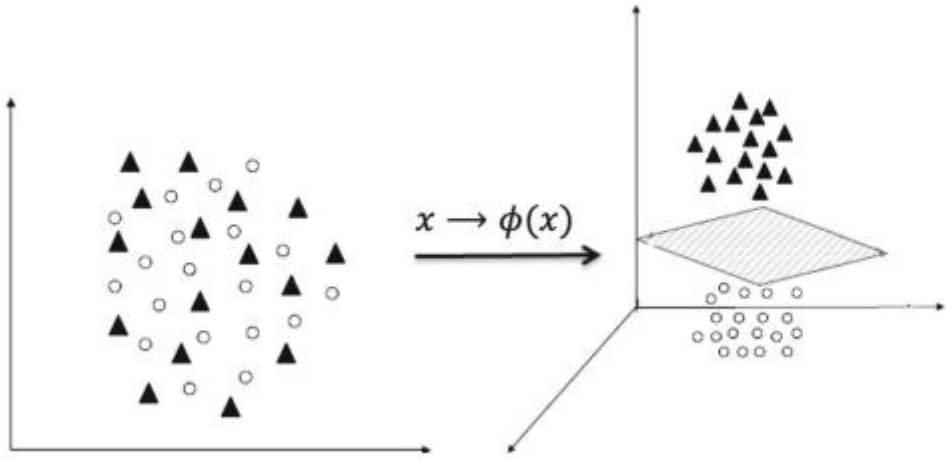

Ilustración 26 Aplicación de Kernel para muestras no lineales

4.2.5.1 One-class SVM

El One-class SVM (Máquinas de vectores de soporte de una sola clase) es una variante de SVM que se utiliza para la detección de anomalías.

A diferencia de las SVM convencionales, en las que se requieren dos clases de datos etiquetados (positivos y negativos), las One-class SVM utilizan solo una clase de datos etiquetados (por lo general, los datos normales o regulares). El objetivo de este algoritmo es construir un modelo que pueda determinar si una nueva instancia de datos se considera una "anomalía" en función de cómo se ajusta al conjunto de datos conocido.

El One-class SVM crea una frontera de decisión que rodea la mayor parte de los datos conocidos, y cualquier punto que se encuentre fuera de esta frontera se considera una anomalía. Este método se utiliza comúnmente en aplicaciones de detección de fraude, detección de intrusiones y monitoreo de procesos industriales para identificar patrones de comportamiento anormales que pueden indicar problemas o riesgos potenciales. En la Ilustración 27 se muestra un conjunto de datos 2D que consiste en puntos azules (datos normales) y un punto rojo (datos anómalos). El One-class SVM crea una frontera de decisión que rodea la mayoría de los puntos azules, creando un área de espacio libre que contiene a los puntos rojos. Cualquier nuevo punto que se coloque en esta área libre se considera anómalo. [24]

One-class SVM es un método basado en el kernel propuesto por Schölkopf y Müller en 2001. La idea principal de one-class SVM es que los datos de entrenamiento en el espacio de entrada se asignan al espacio de características a través de una función del kernel y encuentran un hiperplano con un margen máximo en el espacio de características para separar los datos mapeados del origen. Existen dos enfoques para diseñar un límite de decisión en one-class SVM. El primer enfoque describe un límite de hiperesfera entre los valores atípicos y la clase positiva. La forma del límite está determinada por el parámetro "ν" que se utiliza para definir el equilibrio entre el porcentaje de puntos de datos como positivo y el valor atípico. El segundo enfoque es entrenar el límite de decisión como un hiperplano entre los puntos de datos y el origen (del sistema de coordenadas) y separar un determinado porcentaje de los valores atípicos del resto de los datos. El segundo enfoque ha sido aplicado por muchas implementaciones de one-class SVM debido a su implementación más simple. [25]

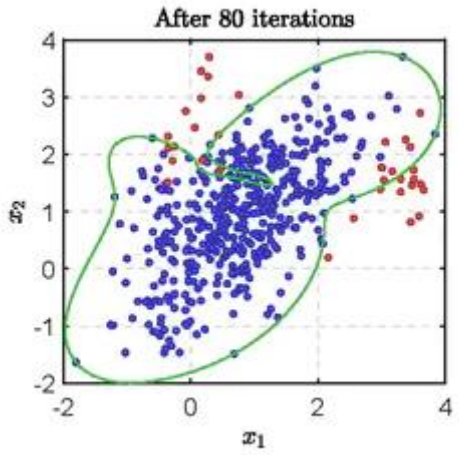

Ilustración 27 Muestras anómalas separadas por un umbral de decisión[24]

Otra idea geométrica similar a SVM más intuitiva se implementa en la formulación de la esfera. Los datos normales se pueden describir sucintamente mediante una esfera (en el espacio de características) que contiene los datos como en la *Ilustración 28*.

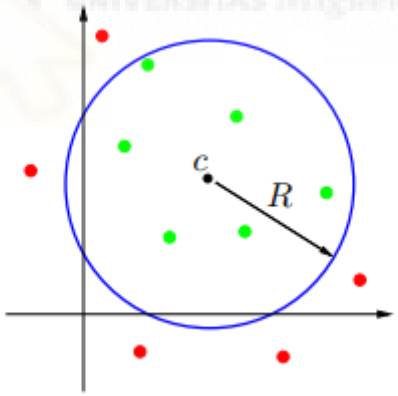

Ilustración 28 La geometría de la formulación de la esfera de one-class SVM [26]

Las anomalías en los datos de entrenamiento se pueden manejar introduciendo variables de holgura ξ_i correspondientes a la fórmula del plan. Matemáticamente, el problema del "ajuste suave" de una esfera a los datos se describe de la siguiente manera:

$$\begin{cases} \min R^2 + \frac{1}{vl} \sum_{i=1}^{l} \xi_i \\ sujeto\ a: ||\Phi(x_i) - c||^2 \leq R^2 + \xi_i \\ \xi_i \geq 0 \end{cases}$$

(19)

No puede resolver directamente el problema elemental de la ecuación *(19)*

con la fórmula de la esfera porque el centro c pertenece al posible espacio de características de alta dimensión. Se puede usar el siguiente truco: encontrar una solución al problema dual:

$$\begin{cases} \min \sum_{ij=1}^{l} \alpha_i \, \alpha_j \, k(x_i, x_j) - \sum_{ij=1}^{l} \alpha_i \, k(x_i, x_j) \\ sujeto\ a: \sum_{i=1}^{l} \alpha_i = 1 \\ 0 \le \alpha_i \le \frac{1}{vl} \end{cases}$$

(20)

La función de decisión se puede computar como:

$$f(x) = sgn\left[R^2 - \sum_{ij=1}^{l} \alpha_i \, \alpha_j \, k(x_i, x_j) + 2\sum_{i=1}^{l} \alpha_i \; k(x_i, x_j) - k(x_i, x) - k(x, x)\right]$$

(21)

El radio R2 juega el papel de un umbral, se puede calcular igualando la expresión bajo el "sgn" a cero para cualquier vector de soporte

4.2.5.2 SVDD

SVDD (Support Vector Data Description) es un método de aprendizaje no supervisado que se utiliza para detectar anomalías en conjuntos de datos. Es similar a SVM de clase única, pero en lugar de usar una clase única de datos etiquetados, SVDD usa descriptores de datos para crear regiones esféricas o elípticas alrededor de datos normales. El objetivo de SVDD es encontrar una región esférica o elíptica que cubra la mayor cantidad posible de datos normales y excluya la menor cantidad posible de datos anormales. Una región esférica o elíptica se define mediante la optimización de una función de costo que mide la distancia entre los datos y una región esférica o elíptica. [10]

Las regiones esféricas o elípticas se pueden definir de dos maneras diferentes:

Método de límite interno: un rango esférico o elíptico se ajusta a los datos normales, lo que permite que algunos valores atípicos se encuentren dentro del rango. Este método es útil cuando los datos anómalos son grandes o cuando los datos anómalos se superponen con los datos normales. Método de la bola exterior: se ajusta una región esférica o elipsoidal justo fuera de los datos normales, eliminando todas las desviaciones. Este método es útil cuando hay pequeños valores atípicos o cuando los valores atípicos están claramente separados de los datos normales. En resumen, SVDD es una técnica de detección de anomalías que utiliza descriptores de datos para crear regiones esféricas o elípticas alrededor de datos normales. Esta tecnología se puede utilizar en una variedad de aplicaciones, como detección de fraude, detección de intrusos y monitoreo de procesos industriales.

