UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA
DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

DISEÑO Y DESARROLLO DE TÉCNICAS AVANZADAS DE INTELIGENCIA ARTIFICIAL PARA LA IDENTIFICACIÓN DE FACTORES DE RIESGO DEL SÍNDROME METABÓLICO

ÁNGELA BURGALETA LEDESMA

2022

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

Título: Diseño y desarrollo de técnicas avanzadas de inteligencia artificial para la identificación de factores de riesgo del Síndrome Metabólico.

Autor: Ángela Burgaleta Ledesma

Tutor: Beatriz Merino

Cotutor: Álvaro Belmar Mas

Ponente: Giuseppe Fico

Departamento: Ist, tfb

MIEMBROS DEL TRIBUNAL

Presidente: D. ……

Vocal: D. ……..

Secretario: D. ……..

Suplente: D. ……..

Los miembros del tribunal arriba nombrados acuerdan otorgar la calificación de: ……

Madrid, a de de 20…

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

DISEÑO Y DESARROLLO DE TÉCNICAS AVANZADAS DE INTELIGENCIA ARTIFICIAL PARA LA IDENTIFICACIÓN DE FACTORES DE RIESGO DEL SÍNDROME METABÓLICO

ANGELA BURGALETA LEDESMA

2022

RESUMEN

El Síndrome Metabólico (SM) es un conjunto de trastornos metabólicos que suponen un factor de riesgo para sufrir diabetes y enfermedades cardiovasculares y cerebrovasculares, patologías que además de ser de las principales causas de muerte a nivel mundial, reducen drásticamente la calidad de vida de los que las padecen. Está conformado por una serie de factores de riesgo, como la hipertensión arterial, la dislipidemia, la intolerancia a la glucosa por la resistencia a la insulina y la obesidad visceral.

La prevalencia del SM aumenta de forma alarmante, diagnosticándose en edades más tempranas cada vez, especialmente en países desarrollados. Por ello ha cobrado paulatinamente más importancia en los últimos años, destinándose progresivamente más recursos a estudiarlo y a tratar de prevenirlo.

Desde los años setenta, el Centro Nacional De Estadística de la Salud de los Estados Unidos, mediante su programa NHANES, realiza encuestas y monitoreo de diversas constantes a una muestra representativa de la población. Anualmente esta muestra puede oscilar entre 5.000 y 10.000 participantes. Estas encuestas incluyen desde datos demográficos, socioeconómicos y dietéticos hasta pruebas de laboratorio y mediciones físicas.

En este Trabajo de Fin de Máster se propone el uso de técnicas de aprendizaje automático aplicadas sobre la Base de Datos NHANES con el objetivo de predecir si una persona es susceptible de sufrir Síndrome Metabólico. Se escoge esta base de datos debido a la gran cantidad de información que recopila anualmente de un grupo muestral suficientemente grande para poder entrenar a los modelos de Inteligencia artificial. Por otro lado, existe una extensa bibliografía de estudios previos que han utilizado NHANES.

En primer lugar, se realizó un estudio sobre el estado del arte de las tecnologías existentes para el desarrollo de investigaciones de Inteligencia artificial predictiva. A continuación, se procedió a desglosar los atributos más relevantes del NHANES, identificando aquellos que puedan resultar más interesantes de formar parte del presente estudio. A su vez, se efectuó una limpieza y un preprocesado de los datos con el fin de llevar a cabo un análisis estadístico entre las variables escogidas. Más tarde, se emplearon y optimizaron distintos modelos de Aprendizaje automático que posibiliten la predicción de padecer Síndrome Metabólico. Finalmente, se realizaron pruebas de validación que verificaron que el desarrollo llevado a cabo cumplía con las especificaciones previstas y lograba su cometido.

En último lugar, se detallaron las conclusiones obtenidas y se planteó la viabilidad de los modelos para ser aplicados en simulaciones médicas que permitan reducir el tiempo de detección del Síndrome Metabólico.

SUMMARY

The Metabolic Syndrome (MS) is a set of metabolic abnormalities that are a risk factor for diabetes and cardiovascular and cerebrovascular diseases, pathologies that besides being the leading causes of death worldwide, drastically reduce the quality of life of those who suffer from them. It consists of a series of risk factors, such as arterial hypertension, dyslipidemia, glucose intolerance due to insulin resistance, and visceral obesity.

The prevalence of MS is increasing at an alarming rate and is being diagnosed at increasingly younger ages, especially in developed countries. For this reason, it has become increasingly important in recent years, with more and more resources being devoted to studying it and trying to prevent it.

Since the 1970s, the National Center for Health Statistics of the United States, through its NHANES program, has been conducting surveys and monitoring of various constants in a representative sample of the population. Annually, this sample can range from 5,000 to 10,000 participants. These surveys range from demographic, socioeconomic and dietary data to laboratory tests and physical measurements.

In this Master Thesis, we propose the use of Machine learning techniques applied to the NHANES database to predict whether a person is susceptible to suffer from Metabolic Syndrome. This database was chosen because of the large amount of information it collects annually from a sufficiently large sample group to be able to train the artificial intelligence models. On the other hand, there is an extensive bibliography of previous studies that have used NHANES.

First, a study was conducted on the state of the art of existing technologies for the development of predictive artificial intelligence research, especially those focused on preventive medicine. Next, the most relevant attributes of NHANES were broken down, identifying those that could be of most interest for this study. At the same time, the data were cleaned and preprocessed in order to carry out a statistical analysis among the chosen variables. Later, different Machine Learning models were used and optimized to enable the prediction of suffering from Metabolic Syndrome. Finally, validation tests were carried out to verify that the development carried out complied with the specifications foreseen and achieved its purpose.

Finally, the conclusions obtained will be detailed and the feasibility of the models to be applied in medical simulations to reduce the detection time of the Metabolic Syndrome will be discussed.

PALABRAS CLAVE

Inteligencia Artificial, Síndrome Metabólico, red neuronal, aprendizaje automático, modelo predictivo.

KEYWORDS

Artificial intelligence, metabolic syndrome, Neural Network, Machine learning, predictive modelling.

INDICE DEL CONTENIDO

GLOSARIO

- **ATPIII:** Adult Treatment Panel III
- **BBDD:** Base de Datos
- **BMI:** Body Mass Index
- **DL:** Deep Learning
- **DT:** Decision Tree
- **IA:** Inteligencia Artificial
- **IMC:** Índice de Masa Corporal
- **ML:** Machine Learning
- **NHANES:** The National Health and Nutrition Examination Survey
- **NN:** Neural Network
- **ODS:** Objetivos de Desarrollo Sostenible
- **OMS:** Organización Mundial de la Salud
- **PCA:** Principal Component Analysis
- **RF:** Random Forest
- **SM:** Síndrome Metabólico
- **SVM:** Support Vector Machine
- **TFM:** Trabajo de Fin de Máster
- **XGBOOST:** Extreme Gradient Boosting

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN Y OBJETIVOS

1.1. ANTECEDENTES: SÍNDROME METABÓLICO

El Síndrome Metabólico (SM) es un conjunto de alteraciones metabólicas constituido por la obesidad abdominal, la disminución de las concentraciones de colesterol unido a las lipoproteínas de alta densidad, la elevación de las concentraciones de triglicéridos, el aumento de la presión arterial y la hiperglucemia [1]

La suma de hipertensión, obesidad, dislipemia y diabetes es un concepto que lleva estando presente mucho tiempo en los entornos clínicos. Se trata de una asociación de problema de salud que pueden aparecer simultáneamente o de manera secuencial en un mismo individuo. Están causados por una combinación de factores genéticos y un estilo de vida poco saludable. [2]

No obstante, a lo largo de la historia la definición del SM ha ido sufriendo modificaciones. Oficialmente, surge en 1988 cuando Gerald Reaven llamó "síndrome X" a la agrupación de resistencia a la insulina, dislipidemia e hipertensión. Pero la presentación conjunta de estos síntomas ya se recoge bajo diferentes términos en épocas todavía más antiguas, tales como "síndrome plurimetabólico" u "obesidad diabetógena". Toda esta compleja terminología sumada a una falta de consenso ha conllevado una gran dificultad a la hora de comparar estudios. En el 1998, la OMS (Organización Mundial de la Salud) acuña el término de Síndrome Metabólico como entidad diagnóstica con criterios definidos para referirse a estas alteraciones [2]. Aunque la definición más utilizada es la propuesta por el ATP III [3]. En la Figura 1 se muestra esta comparativa según diferentes organizaciones.

Organismo	OMS (1998)	EGIR (1999)	ATP III (2001)	AACE (2003)	IDF (2005)	AHA/NHLBI (2005)
Criterio principal	AGA, DM2 o RI Dos o más de los siguientes:	Hiperinsulinemia ayunas: >P75 (no diabéticos) Dos o más de los siguientes:	Ninguno Tres o más de los siguientes:	AGA Más cualquiera de los siguientes según juicio clínico:	Obesidad abdominal	Ninguno Tres o más de los siguientes:
Obesidad Abdominal	H: CCC>0.9 M: CCC>0.85 y/o IMC >30 Kg/m²	H: PC ≥94 cm M: PC ≥80 cm	H: PC >102 cm M: PC >88 cm	H: PC >102 cm M: PC>88 cm IMC ≥25 Kg/m²	PC elevado según la población (Tabla 1) Dos o más de los siguientes:	H: PC ≥102 cm M: PC ≥88 cm
Dislipemia	TG ≥150 mg/dL H: c-HDL<35 mg/dL M: c-HDL<39mg/dL	TG ≥177 mg/dL c-HDL<39 mg/dL o tratamiento para dislipemia	TG ≥150 mg/dL H: c-HDL<40 mg/dL M: c-HDL <50 mg/dL o tratamiento específico	TG ≥150 mg/dL H: c-HDL<40 mg/dL M: c-HDL <50 mg/dL	TG ≥150 mg/dL H: c-HDL<40 M:c-HDL<50 o tratamiento específico	TG ≥150 mg/dL H: c-HDL<40 M:c-HDL<50 o tratamiento específico
Presión Arterial (PA)	≥140/90 mmHg	≥140/90 mmHg o con anti-hipertensivos	≥130/85 mmHg o con anti-hipertensivos	≥130/85 mmHg	≥130/85 mmHg o con anti-hipertensivos	≥130/85 mmHg o con anti-hipertensivos
Glicemia	AGA o DM2	≥110 mg/dL	≥110 mg/dL o tratamiento antidiabético	AGA, pero no diabetes mellitus	≥100 mg/dL incluyendo diabéticos	≥100 mg/dL o con antidiabéticos
Otros	Microalbuminuria			Otras (Definición AACE)		

AGA: Alteración de la glucemia en ayunas; DM2: diabetes mellitus tipo 2; RI: resistencia insulínica; CCC: Cociente entre el perímetro de la cintura y el perímetro de la cadera; PC: Perímetro de cintura; c-HDL: Colesterol unido a proteínas de alta densidad; PA: Presión arterial; TG: Triglicéridos; P75: Percentil 75; H: hombres; M: Mujeres.

Figura 1: Diferentes criterios para SM según distintas organizaciones [1]

1.1.1. SELECCIÓN DE CRITERIOS

Hoy en día no existe un consenso universal para definir el Síndrome Metabólico. En este trabajo los criterios se eligen conforme a la definición publicada en la declaración científica conjunta sobre el Síndrome Metabólico [4]. Para ello, se tienen que dar en un individuo tres de los cinco requisitos expuestos a continuación:

1) Perímetro abdominal mayor o igual de 88cm para mujeres y de 102cm para hombres.
2) Triglicéridos con una concentración mayor o igual a 150mg/dL o encontrarse en tratamiento farmacológico para los triglicéridos elevados.
3) Colesterol HDL de menos de 40mg/dL para los hombres y de 50mg/dL para las mujeres o encontrarse en tratamiento farmacológico para el HDL bajo.
4) Presión arterial sistólico de mayor o igual a 130mmHg, diastólica de más de 85mmHg o ambas, o encontrarse en tratamiento farmacológico antihipertensivo.
5) Glucosa en ayunas de más de 100mg/dL o encontrarse en tratamiento farmacológico por niveles elevados de glucosa en sangre.

1.2. DESCRIPCIÓN DEL PROBLEMA

El SM se ha convertido en un problema de salud pública generalizado, especialmente en los países desarrollados. Esta enfermedad se asocia directamente con un incremento de hasta cinco veces en la prevalencia de diabetes tipo dos y de hasta tres veces en enfermedad cardiovascular [5].

La evolución biológica no sufre un avance tan rápido como la cultural, los genes humanos siguen adaptados al entorno de hace miles de años. Nuestros antepasados se enfrentaban en numerosas ocasiones a complicaciones meteorológicas o ambientales que provocaban que conseguir comida mediante la caza, pesca o recolección fuera un trabajo arduo que implicaba bastante actividad física. Actualmente, los circuitos de recompensa del cerebro siguen experimentando gran activación ante la ingesta de alimentos (especialmente los que se asocian con un aporte energético alto) [7], [8]con la ventaja añadida de que se encuentran siempre disponibles, sin necesidad de realizar prácticamente ningún esfuerzo [6], [7].

La obesidad es tan antigua como la humanidad, durante la prehistoria, cuando una de las principales causas de muerte era la hambruna, la selección natural eligió a los individuos que podían “ahorrar” la mayor cantidad de energía de los alimentos en forma de reservas de grasa [8]. Es evidente que conforme hemos ido modificando el entorno para acomodarlo a nuestras necesidades, esta particularidad se ha vuelto desadaptativa. Ya Hipócrates (460-370 a.C.) [9] puntualizó en su momento que la muerte súbita es más común entre los que padecen sobrepeso.

La prevalencia global de Síndrome Metabólico según criterios de la Organización Mundial de la Salud es del 36,8% [10]. La presentación de dicho síndrome es ligeramente superior en los varones (54,1% frente al 52,8%), y aumenta de forma paralela con la edad, así, supera el 64,1% en los mayores de 59 años, resultando esta asociación estadísticamente significativa.

Para llegar a presentar un cuadro clínico grave de SM es necesario perpetuar unos hábitos de vida poco saludables (junto a una serie de combinaciones genéticas) durante muchos años. La

manifestación en cadena de los síntomas no se puede predecir con seguridad, debido a que los pacientes que los sufren no se percatan de ello mediante mecanismos endógenos obvios, como por ejemplo el dolor. A pesar de esto, es una patología que se previene mediante alimentación adecuada, ejercicio físico, control del estrés, un descanso suficiente y evitación, en la medida de lo posible, el tabaco y el alcohol. De hecho, también se puede revertir con las mismas herramientas si se diagnostica a tiempo [11].

La dificultad en el pronóstico del Síndrome Metabólico hace que resulte muy útil identificar factores de riesgo en esta patología. Un tratamiento temprano de la enfermedad aumenta la probabilidad de éxito junto con un ahorro en la sanidad pública, hay que tener en cuenta que implementar un estilo de vida saludable no supone prácticamente ningún coste adicional, pero si la enfermedad se agrava es bastante probable que el individuo necesite asistencia médica de por vida [12].

Por otro lado, de los estudios de Yang et al. [13], Gregory E Miller et al. [14] y María del Carmen Navarro et al. [15] se concluye que el impacto individual, familiar y social es de gran relevancia, y es necesario subrayar que la obesidad (factor de riesgo principal para padecer SM) como enfermedad de carácter epidémico no se distribuye de manera homogénea por razón de sexo, clase social, nivel de estudios ni zona de residencia (rural/urbana); por tanto, sus consecuencias negativas tampoco se distribuyen homogéneamente [13].

La identificación de la obesidad como problema y desafío de las próximas décadas está empezando a reconocerse, tanto entre la sociedad como políticamente. A pesar de esto, hay una falta de políticas públicas en este contexto debido al desconocimiento generalizado de la enfermedad. No hay una conciencia social de prevención del SM. Por otro lado, el gasto sanitario históricamente se ha destinado al tratamiento de las enfermedades cuando estas ya se encuentran en un estado avanzado [16]–[18] . Llevar a cabo la formulación y la puesta en marcha de planes de acción en el contexto de una política para la nutrición y la actividad física exige un claro y actualizado conocimiento de los patrones de consumo alimentario y de la actividad física de la población, así como de las múltiples políticas, directas, indirectas y no intencionadas, a las cuales se dirigen estos instrumentos de salud pública [19], [20].

1.3. OBJETIVOS

Este trabajo pretende identificar la probabilidad de que un individuo padezca SM implementando técnicas de inteligencia artificial.

La finalidad principal es abordar una serie de modelos predictivos basados en aprendizaje automático (ML, por sus siglas en inglés) que permitan detectar los factores de riesgo asociados con el Síndrome Metabólico. Este desarrollo comenzará con modelos sencillos que se refinarán hasta alcanzar el mejor rendimiento posible y el máximo nivel de aciertos.

Los objetivos específicos se presentan a continuación:

1) Búsqueda y exploración de una base de datos con información detallada a nivel de paciente que contenga los suficientes registros como para ser adecuada en el entrenamiento de modelos de ML.

2) Obtención de la prevalencia de SM en la muestra recogida por la base de datos.
3) Selección de los factores de interés de la base de datos para inyectarlas en los modelos predictivos.
4) Análisis, limpieza y preprocesado de los campos seleccionados de la base de datos para establecer su relación con el SM.
5) Estudio, selección, implementación y optimización de los diferentes algoritmos de aprendizaje automático para resolver el problema de predicción del SM.
6) Validación técnica y de experiencia de usuario. Observación y obtención de la evaluación de profesionales de la salud sobre el potencial interés que puede suponer esta herramienta aplicada a su entorno laboral.

Este trabajo contribuye a los siguientes Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 [21]:

- Garantizar una vida sana y promover el bienestar para todos en todas las edades. Este trabajo supone un avance directo a la hora de aumentar la esperanza de vida de la población y reducir algunas de las causas de muerte más comunes [22].
- Fomentar la innovación promoviendo nuevas tecnologías que permitan el uso eficiente de recursos. La innovación y el progreso tecnológico, mediante técnicas como las expuestas en el presente documento, son claves para descubrir soluciones duraderas para los desafíos económicos [23].

2. ESTADO DEL ARTE

En este apartado se procede a hacer un breve repaso de las bases teóricas del TFM (Trabajo de Fin de Máster). Se enfoca en las tecnologías sobre las que se sustenta el desarrollo: análisis de datos, *big data* y aprendizaje automático, así como estudios recientes sobre el Síndrome Metabólico.

2.1. ESTADO DEL ARTE DEL ANÁLISIS DE DATOS

El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos para obtener información útil. Sirve principalmente para comprender situaciones y extraer conclusiones. Se centra en la inferencia estadística y permite tomar Decisiones con cierto grado de confianza. Tiene múltiples facetas y enfoques, se aplica en numerosas situaciones tales como los negocios, la mercadotecnia, las ciencias sociales y puras o la medicina.

En la era de la información el análisis de datos cobra un papel fundamental. Ya lo decía el matemático Clive Humby en 2006, *"los datos son el nuevo petróleo"*, sentencia que amplió Michael Palmer diciendo que *"los datos son valiosos, pero si no están refinados, en realidad no se pueden usar"* [24]. Lo cierto es que existe una diferencia entre el petróleo y los datos. En su momento el petróleo era controlado por un reducido grupo de personas; sin embargo, contamos con una gran cuantía de datos disponibles para cualquier persona, empresa u organización. La tecnología permite recopilar una cantidad masiva de datos que luego pueden utilizarse para obtener una compresión general de diversas situaciones, anticiparse y tomar mejores Decisiones.

Todo esto se debe a la aparición de diversos factores como Internet, las plataformas en la nube, el aumento de la potencia de computación, las redes sociales, el descenso en los precios de almacenamiento y de la potencia de cálculo y los algoritmos de aprendizaje automático que se encuentran a disposición de cualquier usuario [25].

Se podría establecer una primera división entre los tipos de análisis de datos [26] [27].

- **Cualitativos**: se presentan, generalmente, en forma de gráficas y se basan en la interpretación. Analizan patrones mediante la observación a lo largo del proceso de la recolección de datos.
- **Cuantitativos**: se presentan en forma numérica y se basan en resultados tangibles.

Además de esta clasificación, existen diferentes métodos que se emplean en función de los objetivos que se deseen alcanzar [28].

- **Análisis descriptivo**: es el punto de partida para cualquier reflexión analítica. Permite organizar los datos y prepararlos para llevar a cabo nuevas investigaciones.
- **Análisis diagnóstico**: diseñado para proporcionar respuestas directas y procesables a preguntas concretas. Permite encontrar conexiones y generar hipótesis.
- **Análisis predictivo**: permite descubrir tendencias futuras.
- **Análisis prescriptivo**: se enfoca en la identificación y uso de patrones o tendencias para desarrollar estrategias empresariales prácticas y con alta capacidad de respuesta.

2.1.1 INTELIGENCIA ARTIFICIAL

El campo de la inteligencia artificial es un conjunto de algoritmos que trata de simular la inteligencia humana. Con el transcurso de los años el interés por la IA ha aumentado de manera exponencial, dando lugar al desarrollo de novedosas aplicaciones en multitud de campos. No obstante, este concepto se remonta al 1860 con el piano lógico de Jeavons, aunque el término de Inteligencia artificial nace en 1956 de la mano del informático John McCathy [29] . En la Figura 2 se muestran los acontecimientos más relevantes en orden cronológico de la IA.

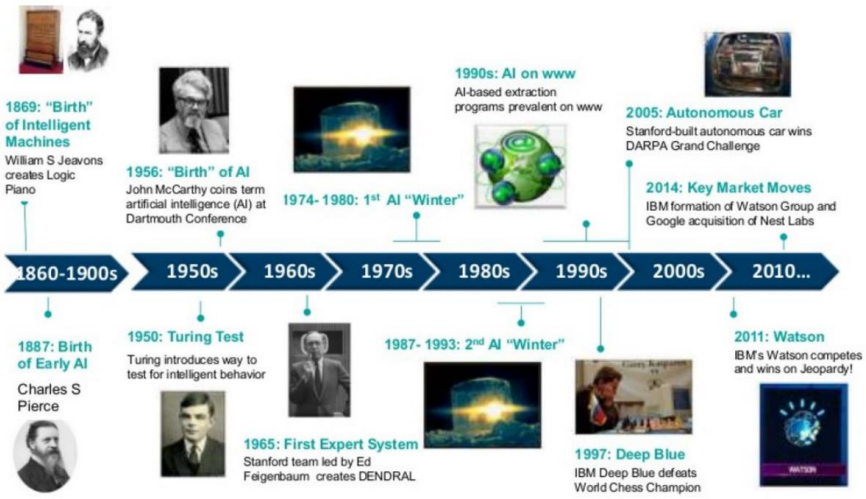

Figura 2: Historia de la IA [140]

Existen diferentes campos de aprendizaje en IA, se diferencian en la forma de analizar y manipular los datos. La Inteligencia artificial engloba al aprendizaje automático y al aprendizaje profundo [30], tal y como se muestra en la Figura 3.

El aprendizaje automático es una rama de la inteligencia artificial que combina grandes volúmenes de datos mediante un procesamiento rápido e iterativo y algoritmos inteligentes. De esta forma el programa es capaz de aprender automáticamente los patrones existentes entre los mismos y realizar predicciones, sin estar expresamente programado para ello [31]. Se observa el diagrama de flujo del aprendizaje automático en la Figura 4.

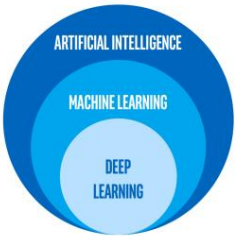

Figura 3: Campos de la IA [140]

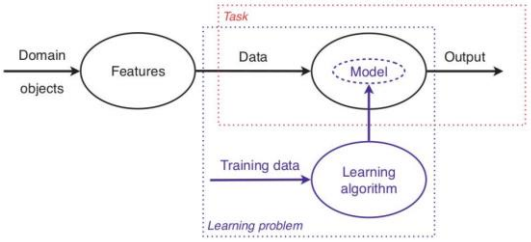

Figura 4: Diagrama de flujo del aprendizaje automático [140]

El aprendizaje profundo usa redes neuronales artificiales para llevar a cabo el proceso de ML citado antes. Posee diferentes niveles de aprendizaje. Al inicio la red neuronal aprende algo simple y lo manda al siguiente nivel, a medida que se avanza en los niveles se elaboran informaciones más complejas. Se denomina profundo porque se constituye de varias capas, como se observa en la Figura 5 y requiere tiempos de aprendizaje más largos que su predecesor [32], [33].

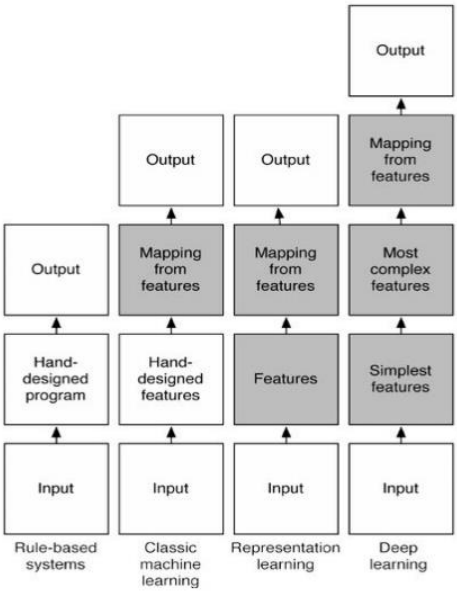

Figura 5: Esquema en aprendizaje profundo [140]

2.2. APLICACIÓN DEL ANÁLISIS DE DATOS EN EL SECTOR SANITARIO

En 1854 el doctor John Snow logró establecer la asociación entre el cólera y el consumo de agua contaminada utilizando técnicas de cartografía y matemáticas relativamente sencillas, sentando así las bases de la epidemiología [34] .

Esta relativa simplicidad contrasta con los desarrollos de los últimos años dados por la disponibilidad de una masiva cantidad de datos de diversas fuentes, así como el desarrollo de tecnologías, incluida la Inteligencia artificial, apalancados por los avances en la biología molecular y la genómica, los cuales, dependen en gran medida del soporte de la ingeniería y de la infraestructura computacional.

Este nuevo entorno del acceso a diferentes herramientas tecnológicas se presenta como una oportunidad para su uso en epidemiología y salud pública, con beneficios tales como la integración de diferentes fuentes de datos, la eficiencia en su análisis, y en muchos casos en los costos derivados de su utilización [35]

La implementación del análisis de datos en salud da lugar a diversas aplicaciones tales como: estudio de la transmisión de enfermedades infecciosas, seguridad farmacológica, gestión de recursos humanos y suministros, vigilancia de la salud, diagnóstico automatizado por imagen o predicción de factores de riesgo [36].

En el sector salud existen numerosas fuentes de datos heterogéneas que arrojan una gran cantidad de información relacionada con los pacientes, las enfermedades y los centros sanitarios. La aplicación de las denominadas técnicas de análisis de datos permite inferir una capa de inteligencia, en la que resulta de especial relevancia la aplicación de modelos predictivos que ayuden a anticiparse a las necesidades sanitarias para posibilitar que se ofrezca una atención médica más eficaz [37].

El análisis de datos es crucial para la investigación en salud. Ha permitido acelerar el desarrollo de nuevos fármacos y optimizar la eficiencia de los ensayos clínicos. La implementación continuada del análisis de datos en todos los aspectos del sector supone una clara ventaja para su gestión y mejora, impactando finalmente de forma positiva en la calidad de vida de los pacientes. Esto ha dado lugar a diversas aplicaciones [38]:

- **Mejoras en la gestión de salud:** Los hospitales y centros de atención sanitaria se han beneficiado de la combinación de datos administrativos y financieros que proporciona el análisis de datos en varios aspectos.

  -Detectando qué procesos no producen los resultados esperados y cómo mejorarlos.

  -Mejorando la coordinación entre departamentos y recursos, lo cual hace que los datos del paciente sean más accesibles para el mismo y se compartan de manera más rápida, segura y precisa.

  -Proponiendo soluciones individualizadas para cada paciente de forma que mejore su calidad de vida.

-Realizando mejoras con el objetivo de adelantarse a posibles crisis, como la pandemia del COVID-19, mediante el análisis predictivo.