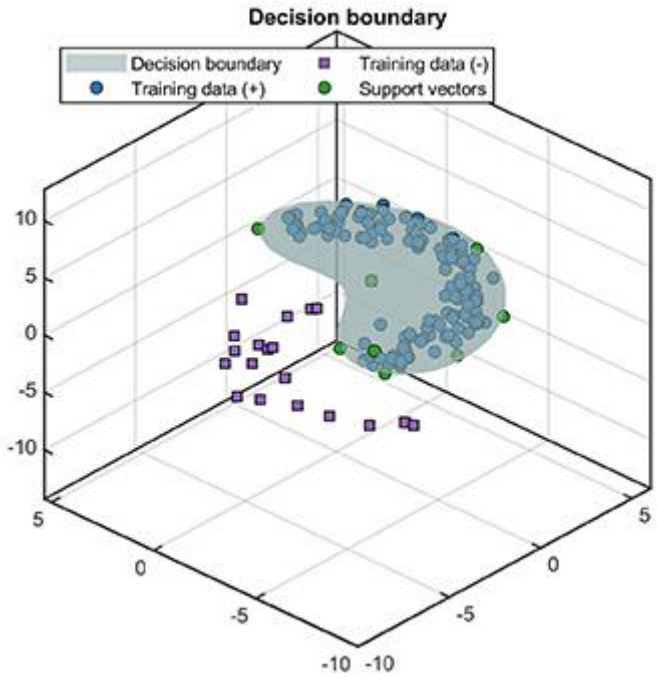

Ilustración 29 Aplicación de SVDD a partir de unos datos de entrenamiento.[27]

Como variante de SVM, SVDD es un algoritmo típico de clasificación de clases. SVDD identifica anomalías principalmente al determinar la hiperesfera con el volumen más pequeño alrededor de las muestras positivas en el espacio del objeto. Los modelos SVDD tradicionales se dividen en dos categorías: SVDD lineal y SVDD no lineal de kernel. Se describen a continuación.

A partir de unas muestras normales de entrenamiento dadas como $\{x_1, x_2, \ldots, x_n\}$ donde $x_i \in R^m$, i=1, 2, …, n, el objetivo de la optimización lineal SVDD es buscar por una descripción de datos ajustada descrita de la siguiente manera:

$$\begin{cases} \min R^2 + \dfrac{\gamma}{n} \sum_{i=1}^{n} \sigma_i \\ ||x_i - o\,||^2 \le R^2 + \sigma_i \\ \sigma_i \ge 0 \end{cases}$$

(22)

Donde || x || representa la norma euclídea, "o" es el centro de la esfera, "σ_i" es la variable de relajación, "R" es el radio de la hiperesfera, y "γ" es el parámetro de compensación que coordina el volumen de la hiperesfera y el error de modelado. $\sum_{i=1}^{n} \sigma_i$ es el término de penalización que permite valores atípicos.

La optimización anterior funciona en el caso lineal. Cuando existen relaciones de datos no lineales, los datos de entrenamiento iniciales no se distribuyen esféricamente, por lo que las hiperesferas no pueden aislar anomalías de manera efectiva. Es por eso que se usa el núcleo SVDD. Primero, la función de mapeo no lineal φ (.) asigna estas muestras a un nuevo espacio de características: $x_i \rightarrow \varphi(x_i)$, donde todas las muestras tienen una relación lineal. Entonces la aplicación básica SVDD se aplica de la siguiente manera:

$$\begin{cases} \min R^2 + \frac{\gamma}{n} \sum_{i=1}^{n} \sigma_i \\ ||\varphi(x_i) - o\,||^2 \leq R^2 + \sigma_i \\ \sigma_i \geq 0 \end{cases}$$

(23)

Resolviendo la optimización anterior se obtiene la siguiente expresión dual:

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i \ \langle \varphi(x_i), \varphi(x_i) \rangle - \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle \\ 0 \leq \alpha_i \leq \ \gamma \\ \sum_{i=1}^{n} \alpha_i = 1 \end{cases}$$

(24)

Donde α_i es el multiplicador de Lagrange, $\langle \varphi(x_i), \varphi(x_j) \rangle$ representa el producto interno en el espacio de mapeo Dado que la función de mapeo generalmente se desconoce, la función kernel se usa para calcular el producto interno como:

$$< \ \varphi(x_i), \varphi(x_i) > = Ker(x_i, x_j)$$

(25)

Donde Ker es alguna función kernel que satisface el teorema de Mercer. Una función kernel comúnmente utilizada es el kernel gaussiano. Con el uso de la función de Kernel la optimización en la ecuación (24) se puede resolver y el centro de la esfera se obtiene como $\sum_{i=1}^{n} \alpha_i \ \varphi(x_i)$. Para determinar si un vector de prueba x es un valor atípico, se desarrolla una métrica de decisión, también conocida como métrica de evaluación:

$$D(x) = ||\Phi(x) - o\,||^2$$

(26)

Una muestra de prueba x con D(x) > R indica una anomalía. De lo contrario, la prueba es normal. [28]

5 Experimentos y resultados

Para el análisis de nuestra base de datos podemos dividir las técnicas que vamos a utilizar en dos tipos, modelos supervisados y modelos no supervisados. Esta distinción se debe a que en unos casos se ha entrenado el algoritmo para identificar a que grupo (grupos del 0 al 3) pertenecen los datos de entrenamiento (supervisado) y en otro caso los datos de entrenamiento no asumen pertenecían a ningún grupo (no supervisado).

5.1 Modelos no supervisados

El aprendizaje no supervisado es una rama del aprendizaje automático que se enfoca en analizar datos sin etiquetas o respuestas previas conocidas. Sus modelos se utilizan para descubrir patrones y estructuras ocultas en conjuntos de datos, sin la guía de información externa. Hay tres tipos principales de modelos no supervisados: agrupamiento (para dividir datos en grupos similares), reducción de dimensionalidad (para simplificar datos manteniendo la información relevante), y asociación (para encontrar patrones de relación entre elementos de datos).

Estos modelos se aplican en diversas áreas, como la detección de fraudes en transacciones financieras, la segmentación de clientes para estrategias de marketing personalizadas, y la reducción de la dimensionalidad en análisis de datos complejos. Aunque la evaluación puede ser desafiante debido a la falta de respuestas conocidas, el aprendizaje no supervisado es esencial para explorar datos no etiquetados y descubrir información valiosa, lo que lo convierte en una herramienta esencial en la exploración de datos y en una etapa preliminar en problemas supervisados de aprendizaje automático.