- **Tecnología *wearable*:** Conjunto de aparatos y dispositivos electrónicos que se incorporan en alguna parte del cuerpo interactuando de forma continua con el usuario y con otros dispositivos. Algunos ejemplos son los relojes inteligentes, las pulseras que controlan nuestro estado de salud, las zapatillas de deporte con GPS incorporado, los monitores de glucosa o los monitores de electrocardiograma. Además de proporcionar un seguimiento personalizado, facilitan el control de la salud de cada individuo debido a que están diseñados para que su utilización sea sencilla para el público.
- **Gestión de posibles crisis:** Durante periodos de crisis es fundamental tomar rápidas Decisiones a gran escala contando con recursos limitados. El hecho de conocer y analizar los datos médicos en este tipo de escenarios aporta una gran ventaja a la hora de gestionar crisis sanitarias de gran magnitud.
- **Predicción:** Actualmente, la utilización de modelos predictivos en enfermedades como la diabetes [39] está ampliamente extendido, contribuyendo a un diagnóstico precoz que permite atajar la enfermedad lo más temprano posible. Algunos de los modelos de ML que más se emplean para esta causa son: árboles de decisión, Random Forest, Support Vector Machines y redes neuronales[40]. A continuación, se realiza un sucinto resumen sobre estos modelos.

2.2.1 SUPPORT VECTOR MACHINE

Las Máquinas de Vector de Soporte (Support Vector Machine, SVM, en su nombre original en inglés) son un conjunto de métodos supervisados de aprendizaje automático dedicados a tareas de regresión y clasificación [41]. Fueron desarrollados por Vladimir Vapnik y su equipo en AT&T.

La idea básica se sostiene en que dado un conjunto de puntos en el que cada uno de ellos pertenece a una categoría, un algoritmo basado en SVM construye un modelo capaz de predecir a qué categoría pertenece un punto nuevo, se ejemplifica en Figura 6. La SVM busca una línea recta, un plano o un hiperplano (función kernel) que separe de forma óptima a los puntos de una clase de las de otras, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior. Las mayores ventajas de este modelo son:

- Resultan muy eficaces en espacios de muchas dimensiones.
- Resultan muy eficaces en casos donde el número de dimensiones es mayor que el número de muestras.
- Es eficiente en cuanto a memoria: utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados Vectores de soporte).
- Es muy versátil: se pueden especificar diferentes funciones del kernel para la función de decisión. Se proporcionan los kernels comunes, pero también es posible especificar kernels personalizados.

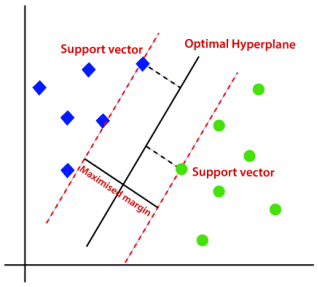

Figura 6: Funcionamiento SVC [140]

2.2.2 DECISION TREE

El árbol de decisión (Decision Tree, en su nombre original en inglés, y DT por sus siglas) es un método de aprendizaje supervisado no paramétrico que se utiliza para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas a partir de las características de los datos, como se ve en la Figura 7. Un árbol puede verse como una aproximación constante a trozos [42]. Algunas de sus mayores ventajas radican en que:

- Son fáciles de interpretar porque son muy visuales.
- Requiere poca preparación de los datos.
- El coste computacional del árbol es logarítmico en el número de puntos de datos de entrenamiento.
- Maneja tanto datos numéricos como categóricos.
- Maneja problemas de salidas múltiples.

No obstante, este modelo también presenta algunas desventajas:

- Se puede crear un árbol demasiado complejo que no generalice bien los datos, generando sobreajuste (*overfitting,* término en inglés).
- Los árboles de decisión pueden ser inestables porque pequeñas variaciones en los datos pueden hacer que se genere un árbol completamente diferente.
- Hay conceptos que son difíciles de aprender porque los árboles de decisión no los expresan fácilmente, como los problemas XOR, de paridad o de multiplexores.
- Los aprendices de árboles de decisión crean árboles sesgados si algunas clases son dominantes. Por lo tanto, se recomienda equilibrar el conjunto de datos antes de ajustarlo con el árbol de decisión.
- Los algoritmos prácticos de aprendizaje de árboles de decisión se basan en algoritmos heurísticos como el algoritmo codicioso, en el que se toman Decisiones localmente óptimas en cada nodo. Estos algoritmos no pueden garantizar la obtención del árbol de decisión globalmente óptimo. Esto puede mitigarse entrenando múltiples árboles en un aprendiz de conjunto, donde las características y las muestras se muestrean aleatoriamente con reemplazo.

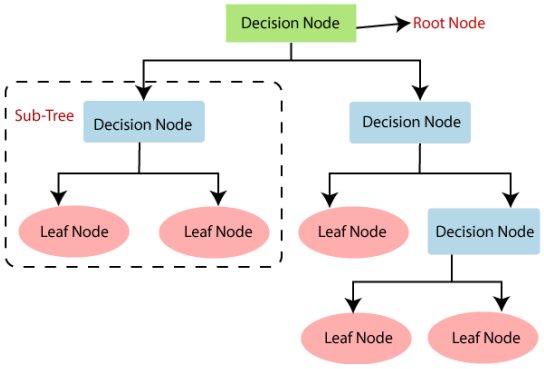

Figura 7: Funcionamiento Decision Tree [140]

2.2.3 RANDOM FOREST

El bosque aleatorio (Random Forest, en su nombre original en inglés) es un conjunto de árboles de decisión que se crean para resolver el problema de sobreajuste, se trata de un algoritmo de aprendizaje automático supervisado empleado para clasificación y regresión [43], [44]. Un Random Forest es un conjunto de árboles de decisión combinados con *bagging*. El *bagging* consiste en crear diferentes modelos usando muestras aleatorias con reemplazo para luego ensamblar los resultados, se puede observar en la Figura 8. Al usar *bagging*, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor [45].

Para problemas de clasificación, se suelen combinar los resultados de los árboles de decisión usando voto suave (*soft-voting,* término original en inglés). En el voto suave, se le da más importancia a los resultados en los que los árboles estén muy seguros.

Para problemas de regresión, la forma más habitual de combinar los resultados de los árboles de decisión es tomando su media aritmética.

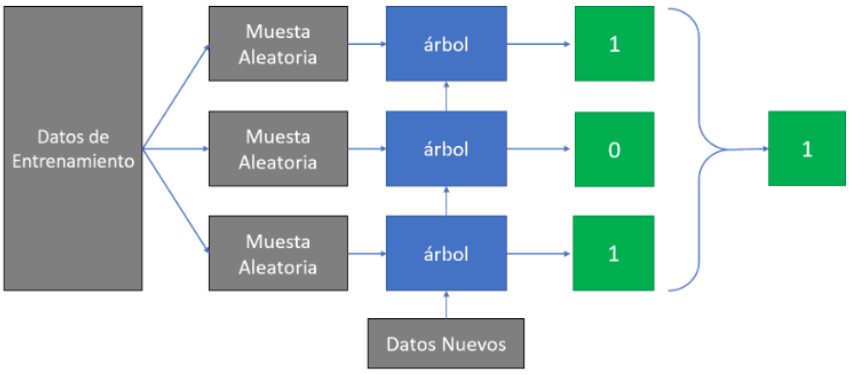

Figura 8: Funcionamiento Random Forest [140]

2.2.4 XGBOOST

Extreme Gradient Boosting es una biblioteca de aprendizaje automático supervisado. Implementa algoritmos en el marco de *gradient boosting*, que consiste en una técnica usada para regresión, clasificación y ordenación. Se basa en la combinación de modelos predictivos débiles, típicamente árboles de decisión, para crear un modelo predictivo fuerte (Figura 9). Se lleva a cabo una generación secuencial de árboles de decisión de forma que los nuevos corrigen los errores de los anteriores. Suele ser el algoritmo que mejores resultados ofrece [46]

De esta forma, el *boosting* puede ser interpretado como un algoritmo de optimización en una función de coste adecuada. Sus principales ventajas son [47]:

- Tiempos de ejecución rápidos.
- Aumenta el rendimiento del modelo.
- Reduce los errores del modelo.

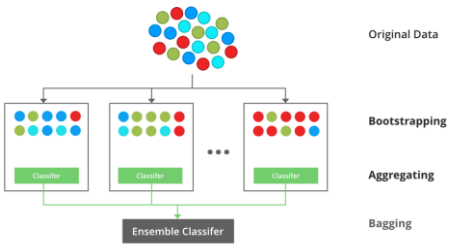

Figura 9: XGBoost esquematizado [43]

2.2.5 NEURAL NETWORK

Las redes neuronales (Neural Network, por su nombre original en inglés) artificiales se engloban dentro del aprendizaje profundo y tratan de imitar el funcionamiento de las redes neuronales de los organismos vivos. Un conjunto de neuronas conectadas entre sí que trabajan conjuntamente para dar respuesta a cierta tarea (Figura 10). Los nodos van creando y reforzando sus conexiones para aprender algo que acabe integrándose en el tejido.

El perceptrón es la unidad fundamental de la red neuronal, es un elemento que tiene varias entradas con cierto peso asociado. Si la suma de esas entradas por cada peso es mayor que un determinado número, la salida del perceptrón es un uno. Si es menor, la salida es un cero. Una red neuronal se forma de varias capas de perceptrones [48], [49].

El alcance de las funciones de las redes neuronales es muy extenso, debido a su funcionamiento, son capaces de aproximar cualquier función existente con el suficiente entrenamiento. Principalmente las redes neuronales son utilizadas para tareas de predicción y clasificación. Su rango de actuación es amplio y de gran utilidad hoy en día, se utilizan para aplicaciones de Industria 4. 0 [50], en otras áreas como la economía, en la que pueden ayudar a predecir cuanto van a variar los precios a lo largo de los años [51], o en medicina [52] donde son de gran ayuda para diagnosticar diversos problemas de salud.

Las redes neuronales están presentes en casos tan famosos como el recomendador de YouTube [53], los precios dinámicos de Amazon [54], la identificación de riesgos en banca o en la personalización de estrategias de marketing [55].

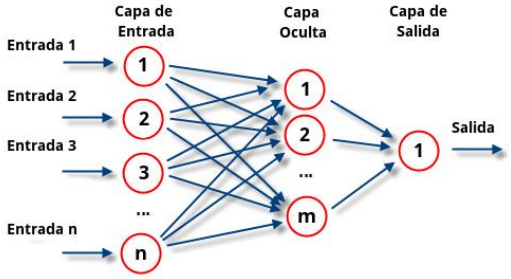

Figura 10: Funcionamiento Neural Network [140]

2.3. ESTADO DEL ARTE DEL SÍNDROME METABÓLICO

El Síndrome Metabólico (SM) es un tema actual y de debate en la comunidad médica, su enfoque es esencial, pues se relaciona con las enfermedades que causan mayor mortalidad a nivel mundial (diabetes y enfermedad cardiovascular [56]). Su incidencia va en aumento [57]. Cuando se inició la transición epidemiológica y principalmente a partir de los primeros hallazgos sobre los estudios de población de Framingham, iniciados por el Servicio de Salud Pública en EEUU en el 1948, se puso mucho interés en la identificación temprana de eventos cardiovasculares. Estas investigaciones consisten en estudios de cohorte observacional que evaluaban clínicamente factores de riesgo biológico y de estilo de vida sobre los resultados de enfermedades cardiovasculares [58]

El primer esfuerzo por introducir el SM a la práctica clínica lo hizo en 1998 el grupo de la Organización Mundial de la Salud (OMS). Este grupo enfatizó el papel central de la resistencia a la insulina, que es difícil de medir en la práctica diaria, pero aceptó evidencias indirectas, como la alteración de la glucosa en ayunas[59]

La prevalencia del SM varía según factores como género, edad y etnia, pero se ubica entre 15% a 40%, siendo mayor en las comunidades hispanas [60]. En un estudio español de G. O. Gutiérrez-Esparza et al. [61] se comparó la prevalencia del SM en la misma población utilizando tres de los criterios (apartado 1.1.1). Aquí se encontró que se representaba con más frecuencia en hombres que en mujeres y aumentaba con la edad, sin importar el criterio utilizado.

Cabe destacar el artículo de J. X. Moore et al. [62] publicado en la revista *Preventing Chronic Disease* que analiza la tendencia del Síndrome Metabólico usando la misma base de datos que en el presente trabajo, NHANES. Se hizo uso de los periodos de 1988-1994, 1999-2006 y 2007-2012 para estudiar la tendencia del SM por raza y sexo. Esta investigación concluye que la prevalencia del SM crece sustancialmente en el tercer ciclo hasta llegar a un 34.2% siendo más frecuente en varones negros no hispanos, seguido de mujeres no hispanas blancas y finalizando con mujeres no hispanas negras. Es interesante puntualizar que este síndrome también se observa en personas no obesas. La probabilidad de sufrir la patología se incrementa con la edad y se aprecia un aumento sustancial en enfermedades asociadas.

Existen varias investigaciones que utilizan diferentes algoritmos de aprendizaje automático para aplicarlos en el Síndrome Metabólico [63]. En Taiwán, en el 2018, nos encontramos con un artículo de C. S. Yu et al. [64] que estudia la aplicación de diferentes árboles de decisión para predecir el SM en pacientes examinados con FibroScan, comparando la precisión entre ellos. Otro estudio en Irán, llevado a cabo por Farzaneh Karimi-Alavijeh et al. [65] en 2016, emplea árboles de decisión y SVM para predecir la incidencia en 7 años del SM [62]. Otra investigación de E. K. Chloe et al. de 2018 [66] propone modelos de ML en la predicción de la prevalencia del Síndrome Metabólico en población no obesa. Finalmente, en una revisión sistemática realizada por H. A. Kakaud et al. [67] en las cinco bases de datos científicas principales (PubMed, Science Direct, IEEE Xplore, ACM digital library y SpringerLink) se obtienen 53 estudios en los que se identifican tres tipos principales de técnicas: estadística (n = 10), ML (n = 40) y cuantificación del riesgo (n = 3). Las pruebas sugieren que las técnicas de ML evaluadas, con una precisión que oscila entre el 75,5% y el 98,9%, pueden diagnosticar el SM con mayor precisión que las técnicas estadísticas y de cuantificación del riesgo.

A pesar de que en la bibliografía ya se encuentran modelos de inteligencia artificial aplicadas en la investigación del Síndrome Metabólico, suelen estar orientadas al aumento de prevalencia de la enfermedad a lo largo de los años o la prevalencia de la enfermedad en función de características demográficas, como la etnia. En este TFM se ha llevado a cabo la aplicación de estas técnicas en la predicción de la patología, fundamentalmente basada en el cuestionario de hábitos de vida del paciente. Por otro lado, se observa que, en comparación con otras patologías como la diabetes (con una prevalencia del 8,8% [68] o el cáncer, con una prevalencia del 35% [69], el Síndrome Metabólico (afectando a un gran número de personas a nivel mundial, aproximadamente del 30% [70]) está menos presente en los estudios.

En lo que tiene que ver con la prevención de la enfermedad, los hábitos dietéticos juegan un papel muy importante en el desarrollo del Síndrome Metabólico. Las recomendaciones generales clásicas incluyen el control de la obesidad, aumento de la actividad física, disminución de ingesta de grasas saturadas, trans y colesterol, reducción en la ingesta de azúcares simples y aumento en la ingesta de frutas y vegetales [71]. Se ha estudiado la influencia de dietas bajas en hidratos de carbono, dietas ricas en ácidos grasos poliinsaturados y monoinsaturados, la ingesta de fibra, la dieta mediterránea y el índice glucémico con relación al Síndrome Metabólico [71]. Otros nutrientes estudiados recientemente han sido micronutrientes (magnesio y calcio entre otros), soja y otras sustancias fitoquímicas. La evidencia sugiere que una dieta saludable como la dieta mediterránea, protege frente al Síndrome Metabólico [71]

3. MATERIALES Y MÉTODOS

3.1. BASE DE DATOS NHANES

La base de datos National Health and Nutrition Examination Survey (NHANES) [72] lleva operativa desde el 1959 recopilando, analizando y difundiendo información sobre el estado de salud de los residentes de los Estados Unidos. Cada año cerca de 5.000 individuos participan en sus encuestas (anualmente el conjunto muestral sufre un aumento, hasta acercarse a los 10.000 participantes en los ciclos más actuales) que incluyen preguntas demográficas, socioeconómicas, dietéticas, de actividad física, hábitos y otros parámetros relacionados con su salud. Además, cuenta con un reconocimiento médico que consiste en mediciones físicas, dentales y fisiológicas, así como pruebas de laboratorio administradas por personal médico altamente capacitado. Los resultados de esta encuesta han sido ampliamente utilizados para determinar la prevalencia de las principales enfermedades y los factores de riesgo de las mismas. La muestra de la encuesta se selecciona para representar a la población estadounidense de todas las edades y etnias, con la finalidad de generar estadísticas lo más fiables posibles.

Se escoge esta base de datos debido a distintos motivos:

- Por la variedad y la cantidad de datos que aporta. El grupo muestral es lo suficientemente grande como para realizar un análisis estadístico en condiciones y para poder entrenar los modelos de aprendizaje automático desarrollados.
- El procedimiento que utiliza para la recopilación de datos es un seguimiento anual y exhaustivo de los participantes.

- El hecho de llevar desde 1959 elaborando información sobre el estado de salud de la población estadounidense hace que esté lo suficientemente asentada y que cuente con la experiencia necesaria como para confiar en la precisión de sus datos.
- Esta base de datos ya se ha utilizado previamente para elaborar estudios relacionados con diferentes patologías y problemas de salud. Por el tipo de información que ofrece NHANES, ha contribuido especialmente en el estudio de las siguientes cuestiones: Anemia [73], enfermedades cardiovasculares, diabetes, exposiciones ambientales, enfermedades oculares, pérdidas de audición, enfermedades infecciosas, enfermedades renales, nutrición, obesidad [74], salud bucodental, osteoporosis [75], condición física, historial reproductivo, enfermedades respiratorias y enfermedades de transmisión sexual.

3.1.1 EXTRACCIÓN DEL DATAFRAME

Para recolectar los datos, en primer lugar, se localiza el componente del ciclo y los archivos que contengan las variables seleccionadas. Un script de Python se encarga de descargar los archivos en su formato *.xpt* original, convertirlos a *.csv* y realizar la unión de todas las columnas por número de secuencia [76]–[79]

Para llevar a cabo el desarrollo de este trabajo, se han considerado cuatro ciclos de la base de datos: del 2011 al 2012, del 2013 al 2014, del 2015 al 2016 y del 2017 al 2018. En total se han recopilado los resultados de **37.606** individuos tal y como se muestra en la Tabla 1.

Tabla 1: Número de participantes por ciclo

Periodo	Número de participantes
2011 - 2012	9.364
2013 - 2014	9.770
2015 -2016	9.575
2017 - 2018	8.897
TOTAL	**37.606**

A continuación, se muestra la Tabla 2 con las columnas que se han tenido en cuenta para este trabajo, basadas en el interés que han supuesto para los análisis plasmados en el estado del arte y que guardan relación con el objetivo del presente proyecto. NHANES engloba los tipos de datos en diferentes categorías: demográficos, de examinación, de laboratorio y resultados del cuestionario.

Cabe destacar que la variable DMDEDUC2 solo aplica a personas mayores de 20 años y que las pruebas de laboratorio solo se han realizado a personas mayores de 12 años, según la documentación oficial de los datos demográficos de la base de datos NHANES [80] .

Tabla 2: Descripción de las características extraídas de la BBDD NHANES

Nº	Columna	Tipo	Long	Descripción	Valores	Grupo
1	SEQN	Num	8	Número de secuencia	**62161 - 102956**	Demográficos
2	RIAGENDR	Num	8	Género	**1**: Hombre **2**: Mujer	Demográficos
3	RIDAGEYR	Num	8	Edad en años en el momento de la encuesta	**0 – 80, .: Falta** media: 41.76 desviación: 22.55	Demográficos
4	DMDEDUC2	Num	8	Nivel educativo (adultos +20)	**1**: Menos de 9º grado, **2**: 9º-11º grado **3**: Graduado de secundaria **4**: Educación superior **5**: Grado universitario o superior **7**: Se niega a responder, **9**: No sabe **.**: Falta	Demográficos
5	RIDRETH3	Num	8	Etnia	**1**: Mexicano americano **2**: Otra etnia hispana **3**: No hispano blanco **4**: No hispano negro **6**: No hispano asiático **7:** Otra raza incluyendo multirracial	Demográficos
6	BPXSY1	Num	8	Presión sanguínea sistólica (1st rdg) mm Hg	**66 - 238,** **.**: Falta Media: 122.56 Desviación: 18.78	Examinación
7	BPXDI1	Num	8	Presión sanguínea diastólica (1st rdg) mm Hg	**0 - 136**, .: Falta Media: 69, Desviación: 13.6	Examinación
8	BMXWT	Num	8	Peso(kg)	**3.6 - 242**, .: Falta Media: 74.38, Desviación: 27.75	Examinación

Tabla 2b: Descripción de las características extraídas de la BBDD NHANES

Nº	Columna	Tipo	Long	Descripción	Valores	Grupo
9	BMXBMI	Num	8	Índice de masa corporal (kg/m**2)	**11.5 - 86.2**, .: Falta Media: 27.95, Desviación: 7.59	Examinación
10	BMXWAIST	Num	8	Perímetro abdominal (cm)	**40-177.9**, .: Falta Media: 94.8, Desviación: 19.76	Examinación
11	BMXHT	Num	8	Altura (cm)	**80.7 - 204.5**, .: Falta Media: 162.98, Desviación: 16.62	Examinación
12	LBXTR	Num	8	Triglicéridos (mg/dL)	**10 - 4233**, .: Falta Media: 114.94, Desviación: 102.77	Laboratorio
13	LBXIN	Num	8	Insulina (uU/mL)	**0.14 - 682.48**, .: Falta Media: 14.11, Desviación: 19.43	Laboratorio
14	LBXGLU	Num	8	Glucosa en ayunas (mg/dL)	**21 - 479**, .: Falta Media: 109.35, Desviación: 36	Laboratorio
15	LBDHDD	Num	8	HDL colesterol (mg/dL)	**6 - 226**, .: Falta Media: 53.25, Desviación: 15.73	Laboratorio
16	BPQ040A	Num	8	Toma prescripción para la hipertensión	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
17	BPQ030	Num	8	Le han dicho que tiene presión sanguínea elevada más de dos veces	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
18	BPQ050A	Num	8	Toma medicación por presión sanguínea elevada	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario

Tabla 2c: Descripción de las características extraídas de la BBDD NHANES

Nº	Columna	Tipo	Long	Descripción	Valores	Grupo
19	BPQ090D	Num	8	Le han prescrito medicación para el colesterol	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
20	DIQ010	Num	8	El doctor te ha dicho que tienes diabetes	**1**: Si, **2**: No, **3**: Prediabetes, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
21	DIQ070	Num	8	Toma pastillas para la diabetes	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
22	ALQ130	Num	8	Media de bebidas alcohólicas por mes en el último año	**1 a 13**: Rango de valores, **15**: 15 bebidas o más, **777**: Se niega a responder, **999**: No sabe, **.**: Falta.	Cuestionario
23	HSD010	Num	8	Condición de salud general	**1**: Excelente, **2**: Muy buena, **3:** Buena **4:** Mediocre, **5**: Pobre **7**: Se niega a responder, **9**: No sabe, .: Falta.	Cuestionario
24	DBQ700	Num	8	Como de saludable es la dieta	**1**: Excelente, **2**: Muy buena, **3:** Buena **4:** Mediocre, **5**: Pobre **7**: Se niega a responder, **9**: No sabe, . : Falta.	Cuestionario
25	HIQ011	Num	8	Cubierto por seguro de salud	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
26	MCQ080	Num	8	El doctor te ha dicho que tienes sobrepeso	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
27	MCQ010	Num	8	Alguna vez te han dicho que tienes asma	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
28	MCQ220	Num	8	Alguna vez te han dicho que tienes cáncer	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, .:Falta	Cuestionario

Tabla 2d: Descripción de las características extraídas de la BBDD NHANES

Nº	Columna	Tipo	Long	Descripción	Valores	Grupo
29	MCQ300C	Num	8	Cercano a padecer diabetes	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
30	DPQ020	Num	8	Te sientes triste, deprimido o desesperanzado	**0:** Nada, **1**: Algunos días, **2**: Más de la mitad de los días, **3:** Casi todos los días, **7**: Se niega a responder, **9**: No sabe, **.**: Falta	Cuestionario
31	DPQ030	Num	8	Problemas para dormir o dormir demasiado	**0:** Nada, **1**: Algunos días, **2**: Más de la mitad de los días, **3:** Casi todos los días, **7**: Se niega a responder, **9**: No sabe, **.**: Falta	Cuestionario
32	DPQ040	Num	8	Te sientes cansado o con poca energía	**0:** Nada, **1**: Algunos días, **2**: Más de la mitad de los días, **3:** Casi todos los días, **7**: Se niega a responder, **9**: No sabe, **.**: Falta	Cuestionario
33	DPQ050	Num	8	Poco apetito o comer en exceso	**0:** Nada, **1**: Algunos días, **2**: Más de la mitad de los días, **3:** Casi todos los días, **7**: Se niega a responder, **9**: No sabe, **.**: Falta	Cuestionario
34	PAQ605	Num	8	Ejercicio físico vigoroso	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario

Tabla 2e: Descripción de las características extraídas de la BBDD NHANES

Nº	Columna	Tipo	Long	Descripción	Valores	Grupo
35	PAQ620	Num	8	Ejercicio físico moderado	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
36	PAQ635	Num	8	Caminar o bicicleta	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
37	PAQ650	Num	8	Actividades físicas recreativas vigorosas	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
38	PAQ665	Num	8	Actividades físicas recreativas moderadas	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
39	SMQ020	Num	8	Has fumado por lo menos 100 cigarrillos a lo largo de tu vida	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
40	INQ020	Num	8	Beneficios de actividades laborales	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario
41	INDFMMPI	Num	8	Índice de pobreza familiar mensual	**0 a 4.98**: Rango de valores, **5**: Más o igual a 5, **.**: Falta.	Cuestionario
42	INDFMMPC	Num	8	Categoría de pobreza familiar mensual	**1**: Índice de pobreza mensual <=1.30, **2**: 1.30<=: Índice de pobreza mensual <=1.85, **3:** Índice de pobreza mensual >1.85 **7**: Se niega a responder, **9**: No sabe, **.**: Falta	Cuestionario
43	SLQ050	Num	8	¿Alguna vez has informado al doctor de problemas para dormir?	**1**: Si, **2**: No, **7**: Se niega a responder, **9**: No sabe, **.**: Falta.	Cuestionario

Debido a que no existe una variable como tal que aporte información sobre si un individuo presenta o no Síndrome Metabólico, a partir de los criterios comentados en el apartado *1.1.1* se ha añadido una columna denominada MET_SYM, codificada como 0 (no presenta Síndrome Metabólico) y 1 (sí presenta Síndrome Metabólico). Para ello se tienen que cumplir como mínimo 3 de las siguientes 5 condiciones reflejadas en la Tabla 3:

Tabla 3: Requisitos del conjunto de datos para SM

Requisito	Descripción	Criterio según columnas del dataframe
1	Perímetro abdominal mayor o igual a 88cm en mujeres Perímetro abdominal mayor o igual a 102cm en hombres	BMXWAIST => 88cm si RIDAGENDR = 2 (mujer) BMXWAIST => 102cm si RIDAGENDR = 1 (hombre)
2	Triglicéridos mayor o igual a 150mg/dL o Tomar medicamentos para el colesterol elevado	LBXTR => 150|| BPQ090D = 1 (sí)
3	HDL menor o igual a 50 mg/dL en mujeres HDL menor o igual a 40 mg/dL en hombres o Tomar medicamentos para el colesterol HDL bajo	LBDHDD <= 50mg/dL si RIDAGENDR = 2 (mujer) LBDHDD <= 40mg/dL si RIDAGENDR = 1 (hombre) || BPQ040A = yes
4	Presión sanguínea sistólica mayor o igual a 130mmHg y/o Presión sanguínea diastólica mayor o igual a 85mmHg o Tomar medicamentos para la presión sanguínea elevada	BPXSY1 >= 130mmHg || BPXDI1 >= 85mmHg || BPQ050A = 1 (sí)
5	Glucosa en ayunas mayor o igual a 100mg/dL o Tomar medicamentos para la glucosa alta	LBXGLU >= 100mg/dL || DIQ070 = 1 (sí)

Además, se añade una columna adicional que refleja el índice cintura-altura de los participantes (WHI, del inglés *waist height index*), debido a que cada vez hay más evidenciade que podría suponer un marcador incluso más fiable que el IMC a la hora de valorar el grado de salud de un paciente [81]. El índice cintura-altura de un sujeto es la división de la circunferencia de su cintura en centímetros entre su estatura en centímetros. A continuación, en la Tabla 4, se muestra la categoría a la que pertenecería un paciente en función del rango en el que se encontrara este indicador [82], [83].