Es importante mencionar que los valores en nuestra base de datos ya se encontraban normalizados por lo que no ha sido necesario aplicar este proceso.

Los modelos de aprendizaje no supervisado que vamos a utilizar son los siguientes: Reducción de dimensionalidad y SVDD

5.1.1 Reducción de dimensionalidad

Los primeros modelos no supervisados que generaron espacios latentes consistieron en modelos lineales que utilizaban PCA. El modelo de datos se utiliza para encontrar las direcciones mutuamente ortogonales de varianza máxima de las muestras de datos de la encuesta que explican la mayor parte de la varianza de la muestra y comprimirlas en un espacio tridimensional. Esto significa que el modelo PCA nos permite proyectar las muestras en un espacio tridimensional para explorar las relaciones espaciales entre las muestras de entrada.

A continuación, en la Ilustración 30 se muestra el resultado después de aplicar PCA a nuestra base de datos.

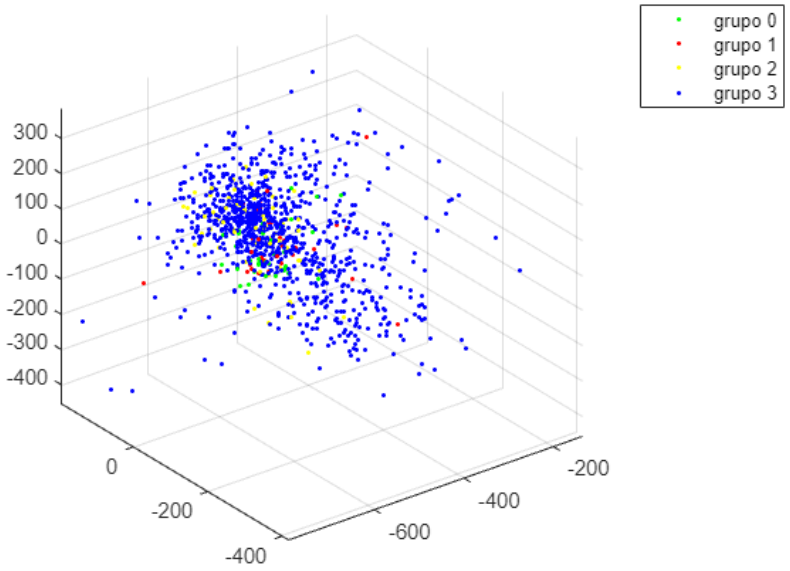

Ilustración 30 Agrupación de pacientes mediante PCA

A pesar de la aplicación adecuada de esta técnica de procesamiento y la limpieza de datos no se ha logrado una distinción clara entre los 4 grupos a analizar.

El siguiente algoritmo de reducción de dimensionalidad que hemos utilizado es tSNE. En la Ilustración 31 se muestra el resultado después de aplicar este algoritmo. Se puede apreciar como el algoritmo tSNE intenta agrupar las muestras en grupos y se pueden apreciar dos conjuntos claramente separados. Sin embargo, esta distinción no depende de los grupos que estamos intentando separar.

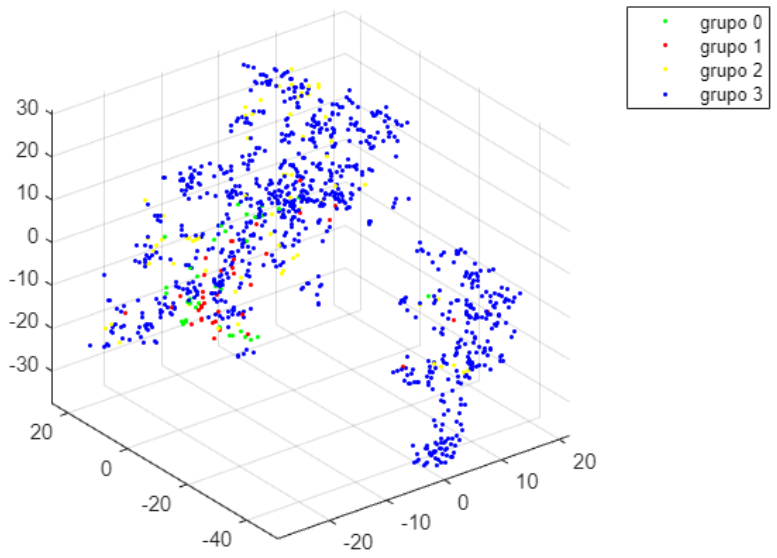

Ilustración 31 Agrupación de pacientes mediante t-SNE

Otro algoritmo de reducción de dimensionalidad es el uso de un Autoencoder. El modelo de datos utilizado consta de un AE con una sola capa oculta, el cuello de botella está formado por 3 neuronas para reducir el tamaño mediante redes neuronales. Lo que significa que la representación de los datos se comprime en un conjunto de datos tridimensional que se puede mostrar posteriormente. La representación esquemática del autoencoder se muestra en la Ilustración 32.

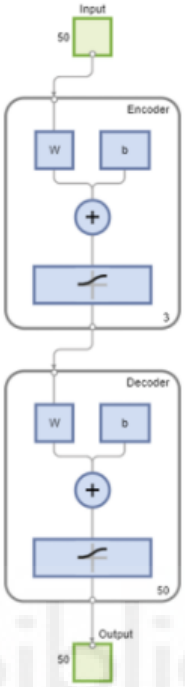

Ilustración 32 Representación esquemática del autoencoder

Mediante esta técnica de reducción de dimensionalidad se obtiene el resultado de la Ilustración 33. En este caso se puede apreciar como los grupos del 0 al 2 están en mayor medida agrupados en un mismo espacio, lo que podría significar que las muestras sanas se agrupan. Sin embargo, fuera de ese espacio también hay otros puntos que corresponden a los grupos del 0 al 2 fuera de ese espacio. Observar, además, que la agrupación de pacientes del grupo 3 no sigue ningún patrón e incluso se mezclan con el espacio del resto de grupos. Por lo tanto no se puede concluir que a raíz de este resultado podamos distinguir de ninguna manera entre pacientes sanos o enfermos.

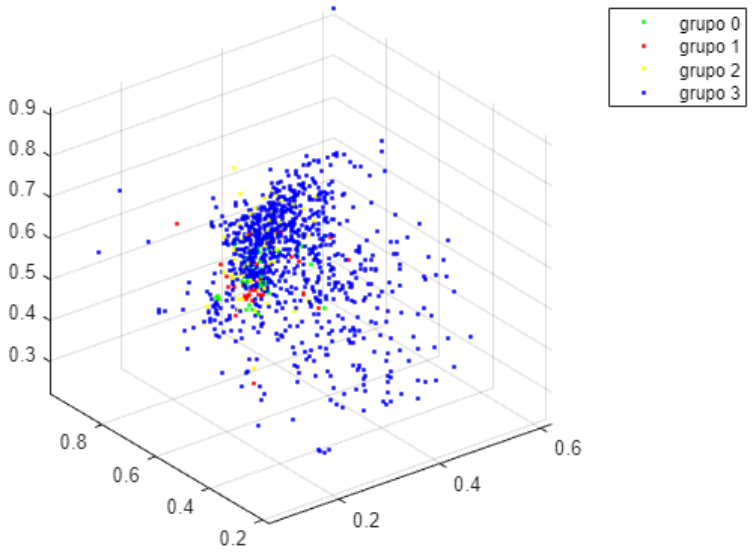

Ilustración 33 Agrupación de pacientes mediante AE

El algoritmo de reducción de dimensionalidad restante es el de UMAP, el resultado se muestra en la Ilustración 34. Este algoritmo obtiene el mejor resultado porque es capaz de agrupar los datos en dos espacios diferentes (al igual que tSNE) y además una buena parte de las muestras de los grupos 0-2 se agrupan en un mismo espacio. Sin embargo, no podemos afirmar que este algoritmo nos permite predecir si un paciente padece la enfermedad ya que los grupos se encuentran demasiado mezclados