Tabla 4: Categorías según el índice cintura altura

Niños y adolescentes (hasta 15 años)	Hombre	Mujer	Categoría
<0.34	< 0.34	< 0.34	Extremadamente delgado
0.35 a 0.45	0.35 a 0.42	0.35 a 0.41	Delgado sano
0.46 a 0.51	0.43 a 0.52	0.42 a 0.48	Sano
0.52 a -.63	0.53 a 0.57	0.49 a 0.53	Sobrepeso
0.64 +	0.58 a 0.62	0.54 a 0.57	Sobrepeso elevado
	> 0.63	> 0.58	Obesidad mórbida

3.2 HERRAMIENTAS EMPLEADAS

Las herramientas que se han empleado para llevar a cabo el desarrollo de este TFM son las siguientes:

- **Python:** es el tercer lenguaje de programación más conocido del mundo y el primero en los casos de uso relacionados con el análisis de datos. Es un lenguaje muy flexible que cuenta con multitud de bibliotecas útiles de aprendizaje automático [84].
  - **Numpy:** biblioteca de Python que ofrece funciones precompiladas para rutinas numéricas [85]
  - **Pandas:** biblioteca de Python que permite manipular, analizar y visualizar datos en grandes colecciones [86].
  - **Matplotlip:** biblioteca de Python para exportar gráficos y otras imágenes a formatos Vectoriales. Muy útil para la visualización de gráficas [87].
  - **Seaborn:** biblioteca de visualización basada en Matplotlib que proporciona una interfaz a alto nivel para dibujar gráficos estadísticos atractivos [88]
  - **Scipy:** biblioteca de Python con funciones algebraicas y estadísticas [89].
  - **Scikit-learn:** biblioteca de Python de software libre para aprendizaje automático. Integra una gran cantidad de algoritmos de ML además de mucha documentación [90].
  - **Imbalanced-learn:** biblioteca de Python que ofrece un conjunto de métodos y técnicas de muestreo. Típicamente empleado a la hora de balancear datos. Se distribuye bajo la licencia del MIT que se basa en Scikit-learn [91].
  - **Keras:** biblioteca de redes neuronales de Python de código abierto. Está diseñada para posibilitar la experimentación en más o menos poco tiempo con redes de aprendizaje profundo. Es intuitiva, modular y extensible [92]
- **Jupyter Notebook:** se trata de un entorno de trabajo interactivo que permite crear y compartir documentos que pueden contener código, ecuaciones, material multimedia y texto. Soporta hasta 40 lenguajes de programación diferentes. Su gran popularidad hace que se encuentre documentación muy completa en la web [93].
- **TensorFlow:** es una plataforma de código abierto para el aprendizaje automático desarrollado por Google. Permite detectar y descifrar patrones y correlaciones, mediante técnicas análogas al razonamiento humano [94]

- **Google Colaboratory:** herramienta que permite programar y ejecutar Python en el navegador sin necesidad de configuración previa. Da acceso gratuito a GPUs y posibilita compartir contenido fácilmente [95].

3.3 METODOLOGÍA

Las etapas del desarrollo del presente trabajo son las que se detallan en la Figura 11:

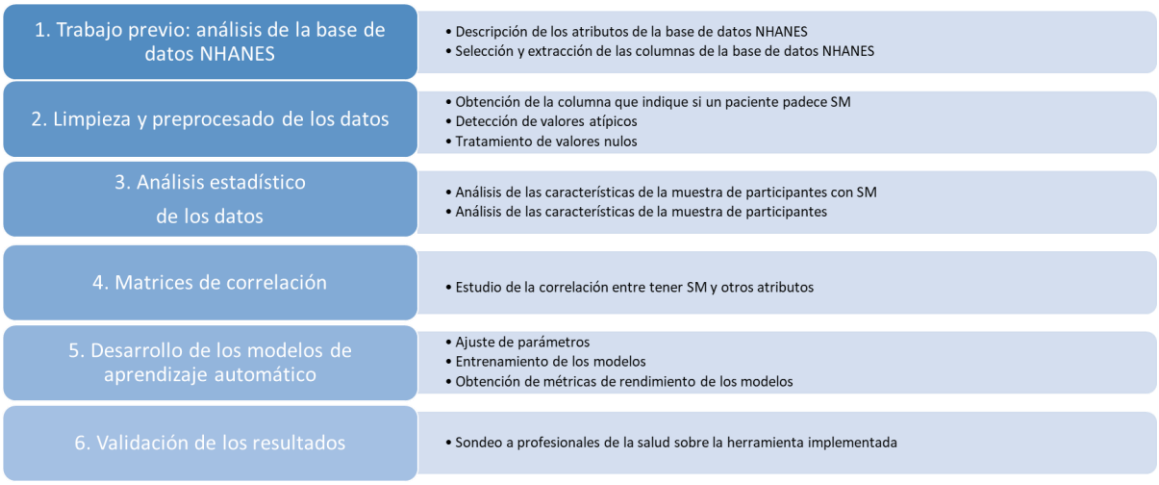

Figura 11: Etapas de la metodología

3.3.1 LIMPIEZA Y PREPROCESADO DE LOS DATOS

El conjunto de datos, por lo tanto, antes del preprocesado cuenta con 37.606 filas y 43 columnas. En primer lugar, se debe realizar una limpieza de los valores nulos. En el campo 'SEQN' correspondiente al id del sujeto no se observan valores duplicados ni valores nulos, es decir, que no hay registros duplicados de pacientes. En la Tabla 5 se muestran las columnas con el número total de valores nulos.

Las decisiones que se han tomado a la hora de procesar los datos y rellenar los campos vacíos para no perder muestras han sido a partir de un criterio optimista. Es decir, no se supone que un paciente presente enfermedad (tanto en respuestas del cuestionario como en presentar niveles atípicos en algún examen médico). Esto puede ocasionar que el porcentaje de individuos con Síndrome Metabólico sea ligeramente mayor en la realidad. Además, la media de los valores perdidos, tanto de cuestionario como de examinación coincide con el criterio empleado.

Tabla 5: Valores nulos de cada columna del dataset

Nº	Columna	Valores nulos	Nº	Columna	Valores nulos
1	SEQN	0	23	DBQ700	3645
2	RIAGENDR	0	24	HIQ011	0
3	RIDAGEYR	0	25	MCQ080	3645
4	DMDEDUC2	5795	26	MCQ010	396
5	RIDRETH3	0	27	MCQ220	5796
6	BPXSY1	5081	28	MCQ300C	5796
7	BPXDI1	5081	29	DPQ020	7629
8	BMXWT	1351	30	DPQ030	7631
9	BMXBMI	2075	31	DPQ040	7635
10	BMXWAIST	3309	32	DPQ050	7635
11	BMXHT	2029	33	PAQ605	3311
12	LBXTR	17490	34	PAQ620	3312
13	LBXIN	17135	35	PAQ635	3312
14	LBXGLU	17442	36	PAQ650	3314
15	LBDHDD	4432	37	PAQ665	3314
16	BPQ040A	19669	38	SMQ020	4900
17	BPQ030	19669	39	INQ020	784
18	BPQ050A	20929	40	HSD010	5916
19	BPQ090D	10571	41	INDFMMPI	4228
20	DIQ010	396	42	INDFMMPC	1483
21	DIQ070	396	43	SLQ050	3645
22	ALQ130	14427			

3.3.1.1 AGRUPACIÓN POR EDAD

Algunos parámetros son muy dependientes tanto del género como de la edad. Por lo tanto, se añade una columna adicional que recoge una distribución en 7 grupos.

I. **1**: De 1 a 5 años.
II. **2**: De 5 a 12 años.
III. **3**: De 12 a 20 años.
IV. **4**: De 20 a 35 años.
V. **5**: De 35 a 50 años.
VI. **6**: De 50 a 65 años.
VII. **7**: De 65 a 80 años.

Se decide descartar el grupo 1 y el 2 al ser un conjunto de participantes que no presentan Síndrome Metabólico debido a su corta edad, además de que en el caso de los participantes menores de 12 años no se recogen pruebas de laboratorio. Se considera, por tanto, que no aportan información relevante a los modelos de aprendizaje automático.

3.3.1.2 LIMPIEZA DE VARIABLES DEMOGRÁFICAS

- **DMDEDUC2:** Tanto para los valores "*missing*" como para aquellos con la opción "no lo sé", se reemplazan los campos por el valor 9, que quiere decir "se niega a contestar" tal y como se muestra en la *Tabla 2*.
- **RIDRETH3**: Se realiza una codificación tipo *"One hot encoding"*, creando seis nuevos atributos binarios correspondientes a las diferentes etnias representadas:
  - **Etnia_1.0:** Mexicano americano.
  - **Etnia_2.0:** Hispano no mexicano.
  - **Etnia_3.0:** Blanco no hispano.
  - **Etnia_4.0:** Negro no hispano.
  - **Etnia_6.0:** Asiático no hispano.
  - **Etnia_7.0:** Otra raza.

3.3.1.3 LIMPIEZA DE VARIABLES NULAS DE EXAMINACIÓN

- **BPXSY1 y BPXDI1**: para la presión sanguínea sistólica y la diastólica se reemplazan los valores nulos por la moda (116 mmHg y 68 mmHg)
- **BMXWAIST**: Corresponde al perímetro abdominal. Se reemplazan los valores nulos por la mediana según sexo y grupo de edad.
- **BMXHT**: Corresponde a la altura. Análogo al caso anterior.
- **BMXWT**: Corresponde al peso. Análogo al caso anterior.
- **BMXBMI**: En este caso se puede realizar el cálculo directamente al haber truncado los valores de la altura y el peso como se muestra en la Ecuación 1.

$$BMXBMI = \frac{peso\ (kg)}{altura\ (m^2)}$$

Ecuación 1: Cálculo del índice de masa corporal

3.3.1.4 LIMPIEZA DE VARIABLES NULAS DE LABORATORIO

- **LBXTR:** Los niveles nulos de triglicéridos se reemplazan por la mediana en función del género y el grupo de edad.
- **LBXIN:** Los niveles de insulina nulos se reemplazan con la mediana según género y grupo de edad.
- **LBXGLU:** Los niveles de glucosa nulos se reemplazan con la mediana según género y grupo de edad.
- **LBHDD:** Corresponde a los niveles de colesterol HDL. Se reemplazan los valores nulos con la mediana según género y grupo de edad.

3.3.1.5 LIMPIEZA DE VARIABLES NULAS DEL CUESTIONARIO

En este caso, según la respuesta del cuestionario podemos establecer tres categorías. Una de respuesta sí (1) o no (0), otra por frecuencia temporal (nunca (0), algunos días (1), más que la mitad de los días (2), casi todos los días (3)), otra por niveles de calidad (excelente (1), muy buena (2), buena (3), suficiente (4) y pobre (5)) y finalmente las que son directamente rangos numéricos.

CASO 1: RESPUESTAS DE SÍ O NO

- **BPQ040A**: Indica si el individuo tiene prescripción de medicación para la hipertensión. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **BPQ030:** Indica si al individuo le ha dicho el médico dos veces o más que tiene la presión sanguínea. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **BPQ050A**: Indica si el individuo toma medicación actualmente para la hipertensión. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **BPQ090D**: Indica si el individuo toma medicación para el colesterol. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **DIQ010:** Indica si el doctor ha informado al paciente de tener diabetes. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), se codifica la opción "*borderline*" (es decir, que se sitúa en el límite) como 1 y la opción de sí como 2.
- **DIQ070:** Indica si el individuo toma medicación para bajar los niveles de azúcar en sangre. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **HIQ011:** Indica si el individuo está cubierto por un seguro sanitario. No existen valores nulos.
- **MCQ080:** Indica si el doctor ha informado al paciente de tener obesidad. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **MCQ010:** Indica si el doctor ha informado al paciente de tener asma. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **MCQ220**: Indica si el doctor ha informado al paciente de tener cáncer. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **MCQ300C:** Indica si el paciente ha estado cerca de tener diabetes. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **PAQ605:** Indica si el paciente realiza actividad física vigorosa. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), que es la moda.
- **PAQ620:** Indica si el paciente realiza actividad física moderada. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), que es la moda.
- **PAQ635:** Indica si el suele caminar o montar en bicicleta. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), que es la moda.
- **PAQ650:** Indica si el paciente realiza actividad física vigorosa en sus planes de ocio. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), que es la moda.
- **PAQ665:** Indica si el paciente realiza actividad física moderada en sus planes de ocio. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0), que es la moda.
- **SMQ020:** Indica si el paciente ha fumado al menos 100 cigarrillos en su vida. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).
- **INQ0200:** Indica si el paciente obtiene un dinero de su sueldo. Se rellenan los valores nulos, que no saben o que se niegan a responder con sí (1).
- **SLQ050:** Indica si el paciente ha comunicado al doctor que tiene problemas para dormir. Se rellenan los valores nulos, que no saben o que se niegan a responder con no (0).

CASO 2: RESPUESTAS DE FRECUENCIA TEMPORAL

- **DPQ020:** Indica con cuanta frecuencia el paciente se siente triste, deprimido o desesperanzado. Se rellenan los valores nulos, que no saben o se niegan a responder con nunca (0).
- **DPQ030:** Indica con cuanta frecuencia el paciente tiene problemas para dormir o duerme demasiado. Se rellenan los valores nulos, que no saben o se niegan a responder con nunca (0).
- **DPQ040:** Indica con cuanta frecuencia el paciente se siente cansado o con poca energía. Se rellenan los valores nulos, que no saben o se niegan a responder con nunca (0).
- **DPQ050:** Indica con cuanta frecuencia el paciente se siente con poco apetito o que come en exceso. Se rellenan los valores nulos, que no saben o se niegan a responder con nunca (0).

CASO 3: RESPUESTAS DE NIVELES DE CALIDAD

- **HSD010:** Indica la condición de salud general del paciente. Se rellenan los valores nulos, que no saben o se niegan a responder con buena (3).
- **DBQ700:** Indica como de saludable es la dieta del paciente. Se rellenan los valores nulos, que no saben o se niegan a responder con buena (3).

CASO 4: RANGO DE VALORES

- **INDFMMPI**: Indica el índice de pobreza familiar (mensualmente). Es el ratio de ingresos familiares entre las directrices de pobreza. Se calcula según el estado, el año y el número de individuos presentes en la unidad familiar.

- **INDFMMPC:** Indica la categoría de pobreza familiar según el índice de pobreza familiar descrito en el punto anterior. Los valores nulos, que no saben o se niegan a contestar se codifican con:
  I. **1:** Índice de pobreza mensual <= 1.30
  II. **2:** 1.3 **<** Índice de pobreza mensual <= 1.85
  III. **3:** Índice de pobreza mensual > 1.85

- **ALQ130:** Indica la media de bebidas alcohólicas que ha consumido un paciente en el último año. En este caso se establecen grupos:
  I. **0**: Si el paciente no ha consumido bebidas alcohólicas
  II. **1**: Si el paciente ha consumido de 1 a 3
  III. **2**: Si el paciente ha consumido de 4 a 6
  IV. **3**: Si el paciente ha consumido de 7 a 10
  V. **4**: Si el paciente ha consumido de 11 a 14
  VI. **5**: Si el paciente ha consumido 15 bebidas alcohólicas o más

3.3.1.6 DETECCIÓN DE VALORES ATÍPICOS

Los modelos son sensibles a la presencia de valores atípicos. Los valores atípicos son valores que se salen notablemente de la norma y que suelen representar errores en la recolección de datos o mal procesado. Cuando los datos no cumplen con estos supuestos disminuye la capacidad de detectar efectos reales [96].

Para llevar a cabo este apartado se ha realizado un estudio de los marcadores físicos de los participantes como son la insulina, la glucosa en ayunas y la presión sanguínea, así como sus medidas corporales de peso, altura y perímetro abdominal. Este tipo de medidas pueden ser susceptibles a tener errores, ya sea por una mala toma de la muestra o por algún fallo de los pacientes a la hora de recoger las muestras (como puede ser no presentarse en ayunas para el análisis de sangre).

Mediante la visualización de diagramas de caja o *boxplots* (en su nombre original en inglés) [89] se puede decidir que valores entran dentro de los rangos normales y cuales merece la pena descartar [94].

3.3.1.7 OBTENCIÓN DE LAS COLUMNAS DE MET_SYM Y DE WHI

Una vez se tienen todos los campos limpios, se procede a calcular el índice cintura altura dividiendo directamente los valores de la cintura de los participantes entre su altura, ambas variables en centímetros. Adicionalmente, según los criterios establecidos en Tabla *3* se codifica como 1 la nueva columna MET_SYM para indicar que un participante tiene Síndrome Metabólico y con 0 para indicar lo contrario. De esta forma el conjunto de datos acaba sumando 45 columnas.

3.3.2 DESARROLLO

En este apartado se detalla el desarrollo del proyecto tras la limpieza y el preprocesado de los datos.

En primer lugar, se realiza un análisis de componentes principales con el objetivo de simplificar el entrenamiento de los modelos de aprendizaje automático, después se plantea el modo de llevar a cabo la división entre los datos de entrenamiento y los datos de prueba para finalmente dar paso a la discusión de la problemática de los datos desbalanceados.

Una vez se tengan estas cuestiones resueltas se puede proceder con el análisis estadístico del conjunto de datos y el desarrollo de los modelos con su consecuente ajuste de hiperparámetros.

3.3.2.1 ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales o PCA (por sus siglas en inglés) es un método de reducción de dimensionalidad que permite simplificar la complejidad de espacios con múltiples dimensiones conservando su información [97]. Este método hace posible condensar la información en menos variables para entrenar modelos supervisados de aprendizaje automático, de esta forma se acelera el tiempo de entrenamiento y test del algoritmo.

En primer lugar, obtenemos el conjunto de datos que va a servir para entrenar al modelo, eliminando las columnas que sirven para determinar la presencia de SM en un individuo y la de MET_SYM (que indica si un individuo presenta Síndrome Metabólico). Consecuentemente se obtiene un dataset de 33 columnas y 26.865 filas.

Para eliminar sesgos en las mediciones lo primero es centrar las observaciones. Para lo cual se utiliza la función *StandarScaler* de Scikit-learn que normaliza y escala los datos [98]

No existe un consenso único para identificar el número óptimo de componentes principales. En este caso se evalúa la proporción de varianza explicada acumulada (se refiere a la medida para ver cuanta información se conserva en el cálculo de componentes principales) para seleccionar el número mínimo de componentes a partir del cual el incremento deja de ser sustancial. En el siguiente gráfico, en la Figura 12, se observa que con 26 variables se logra explicar el 99% de la varianza. Por lo tanto, por medio de la clase PCA de la librería Scikit-learn se fijan las 26 componentes [99]

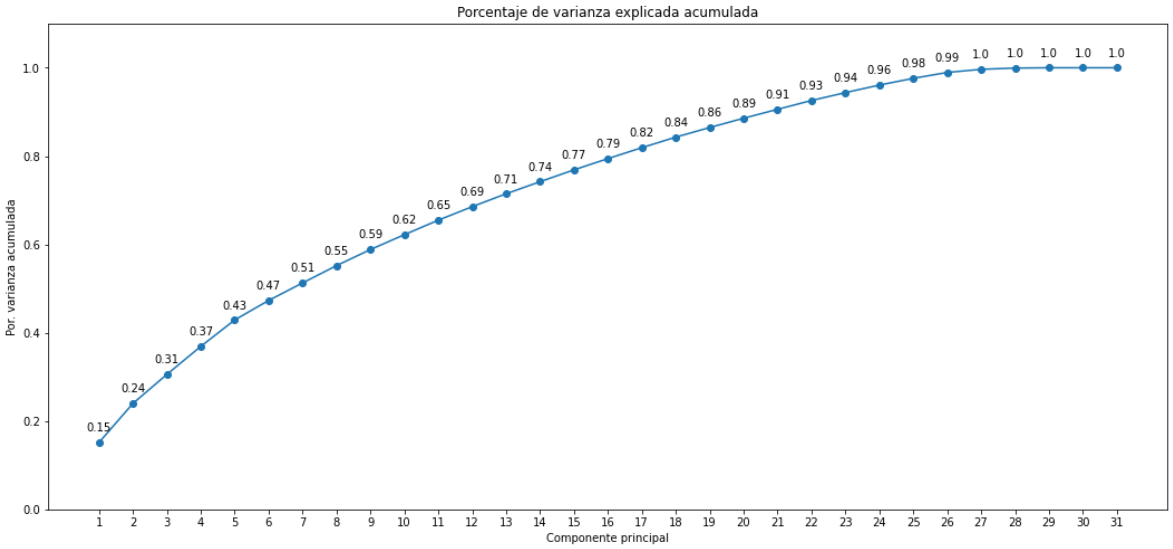

Figura 12: Porcentaje de varianza explicada acumulada para selección de números de PCAs

3.3.2.2 CONJUNTO DE ENTRENAMIENTO Y TEST

Una vez obtenida la división por componentes principales, se procede a dividir el conjunto de datos en una agrupación de entrenamiento y otra de test. Para ello se utiliza la función *train_test_split* de la librería Scikit-learn [100]. La división se realiza de forma que el conjunto de entrenamiento contiene el 80% de los datos mientras que el conjunto de test el 20%, tras diversas pruebas con distintas divisiones del conjunto se reporta una ligera mejora al elegir estos porcentajes.

Esta partición del conjunto de datos permite conocer si el modelo es capaz de generalizar con datos diferentes al de su entrenamiento y obtener sus métricas de rendimiento.

3.3.2.3 PROBLEMA DE CLASIFICACIÓN DE CONJUNTOS DE DATOS DESBALANCEADOS

El problema de clasificación de conjuntos de datos desbalanceados *(imbalanced classification problems*, en su nombre original en inglés) puede resultar conflictivo a la hora de utilizar modelos predictivos de clasificación.

Cualquier conjunto de datos con una distribución de clases desigual está, técnicamente, desequilibrado. Sin embargo, para que se pueda calificar de un desequilibrio grave debe darse una desproporción significativa, del orden de 1:1000 [101].

En el conjunto de datos que se va a utilizar para entrenar los modelos se obtienen 26.865 filas, los pacientes que presentan Síndrome Metabólico son 11.142 frente a 15.723 que no lo tienen, esto se puede calificar de un desbalanceo leve que no precisa de técnicas de submuestreo o sobremuestreo. Los clasificadores no funcionan bien frente a conjuntos de datos desbalanceados por diversas causas [102]:

- Existencia de subclases poco representadas que pueden confundirse con valores atípicos.

- Falta de densidad en los datos de entrenamiento.
- Solape entre clases en las zonas fronterizas.
- Confusión con ruido o dificultad para poder discriminarlos frente a los sobrerrepresentados.
- Instancias de entrenamiento y test con una distribución de probabilidad diferente.

En conclusión, este ligero desbalanceo no se considera problemático para la aplicación expuesta en este trabajo.

3.3.2.4 ANÁLISIS ESTADÍSTICO DE LOS DATOS

En este apartado se realiza un análisis estadístico de los datos recopilados por los participantes que forman parte de los ciclos seleccionados de la base de datos con el objetivo de tener una visión de las características generales del grupo muestral de la encuesta.

Para ello se han estudiado la distribución de los participantes por datos demográficos. Es decir, la división por género, edad y etnia. Por otro lado, se ha llevado a cabo un estudio de la prevalencia del Síndrome Metabólico en función a estos factores.

Finalmente, se estudia el porcentaje de sobrepeso y obesidad que existe entre los individuos de la muestra mediante la visualización de diagramas de caja de sus medidas corporales.

3.3.2.5 MATRICES DE CORRELACIÓN

En este apartado se obtienen las matrices de correlación de los atributos seleccionados de la base de datos para comprender su influencia sobre padecer Síndrome Metabólico. Las matrices de correlación muestran los valores de correlación de Pearson, que miden el grado de relación lineal entre cada par de elementos o variables. Estos valores se ubican entre -1 y +1 en función de si la relación es directa o inversamente proporcional [103].

Se realiza una diferenciación por el tipo de datos que consiste en:

- Correlación por etnia.
- Correlación por variables demográficas: nivel de educación, nivel de pobreza, edad y estrato social.
- Correlación por la frecuencia y la intensidad del ejercicio físico que realizan los participantes (o la ausencia de este).
- Correlación por la calidad de la dieta de los participantes y su valoración del estado general de salud.
- Correlación por problemas en el sueño de los participantes, por exceso o por defecto, y de su estado anímico.
- Correlación por la frecuencia en la toma de alcohol y tabaco

Finalmente, los atributos en los que no se observa una correlación suficiente con el Síndrome Metabólico se descartan.

3.3.2.6 DESARROLLO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO

Con el análisis previo, se procede al desarrollo de los modelos de aprendizaje automático y a su ajuste de hiperparámetros con el objetivo de optimizar su rendimiento.

3.3.2.6.1 DESARROLLO DEL MODELO DECISION TREE

Para este apartado se utiliza *DecisionTreeClassifier* de la biblioteca Scikit-learn [104]. En primera instancia se entrena el modelo con parámetros por defecto para tener una ligera idea de su rendimiento. Después se optimiza la elección de hiperparámetros (parámetros que deben seleccionarse manualmente) con la función *GridSearchCV* de Scikit-learn [105]. Esta función permite recorrer todas las combinaciones de parámetros posibles definidos dentro de un rango a través de bucles, el parámetro que aporte mejor rendimiento es el seleccionado. Su principal desventaja radica en que a nivel computacional es costoso, especialmente cuando se trata de grandes conjuntos de datos y múltiples parámetros.

Los hiperparámetros que se ajustan en esta ocasión son:

1. Criterion: La función para medir la calidad de una división.

    i) Gini (valor por defecto): la impureza de gini mide la frecuencia con la que cualquier elemento del conjunto de datos será mal etiquetado cuando se etiquete al azar, solo es válida para problemas de clasificación binaria. [106]

    ii) Entropy: la división óptima es elegida por el atributo con menos entropía. Obtiene su valor máximo cuando la probabilidad de las dos clases es la misma. [107]

2. Splitter: La estrategia utilizada para elegir la división de cada nodo.
    i) Best (valor por defecto): elige la mejor división.
    ii) Random: elige una división aleatoria.

3. Max_depth: Indica la máxima profundidad del árbol. Es un entero. None es el parámetro por defecto y significa que el árbol se ramifica hasta que todas las hojas son puras.
4. Min_samples_split: Mínimo número de muestras necesarias para dividir un nodo interno. Es un entero y por defecto está ajustado a 2.
5. Min_samples_leaf: Mínimo número de muestras dentro de un nodo hoja.
6. Random_state: Indica el grado de permutación de característica en cada división y en cada ejecución puede variar. Es un entero.
7. Max_leaf_nodes: Máximo número de nodos que puede tener el árbol.
8. Class_weight: Peso asociado a cada atributo con el que se entrena al modelo. De no especificarse se asigna uno a todos.
9. Ccp_alpha: Parámetro de complejidad para la poda del árbol con la finalidad de evitar el sobreajuste.