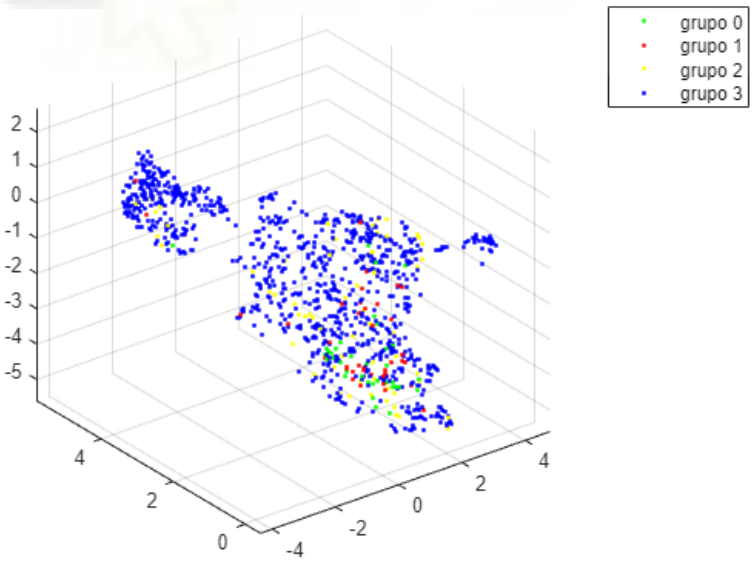

Ilustración 34 Agrupación de pacientes mediante UMAP

5.1.2 SVDD

Dado que el método UMAP es el que mejor funciona en la matriz de entrada y produce pocas diferencias entre las regiones tridimensionales de los datos, es posible utilizar las regiones identificadas como los dominios más comunes de los vectores de casos de pacientes. Esto puede solucionar el problema de detectar anomalías o desviaciones. Utilizando un conjunto de datos mapeado en el espacio latente, utilizamos técnicas de descripción de dominio de vector de soporte (SVDD) para identificar límites de volumen que representan las regiones más comunes en los casos observados hasta ahora.

El proceso de entrenamiento de SVDD se basa en un algoritmo que encuentra una hiperesfera en un espacio de alta dimensión (que representa el espacio de Hilbert del núcleo) que contiene todos los datos no atípicos observados hasta ahora. Al mapear la hiperesfera al espacio latente, se puede obtener una región geométrica que describe un dominio de geometría compleja arbitraria, que siempre depende de la descripción de los datos mismos.

La representación de SVDD con UMAP se muestra en la Ilustración 35. En la figura se muestra que la hiperesfera solamente detecta como valores anómalos los que se encuentran en los bordes de cada uno de los espacios. Por lo tanto, tampoco hace ningún tipo de distinción entre grupos.

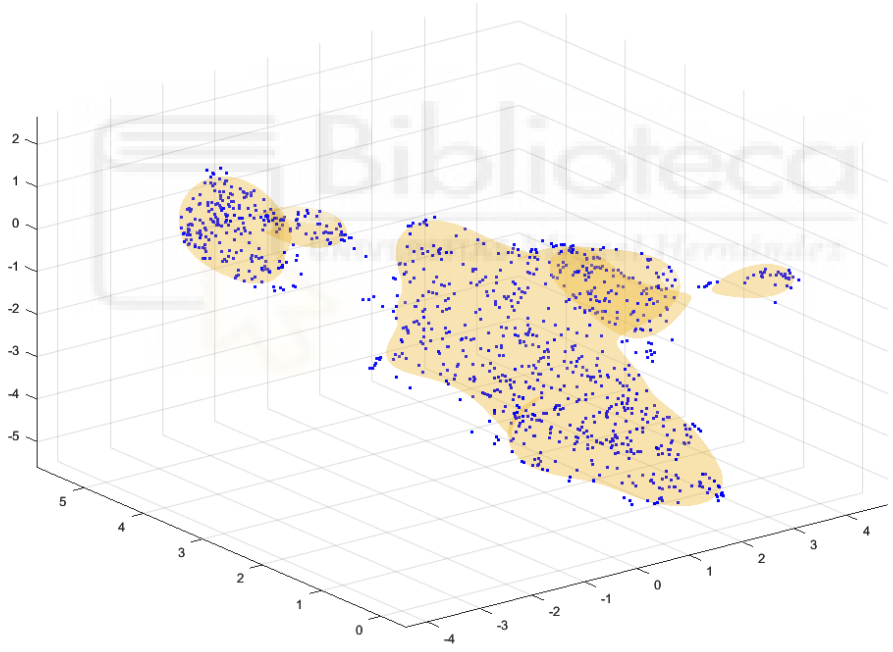

Ilustración 35 SVDD en el espacio latente UMAP

5.2 Modelos supervisados

El aprendizaje supervisado es una rama del aprendizaje automático en la que los algoritmos se entrenan utilizando conjuntos de datos etiquetados, lo que significa que cada ejemplo de entrada tiene una respuesta o etiqueta conocida, en nuestro caso estas etiquetas serán uno de los grupos de pacientes. El objetivo principal es aprender a asignar entradas a salidas deseadas para poder hacer predicciones precisas sobre datos nuevos e invisibles. En el aprendizaje supervisado, los

modelos se dividen en dos categorías amplias: clasificación (asigna una etiqueta o clase a cada entrada) y regresión (utiliza valores continuos para predecir). Tiene aplicaciones en diversos campos como el procesamiento del lenguaje natural, visión por computadora, medicina, economía, etc. El éxito de los algoritmos de aprendizaje supervisado, medido por métricas como la precisión, la puntuación F1 o la raíz del error cuadrático medio, está relacionado con la capacidad de generalizar patrones a partir de datos de entrenamiento y utilizar estos patrones para tomar decisiones informadas en nuevos escenarios.

Los modelos de aprendizaje no supervisado que vamos a utilizar son los siguientes: SVM, UMAP supervisado y one class SVDD.

5.2.1 SVM

Las SVM son particularmente eficientes para conjuntos de datos de alta dimensión y pueden manejar datos no lineales utilizando funciones del kernel que transforman los datos en un espacio de características de mayor dimensión. Esto les permite capturar relaciones complejas en los datos. Las SVM se utilizan ampliamente en diversas aplicaciones, como clasificación de texto, reconocimiento de imágenes, detección de anomalías y muchos otros problemas de aprendizaje automático que requieren una separación eficiente y una buena generalización.

Para este análisis vamos a separar las muestras en dos grupos. El primer grupo el grupo A estará formado por todas las muestras del grupo 0, que se consideran los pacientes completamente sanos. El segundo grupo será el grupo B y estará formado por el resto de los pacientes.

Con este planteamiento se nos presenta un problema y ese es la abismal diferencia de pacientes entre el grupo A y el grupo B. Para solucionarlo se puede optar por dos enfoques, igualar las muestras al alta o igualarlas a la baja.