Las métricas que evalúan los mejores parámetros se obtienen mediante K-Fold Cross-Validation [108]. Se trata de una técnica que previene el sobreajuste (*overfitting,* por su término original en inglés). Se basa en un proceso iterativo que consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño (Figura 13), k-1 grupos se destinan al entrenamiento y uno para validación. El proceso se repite k veces utilizando diferentes agrupaciones en cada iteración. Se obtienen k estimaciones del error cuyo promedio es el resultado final. Su principal ventaja es que consigue una estimación precisa del error de test gracias a un mejor balance entre sesgo y varianza [109]

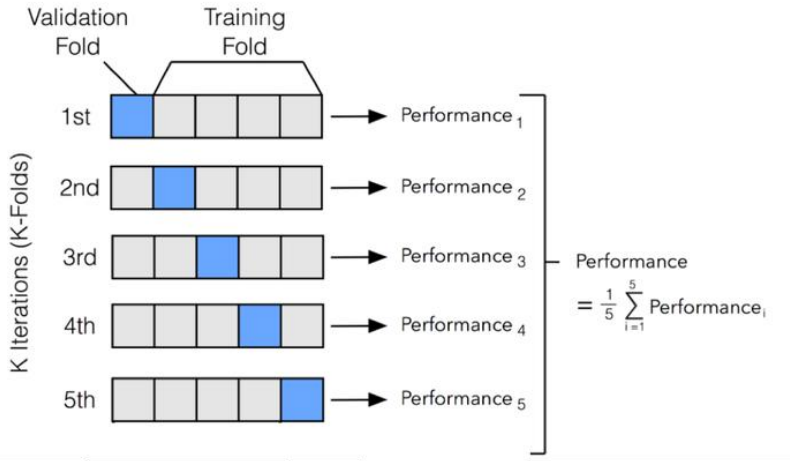

Figura 13: K-Fold Cross-Validation

Con el objetivo de prevenir el *overfitting* (efecto por el que el algoritmo se especializa en predecir los datos con los que ha sido entrenado, pero falla a la hora de generalizar) en árboles de decisión se usa la poda del árbol (*pruning* en su nombre original en inglés) o la aplicación del método de ensamblado Random Forest, que se desarrolla en el siguiente apartado. Nos encontramos con dos tipos de *pruning* [110].

- **Pre-pruning**: consiste en limitar el crecimiento del árbol de decisión. Para ello se deben ajustar los hiperparámetros de max_depth, min_samples_leaf y min_samples_split. Lo cual ya se realiza con el ajuste de hiperparámetros por *GridSearchCV* mencionado anteriormente.
- **Post-pruning**: este método consiste en dejar al árbol crecer hasta su profundidad máxima para luego borrar las ramas que sobran. Para ello se ajusta el parámetro ccp_alpha (cost complexity pruning). Scikit-learn implementa la función cost_complexity_pruning_path[13], Gr para este cometido. Mientras más valor tenga el ccp_alpha, más nodos se podan.
- **Random Forest**: combinación de árboles de decisión tal que cada árbol depende de los valores de un Vector aleatorio probado independientemente. Evita el sobreajuste con técnicas de agregación y bootstrap sampling (método que implica la extracción de datos de muestra repetidamente con reemplazo de una fuente de datos para estimar un parámetro de población). Se implementa en el siguiente punto.

A continuación, se muestra la selección final de hiperparámetros en la Tabla 6.

Tabla 6: Hiperarámetros del Decision Tree

Parámetro	Valor
Max_depth	7
Max_leaf_nodes	None
Splitter	Best
Class_weigth	None
Criterion	Gini
Random_state	1
Min_samples_leaf	5
Min_samples_split	10

3.3.2.6.2 DESARROLLO DEL MODELO RANDOM FOREST

Como se ha mencionado anteriormente el Random Forest se forma de varios árboles de decisión funcionando con diferentes submuestras del conjunto de datos. Se utiliza RandomForestClassifier de Scikit-learn [111]. Al igual que con el árbol de decisión, se ajustan los parámetros utilizando *GridSearchCV* y se obtienen las métricas mediante K-Fold Cross-Validation. N_estimators hace referencia al número de árboles de decisión que se han fijado para el modelo. En la Tabla 7 se muestra la selección final de hiperparámetros.

Tabla 7: Hiperparámetros del Random Forest

Parámetro	Valor
N_estimators	100
Max_depth	9
Criterion	Gini
Random_state	1
Min_samples_leaf	2
Min_samples_split	6

3.3.2.6.3 DESARROLLO DEL MODELO SUPPORT VECTOR MACHINE

En este punto se utiliza el SVC (Support Vector Classifier), función de Scikit-learn. Los hiperparámetros que se van a seleccionar, nuevamente mediante *GridSearchCV*, son los siguientes:

1. C: parámetro de regulación. Debe ser un número positivo. A mayor valor más control de los valores atípicos
2. Kernel: hace referencia al tipo de kernel.
   i) Linear: se emplea cuando los datos pueden ser separados con una simple línea recta.
   ii) Poly: la separación entre los datos se realiza mediante una función polinómica.
   iii) Rbf (valor por defecto): Para añadir más características a los datos se utilizan las características de similitud, lo cual mide la distancia entre un valor de una característica existente y un punto de referencia. El Rbf nos permite obtener los mismos resultados que si añadiéramos un punto de referencia en cada valor del atributo original, sin hacerlo realmente.
   iv) Sigmoid: Las curvas de aprendizaje en sistemas complejos muestran una progresión temporal, desde un nivel bajo al inicio hasta acercarse a un clímax. La función sigmoide describe esta evolución.
   v) Precomputed: kernel personalizado.
3. Gamma: coeficiente del kernel, solo aplicable a rbf, sigmoid y poly.
   i) Scale (valor por defecto):

$$Scale = \frac{1}{n_{features} * X.var}$$

Ecuación 2: Scale

   ii) Auto:

$$Auto = \frac{1}{n_{features}}$$

Ecuación 3: Auto

Las Máquinas de Vector de Soporte son bastante robustas frente al *overfitting*, debido a que incluye un parámetro C, el cual debe ser seleccionado cuidadosamente, que controla la compensación entre errores de entrenamiento y los márgenes rígidos [112]. De esta forma, permite algunos errores en la clasificación a la vez que los penaliza. Incluso en las ocasiones en las que el número de atributos es mayor que el número de observaciones. A continuación, en la Tabla 8, se muestra la elección de hiperparámetros.

Tabla 8: Hiperparámetros SVC

Parámetro	Valor
Kernel	rbf
Gamma	Scale
C	0.3

3.3.2.6.4 DESARROLLO DEL MODELO XGBOOST

Extreme Gradient Boosting es un algoritmo de aprendizaje supervisado que intenta predecir con precisión una variable objetivo combinando las estimaciones de un conjunto de modelos más simples y débiles. Para este apartado se utiliza *GradientBoostingClassifier* de Scikit-learn [113].

Mediante *GridSearchCV* se seleccionan los siguientes hiperparámetros:

1. Learning_rate: parámetro que determina la contribución de cada árbol. Se debe encontrar un equilibrio entre este valor y el número de estimadores.
2. Max_depth: máxima profundidad del modelo.
3. Min_child_weight: suma mínima de peso de instancia (arpillera) necesaria en un hijo. Se utiliza para regular y limitar la profundidad del árbol.
4. Subsample: la fracción de muestras que se utilizará para ajustar los aprendices base individuales. Si es menor que 1,0, el resultado es el refuerzo de gradiente estocástico.
5. Colsample_byTree: fracción de características (seleccionadas al azar) que se utilizará para entrenar cada árbol.
6. N_estimators: número de etapas de refuerzo.
7. Objective: Informa sobre el tipo de problema al que se enfrenta el modelo.
8. Nthreads: número de hebras concurrentes al ejecutar el modelo.
9. Speed: indicador de velocidad de ejecución.
10. Scale_pos_weight: ratio del número de ceros y unos en el atributo a predecir.

A continuación, en la Tabla 9 se muestra la selección final de los hiperparámetros.

Tabla 9: Hiperparámetros XGBoost

Parámetro	Valor
Learning_rate	0.05
Max_depth	5
Min_child_weight	3
Subsample	0.8
Colsample_byTree	0.7
N_estimators	150
Objective	binary:logistic
Nthreads	5
Speed	27
Scale_pos_weight	1

3.3.2.6.5 DESARROLLO DEL MODELO NEURONAL NETWORK

Para crear la red neuronal se emplea la clase *Sequential* de la librería Keras de TensorFlow [114]. Este modelo permite conectar varias capas de la red neuronal en serie. En primer lugar, se define el grafo, creando las capas de la red y estableciendo el número de neuronas de cada una. Después se compila la red para transformar la secuencia en una serie con un formato compatible para ser ejecutadas en una CPU, aquí se especifican el algoritmo de optimización y la función de pérdidas que trata de minimizar el optimizador. Finalmente se ajusta la red, adaptando los pesos al conjunto de datos e indicando el número de exposiciones (*epochs*) para su entrenamiento, cada exposición se divide en grupos de pares de patrones input-output (lotes) [115].

Con objeto de prevenir el sobreajuste se define la función de *callback early stopping*, *Early stopping* es una técnica regularización que detiene el entrenamiento cuando las actualizaciones de los parámetros ya no empiezan a producir mejoras significativas en el conjunto de validación.

Otra técnica que se utiliza para prevenir el sobreajuste es el *Dropout*, consiste en un método que desactiva de forma aleatoria un número de neuronas de la red neuronal de forma aleatoria [116].

Se realizan cuatro etapas de *GridSearchCV* para seleccionar los hiperparámetros, a continuación, se detalla cada uno.

1. Batch size: define el número de muestras que se propagan por la red neuronal en cada iteración
2. *Epochs*: hace referencia al número de iteraciones de la red
3. Layers: estructura de la topología del modelo.
4. Dropout: parámetro que se utiliza para prevenir el *overfitting*. Se seleccionan aleatoriamente neuronas que se ignoran durante el entrenamiento. Su contribución a la red se elimina temporalmente. El dropout solo se utiliza en el entrenamiento, no en la evaluación.
5. Optimizer: función que modifica los atributos de la red neuronal, como los pesos o la tasa de aprendizaje con el objetivo de reducir la pérdida global y mejorar la precisión.

Finalmente, se obtienen los siguientes resultados en la Tabla 10 y ajuste de parámetros:

Tabla 10: Hiperparámetros Neural Network

Parámetro	Valor
Optimizer	Adam
Batch size	32
Epochs	80
Lyrs	[8]
Dr	0.08

3.3.5 VALIDACIÓN DE LOS RESULTADOS

Para la validación desde el punto de vista de usuario, se realiza un sondeo realizado a profesionales de la salud sobre la utilidad, limitaciones y líneas futuras del empleo de técnicas de inteligencia artificial avanzadas aplicadas en el diagnóstico precoz de factores de riesgo de Síndrome Metabólico (SM). Se valora también la aplicación de estas técnicas en diferentes problemas de salud en función de la especialidad del encuestado.

En esta encuesta se realizan tres preguntas sobre el perfil del encuestado:

- Rango de edad
- Género
- Especialidad sanitaria

A continuación, se propone una escala del 1 al 10 para valorar las siguientes cinco cuestiones:

- ¿Cómo valorarías la utilidad en un entorno médico el empleo de estas técnicas?
- ¿Te parece que el empleo de estas técnicas podría ahorrar recursos y tiempo al personal sanitario y a los centros de salud?
- ¿Cuánta mejora piensa que se podría obtener a la hora de prevenir y tratar enfermedades si el empleo de estas técnicas se aplicara en los centros médicos?
- ¿Qué probabilidades habría de que quisiera utilizar esta herramienta en su entorno de trabajo (si se pusiera a su disposición) como ayuda para tomar Decisiones sobre el diagnóstico y tratamiento de un paciente?
- ¿Cuánta desconfianza le producen los modelos de Inteligencia artificial para el diagnóstico del Síndrome Metabólico?

Finalmente, se plantean dos preguntas opcionales de respuesta libre.

- ¿Qué barreras identificas a la hora de adoptar este tipo de soluciones?
- ¿Mejorarías algo de esta herramienta?

4. RESULTADOS

En este apartado se exponen los resultados obtenidos en cada etapa del desarrollo del TFM.

4.1 RESULTADOS DE LA LIMPIEZA Y PREPROCESADO DE LOS DATOS

En cuanto a la detección de valores atípicos, para las variables continuas de laboratorio y mediciones físicas: LBXIN (insulina), LBXGLU (glucosa), LBTR (triglicéridos), LBDHDD (colesterol HDL), BPXSY1 (presión sistólica), BPXDI1 (presión diastólica) y BXBMI (IMC) (que además guardan relación directa con los requisitos del SM) se realiza un estudio para ver en qué casos merece la pena eliminar los valores atípicos. Esto se debe a que hay que valorar en qué situación un valor atípico está relacionado con un estado patológico y cuando es una medida errónea. Dicho error puede suceder o bien porque se ha procesado mal el dato, o bien porque el paciente no ha realizado el test en las condiciones óptimas.

Las pruebas de laboratorio deben realizarse en ayunas, por lo tanto, la presencia de valores atípicos en niveles de glucosa, insulina, triglicéridos y colesterol HDL puede deberse a que los sujetos no hayan respetado esta recomendación. Se eliminan los valores atípicos por encima del cuantil 0.99 y por debajo del 0.1 [117].

A continuación, se muestra la reducción de dispersión de los valores de insulina y glucosa tras haber imputado los valores atípicos. Antes del borrado de valores atípicos de insulina en la Figura 14 y después del borrado en la Figura 15.

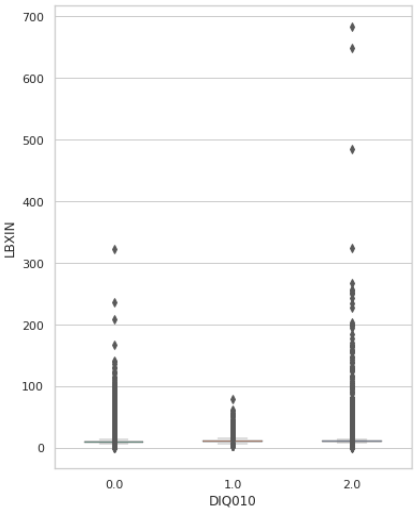

Figura 14: Boxplot de LBXIN

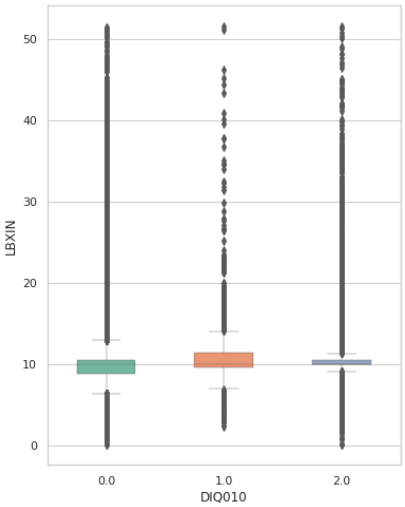

Figura 15: Boxplot de LBXIN tras borrado de valores atípicos

Como referencia se ha usado la variable DIQ010, que indica si el doctor ha hecho saber al paciente que padece diabetes. El 0 es que no, el 1 indica que se encuentra en el límite y el 2 es que sí. En los valores de insulina se aprecia mayor dispersión en 0 y en 2 mientras que en la glucosa se aprecia un nivel de dispersión similar. Por este hecho se puede suponer que algunos pacientes pudieron no ir en ayunas a la prueba o que el dato se procesó mal. Se muestra la dispersión antes del borrado de valores atípicos de glucosa en la Figura 16 y tras el borrado en la Figura 17.