Vamos a hacer un primer análisis igualando las muestras al alta, para ello utilizamos sobremuestreo. Se aumenta la cantidad de muestras en el grupo más pequeño creando nuevas muestras sintéticas basadas en las muestras existentes.

Mas adelante debemos separar un porcentaje de cada uno de los grupos como muestras de entrenamiento y otro porcentaje como muestras de test. El objetivo es que el algoritmo a partir del entrenamiento de una muestra aleatoria entre estos dos grupos sea capaz de predecir con la mayor exactitud posible las muestras de test.

Los resultados de todo este proceso se muestran en la matriz de confusión de la Ilustración 36, el 0 representa al grupo A (sanos, originalmente el grupo 0) y el 1 representa al grupo B. Aunque la matriz de confusión tenga una precisión de un 82% realmente no podemos decir que es un resultado positivo por dos motivos. El primero es que, aunque a primera vista parece un resultado alto no es lo suficientemente preciso como para que pueda ser considerado un diagnóstico médico preciso. El segundo motivo es que el valor de la matriz de confusión de abajo a la izquierda es demasiado alto. Este valor se llama falso positivo y está definido de la siguiente manera: “incorrectamente indica que la condición está presente”. Para nuestro análisis este valor indica que la condición es paciente sano. Por lo tanto, un 26% de los pacientes enfermos están siendo considerados como sanos.

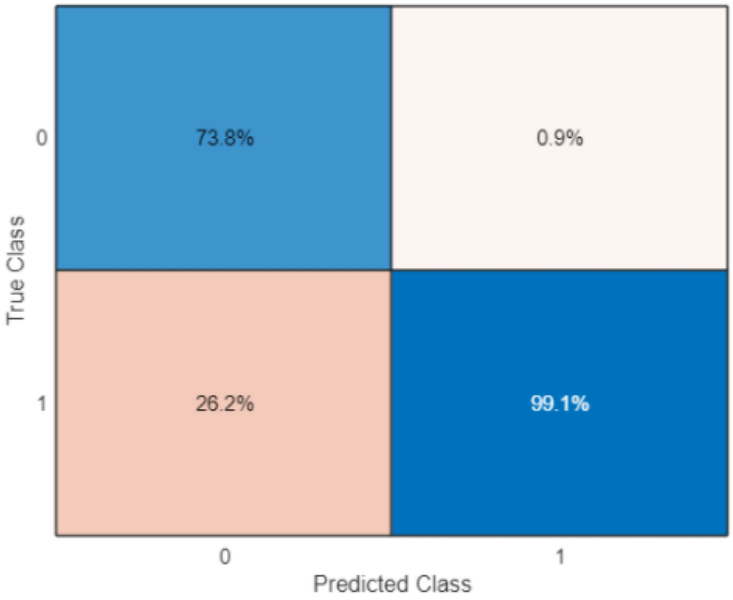

Ilustración 36 Matriz de confusión SVM

Podemos cambiar el enfoque que hemos seguido a la hora de igualar las muestras entre el grupo A y el grupo B. En este caso vamos a probar a igualar las muestras a la baja. Para ello utilizamos un submuestreo, donde se reduce la cantidad de muestras en el grupo más grande, eliminando algunas de las muestras para igualar el tamaño del grupo minoritario.

El resultado se muestra en la Ilustración 37. Debido a haber igualado las muestras a la baja y a haber separado parte para entrenamiento y otra para test apenas hay muestras a considerar y encima SVM no ha sido capaz de separar de ninguna manera las muestras.

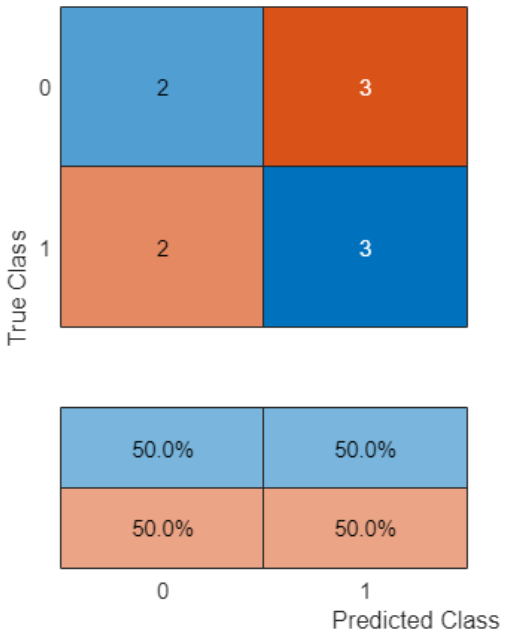

Ilustración 37 Segunda matriz de confusión SVM

5.2.2 One class SVDD

La "descripción de datos vectoriales de soporte de una clase" (SVDD de una clase o simplemente SVDD) es un algoritmo de aprendizaje automático que se utiliza principalmente para la detección de anomalías. Su objetivo principal es describir y encapsular una distribución normal de datos en espacios de alta dimensión. Para lograr esto, entrena una clase de datos (datos normales) y forma una "burbuja" o hiperesfera alrededor de esos datos normales. También conocida como esfera o hiperboloide, esta hiperesfera es más adecuada para ajustar datos normales y es capaz de capturar la estructura subyacente de una clase.

Una vez definida la hiperesfera SVDD, se puede utilizar para identificar puntos de datos que se encuentran fuera de esta estructura, convirtiéndolos en anomalías candidatas. SVDD es particularmente útil cuando tienes un conjunto de datos muy desequilibrado y tu principal interés es detectar anomalías en algunas clases.

Para el entrenamiento de este algoritmo hemos considerado como normal muestras del grupo 0. Por lo tanto, el resto de las muestras de tests deberán quedar dentro de la hiperesfera. En este caso los grupos 1, 2 y 3 deberían estar claramente separados de las muestras del grupo 0

Como algoritmo de reducción de dimensionalidad hemos elegido UMAP ya que es el más preciso entre todos.

A partir de todo lo dicho hasta ahora el resultado es el que se muestra en la Ilustración 38

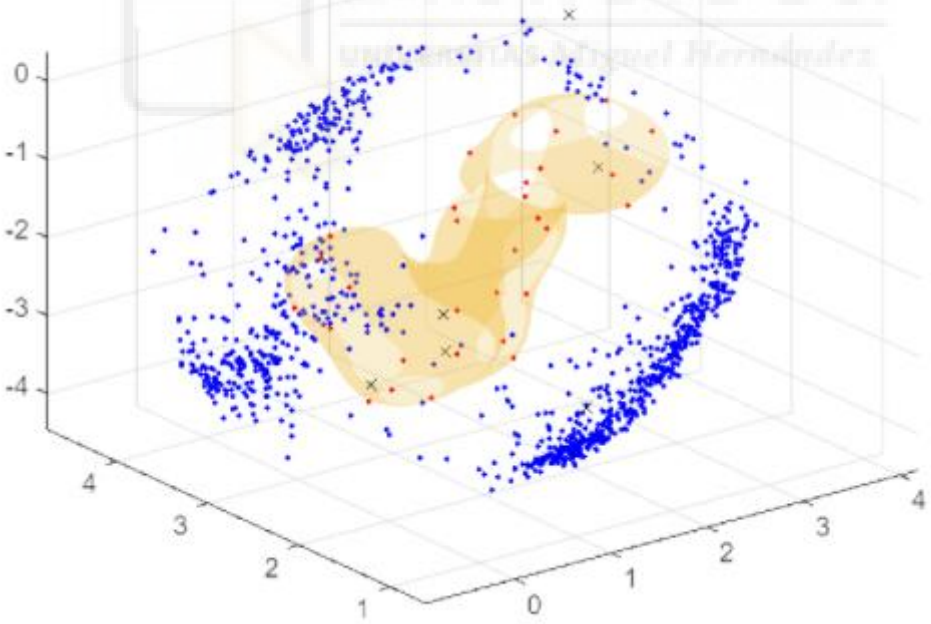

Ilustración 38 Espacio latente SVDD en UMAP

Las muestras representadas con una “X” representan al grupo 0 que no ha sido entrenado y considerado como normal, los puntos rojos son las muestras del grupo 0 y los puntos azules son

las muestras del resto de grupos. Aunque en la figura no aprecia las muestras representadas con una X se encuentran fuera de la hiperesfera, es decir, ha clasificado muestras del grupo 0 como muestras del resto de grupos. Esto se debe a que el algoritmo con lo que tiene únicamente es capaz de separar las muestras de entrenamiento y de las de test.

5.2.3 Umap supervisado

UMAP (Unified Manifold Approximation and Projection) es una técnica de reducción de dimensionalidad desarrollada principalmente como un método no supervisado para visualizar y reducir de manera eficiente datos de alta dimensión. Sin embargo, UMAP no es inherentemente un algoritmo supervisado porque su objetivo principal es preservar las relaciones entre puntos de datos independientemente de las etiquetas de clase o categoría. En su forma original, UMAP se basa en la topología de los datos para generar una representación bidimensional o tridimensional que mantiene la estructura subyacente de los datos, lo que lo hace adecuado para la exploración y visualización de datos sin etiquetas.

En nuestro caso y al igual que en los dos métodos supervisados anteriores, vamos a dividir nuestro conjunto de muestras en una parte de entrenamiento y en otra de test. De esta manera es posible que el algoritmo sea capaz de separar los grupos del 0 al 3.

Las muestras utilizadas en el entrenamiento las representa UMAP de la siguiente manera (ver Ilustración 39)

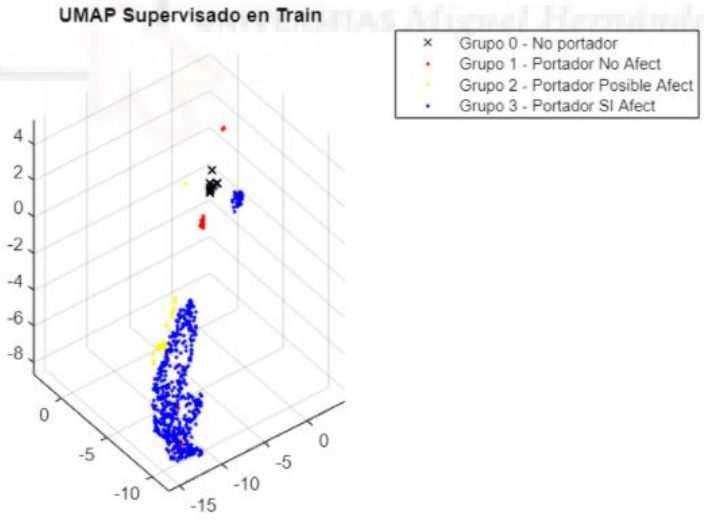

Ilustración 39 UMAP Supervisado en entrenamiento

El algoritmo separa en mayor o menor medida los grupos en la fase de entrenamiento. Sin embargo, no se consigue este resultado con las muestras de test (ver Ilustración 40).

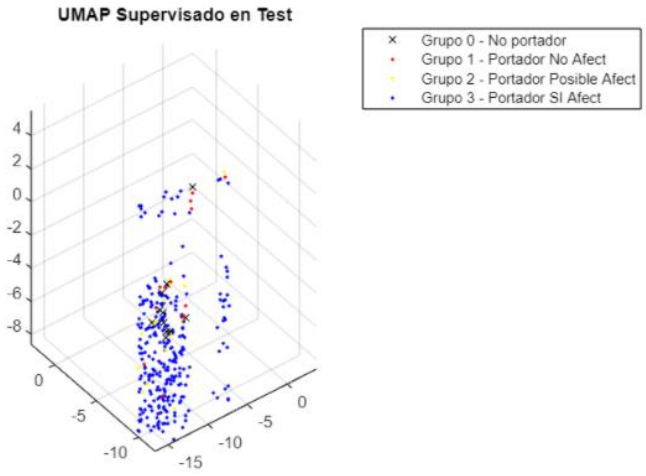

Ilustración 40 UMAP Supervisado en test

La agrupación en el espacio de las muestras de test es similar a la agrupación de las muestras de entrenamiento. Aun así, el algoritmo no puede separar los grupos del 0 al 3 de manera que en una de las agrupaciones se encuentran incluso los 4 grupos.

En lo que se refiere a UMAP supervisado existe otra manera de abordar el problema, podemos variar el número de vecinos. En UMAP, el número de vecinos se refiere a la cantidad de puntos de datos cercanos que se consideran al calcular las relaciones entre los puntos.

Al aumentar el número de vecinos, se considerarán más puntos cercanos al calcular las relaciones entre los datos, lo que puede llevar a una representación más suave y densa. Esto puede ser útil cuando se trabaja con datos que tienen una estructura subyacente compleja o cuando se busca una representación más precisa de los datos.

Por otro lado, reducir el número de vecinos puede conducir a una representación más dispersa y menos densa. Esto puede ser beneficioso cuando se busca simplificar la estructura de los datos o cuando se tiene un conjunto de datos grande y se desea reducir la complejidad computacional.

En las siguientes ilustraciones se muestran varios resultados con diferentes números de vecinos.

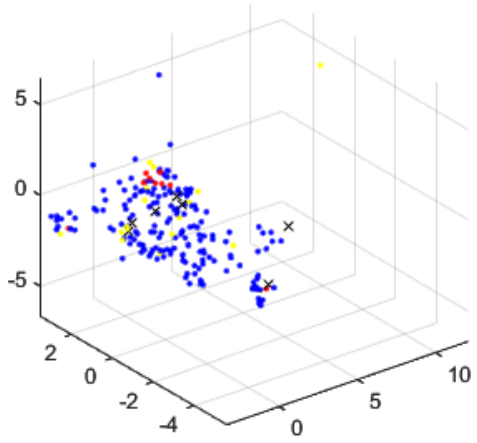

Ilustración 41 UMAP Supervisado con nºvecinos=5

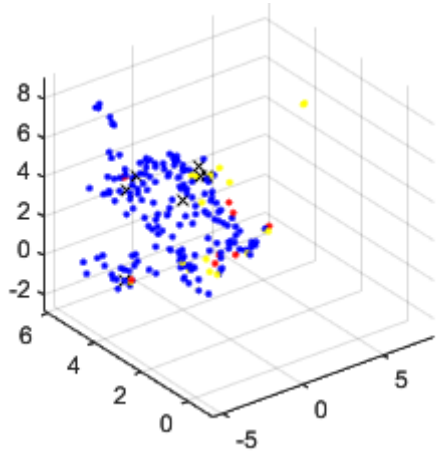

Ilustración 42 UMAP Supervisado con nºvecinos=10

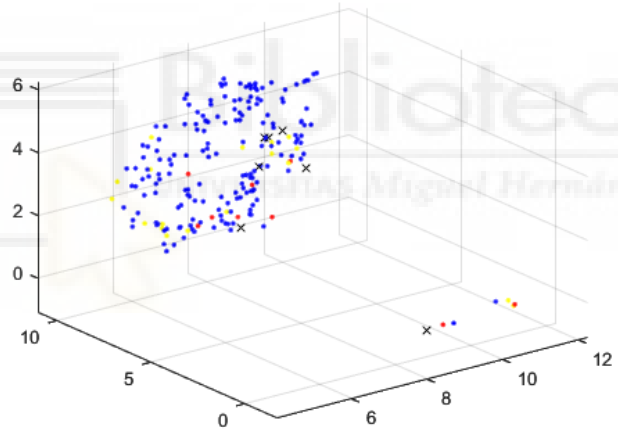

Ilustración 43 UMAP Supervisado con nºvecinos=30

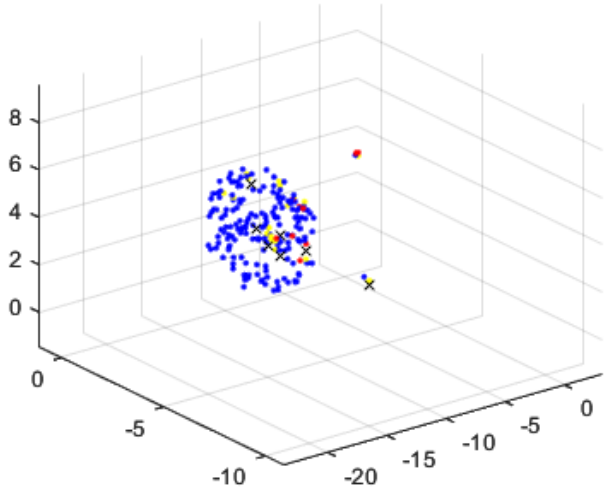

Ilustración 44 UMAP Supervisado con nºvecinos=50

5.3 Justificación de los resultados obtenidos

Los resultados que se han obtenido en este apartado no han sido satisfactorios. Para justificarlos se ha optado por hacer un análisis de la superposición de todas las variables. Con esta finalidad hemos elegido representar cada variable mediante una distribución normal con cada grupo.

La elección de representar las distribuciones de cada variable como distribuciones normales es una estrategia estadística que puede ser justificada en este contexto debido a la observación de que hay escasas diferencias apreciables entre los grupos de datos. Cuando las diferencias entre grupos son mínimas, suponer que las distribuciones siguen una forma normal es una aproximación razonable y simplificadora. La distribución normal, o campana de Gauss, es conocida por su propiedad de ser una distribución simétrica y, en muchos casos, representa una suposición conservadora que facilita el análisis y la interpretación de los datos. Esto puede ayudar a resaltar cualquier diferencia o variabilidad significativa que pudiera existir, ya que las discrepancias con respecto a una distribución normal pueden indicar posibles patrones o anomalías en los datos.

Varios ejemplos de variables de nuestra base de datos se muestran en las ilustraciones siguientes:

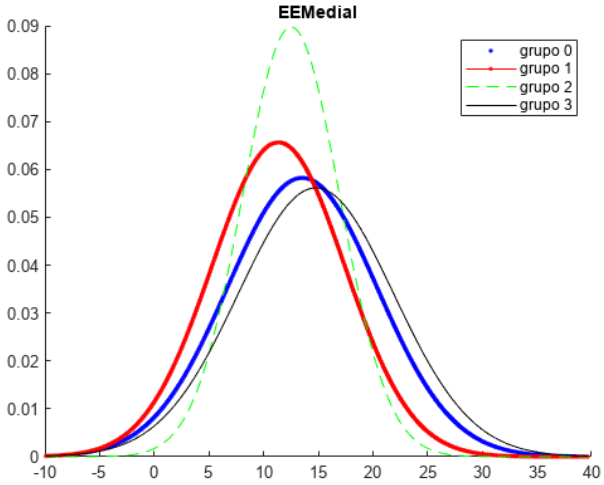

Ilustración 45 Superposición de la variable E/E' medial

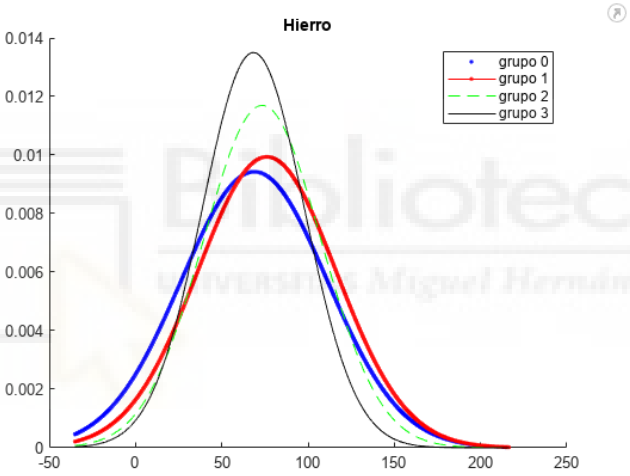

Ilustración 46 Superposición de la variable hierro

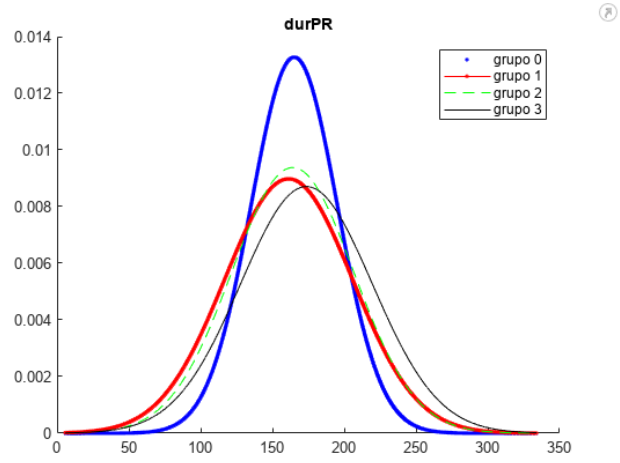

Ilustración 47 Superposición de la variable durPR

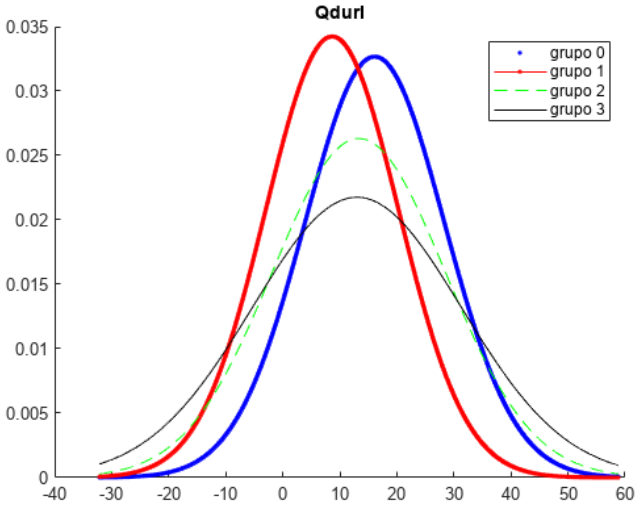

Ilustración 48 Superposición de la variable duración de la onda Q

Como muestra de la superposición de variables se ha elegido algunas de las muchas que forman la base de datos. Aun así la superposición de los grupos es casi idéntica en cada una de ellas por lo que no es necesario representarlas todas.

En las figuras anteriores se puede observar como no hay diferencia entre las variables de cada grupo por lo que es coherente que los algoritmos utilizados en este TFG no hayan sido capaces de separar cada uno de los grupos.

6 Conclusiones

Este estudio ha arrojado importantes conclusiones sobre la detección de la Enfermedad de Cardiomiopatía Hipertrófica (HCM) utilizando algoritmos de aprendizaje automático y deep learning a partir de una base de datos matricial del Hospital de Murcia. Las limitaciones en la cantidad de datos de pacientes sanos y la presencia de valores faltantes en la base de datos presentaron desafíos significativos. Esto impactó negativamente en la capacidad de los modelos para lograr una separación efectiva entre pacientes afectados y no afectados por la HCM.

La principal conclusión es la necesidad crítica de mejorar la calidad y la cantidad de los datos. Futuras investigaciones deben centrarse en la adquisición de datos más completos y precisos, posiblemente a través de la colaboración con múltiples fuentes de datos médicos. También es esencial abordar la corrección de valores faltantes en la base de datos. Además, se debe trabajar en el desarrollo de modelos de machine learning y deep learning más robustos y en la exploración de técnicas avanzadas para la detección de HCM.

Finalmente, se destaca la importancia de llevar a cabo una validación clínica rigurosa de cualquier modelo desarrollado en investigaciones futuras. Esto garantizará que los modelos sean clínicamente útiles y seguros para su implementación en la práctica médica. A pesar de los desafíos actuales, este TFG sienta las bases para futuros avances en la detección temprana y precisa de la HCM, lo que podría tener un impacto significativo en la atención médica y el diagnóstico de esta enfermedad cardíaca hereditaria.

7 Referencias

[1] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, vol. 31, no. 3, 2021, doi: 10.1007/s12525-021-00475-2.

[2] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, 2020, doi: 10.1007/s11263-019-01247-4.

[3] G. Rodas, C. Pedret, J. Ramos-Castro, and L. Capdevila, "Variabilidad de la frecuencia cardiaca"," *Arch. Med. del Deport.*, vol. 25, no. 123, pp. 41–7, 2008, [Online]. Available: http://www.efdeportes.com/

[4] "Sistema de conducción cardíaco - Videos de salud: MedlinePlus enciclopedia médica." https://medlineplus.gov/spanish/ency/anatomyvideos/000021.htm (accessed Sep. 02, 2023).

[5] Federico Ramirez, "Fisiología cardiaca," *Rev. Médica MD*, vol. 1, no. 3, pp. 3–6, 2009, [Online]. Available: https://www.medigraphic.com/pdfs/revmed/md-2009/md093d.pdf

[6] *Capítulo del libro: Ecocardiografía Básica. M.A. García Fernández y col. Si desea descargarse otros capítulos u obtener más información puede hacerlo desde la página www.ecocardio.com*.

[7] I. P. G., J. G., and J. B., *Hematología. Fisiopatología y Diagnostico*. 2009.

[8] M. Pere, "Interpretación clínica de las pruebas analíticas y su aplicación en Atención Farmacéutica," *Var. Fisiol. Anal. y Patol.*, vol. 1, p. 44, 2009.

[9] G. Goldich, "El electrocardiograma de 12 derivaciones," *Nurs. (Ed. española)*, vol. 32, no. 3, 2015, doi: 10.1016/j.nursi.2015.06.013.

[10] "Taller de interpretación del electrocardiograma. | FISIOLOGÍA." https://fisiologia.facmed.unam.mx/index.php/taller-de-interpretacion-del-electrocardiograma/ (accessed Aug. 31, 2023).

[11] H. Zhou *et al.*, "Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images," *Eur. Radiol.*, vol. 31, no. 6, 2021, doi: 10.1007/s00330-020-07454-9.

[12] I. Gómez Arraiz, E. Barrio Ollero, and A. Gómez Peligros, "Pruebas genéticas en la miocardiopatía hipertrófica: beneficios, limitaciones y aplicaciones en la práctica clínica," *Med. Fam. Semer.*, vol. 44, no. 7, pp. 485–491, Oct. 2018, doi: 10.1016/J.SEMERG.2018.03.002.

[13] S. R. Ommen *et al.*, "2020 AHA/ACC Guideline for the Diagnosis and Treatment of Patients with Hypertrophic Cardiomyopathy: Executive Summary: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *Circulation*, vol. 142, no. 25, pp. E533–E557, Dec. 2020, doi: 10.1161/CIR.0000000000000938.

[14] C. V. Tuohy, S. Kaul, H. K. Song, B. Nazer, and S. B. Heitner, "Hypertrophic cardiomyopathy: the future of treatment," *Eur. J. Heart Fail.*, vol. 22, no. 2, pp. 228–240, Feb. 2020, doi: 10.1002/EJHF.1715.

[15] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical,*

Physical and Engineering Sciences, vol. 374, no. 2065. 2016. doi: 10.1098/rsta.2015.0202.

[16] "Algorithm of Principal Component Analysis (PCA)." https://iq.opengenus.org/algorithm-principal-component-analysis-pca/ (accessed May 21, 2023).

[17] U. Michelucci, "An Introduction to Autoencoders," vol. 1, no. 1986, 2022, [Online]. Available: http://arxiv.org/abs/2201.03898

[18] N. Pezzotti, B. P. F. Lelieveldt, L. Van Der Maaten, T. Höllt, E. Eisemann, and A. Vilanova, "Approximated and user steerable tSNE for progressive visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 7, 2017, doi: 10.1109/TVCG.2016.2570755.

[19] "T-SNE Explained — Math and Intuition | by Achinoam Soroker | The Startup | Medium." https://medium.com/swlh/t-sne-explained-math-and-intuition-94599ab164cf (accessed May 01, 2023).

[20] A. Jackson, "The mathematics of UMAP," 2019.

[21] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018, [Online]. Available: http://arxiv.org/abs/1802.03426

[22] G. Friedman, "Survey article: An elementary illustrated introduction to simplicial sets," *Rocky Mt. J. Math.*, vol. 42, no. 2, 2012, doi: 10.1216/RMJ-2012-42-2-353.

[23] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review*, vol. 52, no. 2. Springer Netherlands, pp. 803–855, Aug. 15, 2019. doi: 10.1007/s10462-018-9614-6.

[24] "Incorporating User Feedback Into One-Class Support Vector Machines for Anomaly Detection | Julien LESOUPLE @ SIGNAV." http://perso.recherche.enac.fr/~julien.lesouple/publication/eusipco2020/ (accessed Mar. 17, 2023).

[25] M. Hejazi and Y. P. Singh, "One-class support vector machines approach to anomaly detection," *Appl. Artif. Intell.*, vol. 27, no. 5, 2013, doi: 10.1080/08839514.2013.785791.

[26] P. Laskov, C. Schäfer, and I. Kotenko, "Intrusion detection in unlabeled data with quarter-sphere Support Vector Machines," in *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, 2004, vol. P 46. doi: 10.1515/piko.2004.228.

[27] "Support Vector Data Description (SVDD) - File Exchange - MATLAB Central." https://es.mathworks.com/matlabcentral/fileexchange/69296-support-vector-data-description-svdd (accessed Mar. 17, 2023).

[28] Z. Zhang and X. Deng, "Anomaly detection using improved deep SVDD model with data structure preservation," *Pattern Recognit. Lett.*, vol. 148, pp. 1–6, Aug. 2021, doi: 10.1016/j.patrec.2021.04.020.