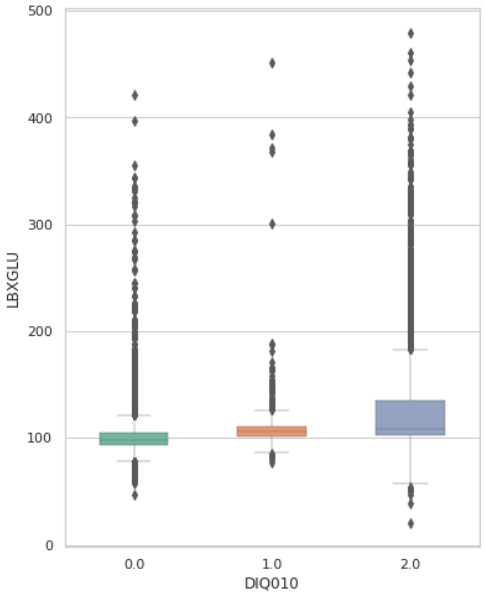

Figura 16: Boxplot de LBXGLU

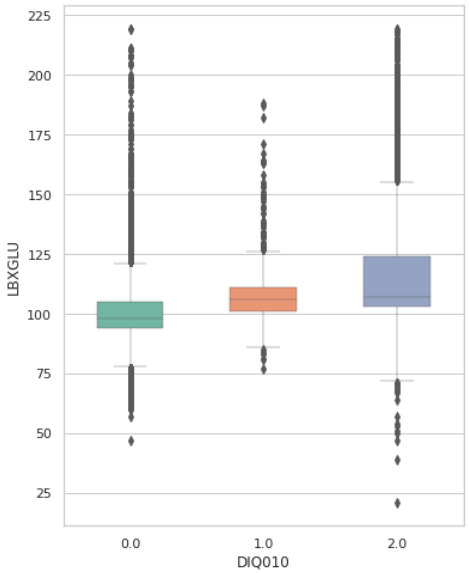

Figura 17: Boxplot de LBXGLU tras borrado de valores atípicos

Asimismo, se realiza el mismo método para los niveles de triglicéridos. Antes del borrado de triglicéridos en Figura 18 y después del borrado en la Figura 19. Son parámetros muy influidos por las hormonas, por lo que pueden diferir entre hombres (RIAGENDR = 1) y mujeres (RIAGENDR = 2).

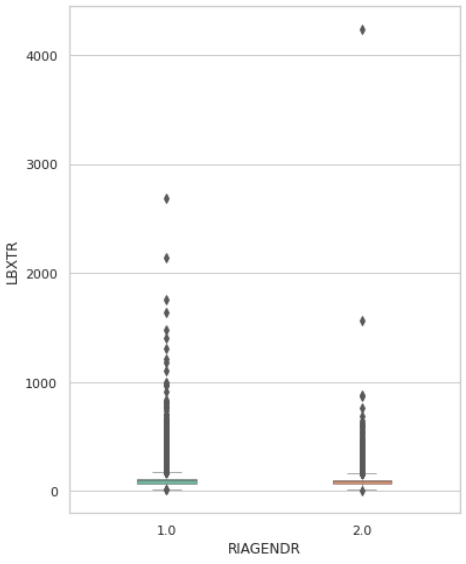

Figura 18: Boxplot de LBXTR

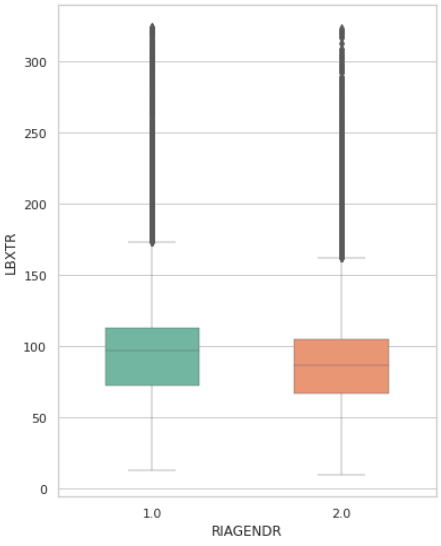

Figura 19: Boxplot de LBXTR tras borrado de valores atípicos

Para la presión sistólica y diastólica, se observan valores que se podrían considerar desmesurados en las Figura 20 y 22. Por lo tanto, se eliminan también los valores atípicos en este caso, en las Figuras 21 y 23 se aprecian los diagramas de caja tras el borrado de los valores atípicos de la presión sanguínea sistólica y diastólica. Se lleva a cabo una diferenciación por grupo de edad en las gráficas (RIDAGEYR_Tramos):

- **3**: De 12 a 20 años.
- **4**: De 20 a 35 años.
- **5**: De 35 a 50 años.
- **6**: De 50 a 65 años.
- **7**: De 65 a 80 años.

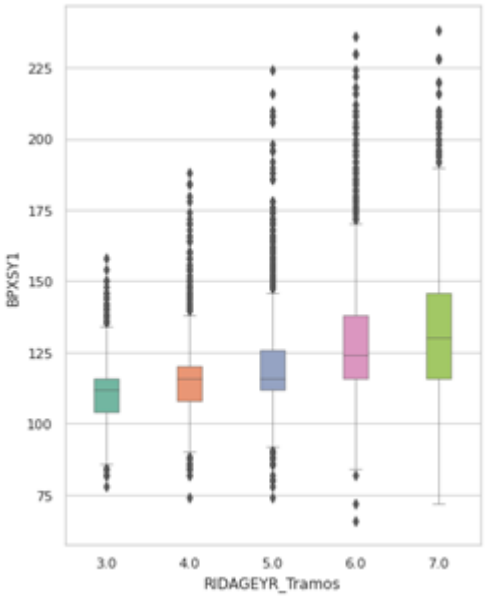

Figura 20: Boxplot de la presión sistólica

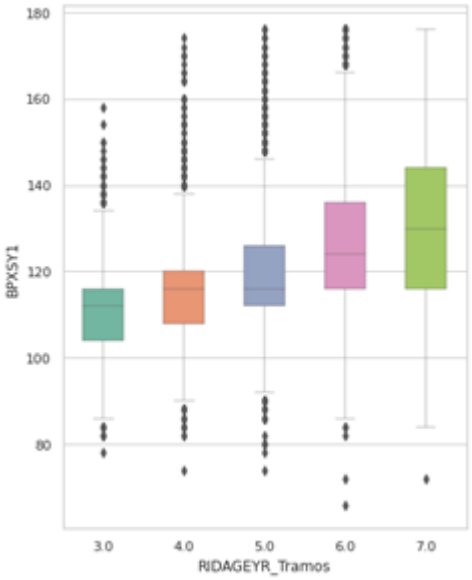

Figura 21: Boxplot de la presión sistólica tras borrado de valores atípicos

Análogamente, se aplica el mismo método para los niveles de HDL, al ser un parámetro muy influenciado por las hormonas se realiza la división por género entre hombres (RIAGENDR = 1) y mujeres (RIAGENDR = 2). Se observa en Figura 24 el diagrama de cajas antes del borrado de valores atípicos de HDL y en la Figura 25 tras el borrado.

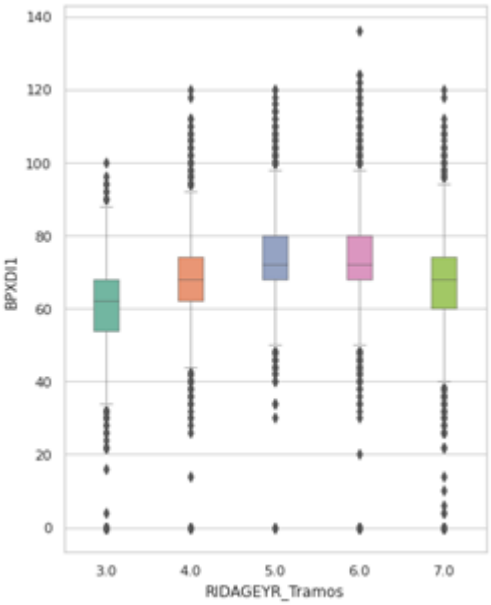

Figura 22: Boxplot de la presión diastólica

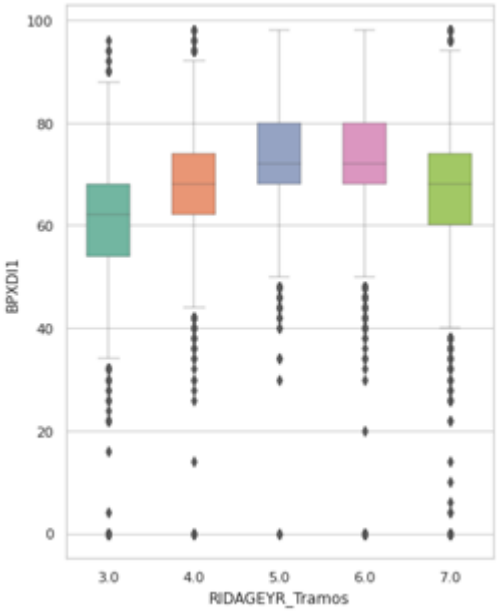

Figura 23: Boxplot de la presión diastólica tras borrado de valores atípicos

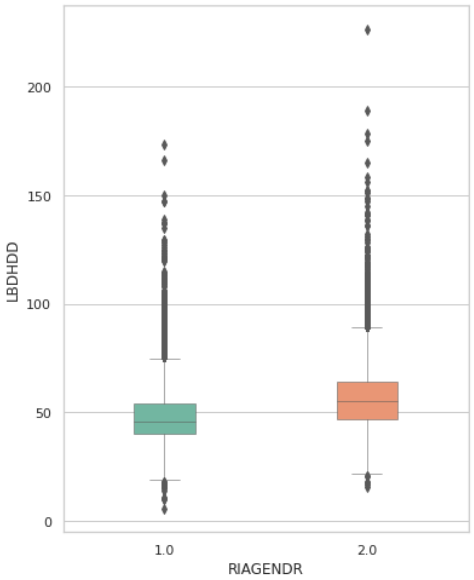

Figura 24: Boxplot del colesterol HDL

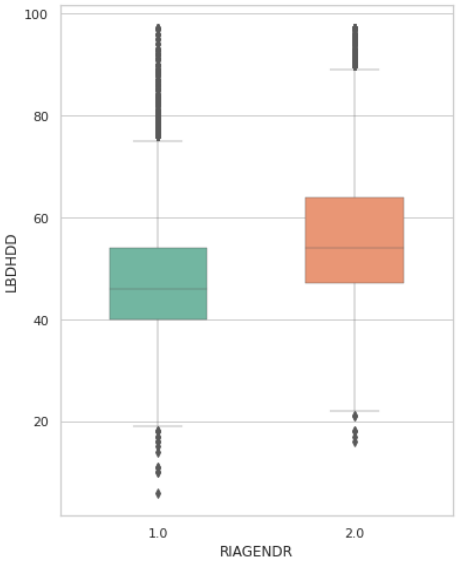

Figura 25: Boxplot del colesterol HDL tras borrado de valores atípicos

Finalmente, en cuanto a las mediciones corporales del sujeto, se puede suponer que es una medida bastante más complicada de que se tome o procese de manera errónea. Por lo tanto, se decide no eliminar los valores atípicos del perímetro abdominal, altura y peso del sujeto.

Tras la detección de valores atípicos contamos con **283** registros menos, lo cual supone una muestra casi despreciable de todo el conjunto de datos. A continuación, se presenta el diagrama de dispersión antes del borrado de valores atípicos, en la Figura 26, y después del borrado de valores atípicos, en la Figura 27 para poder llevar a cabo una comparación visual, se aprecia que la diferencia es irrisoria. Los diagramas de dispersión representan una comparación de los atributos del conjunto de datos, de forma que se pueda identificar visualmente la posible correlación entre todas las variables.

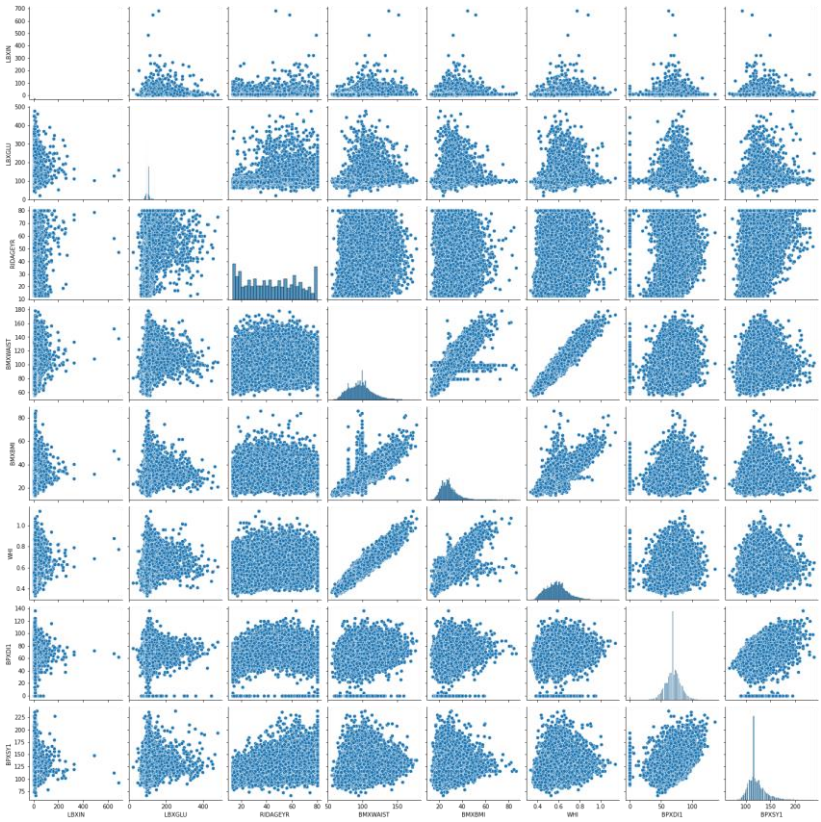

Figura 26: Diagrama de dispersión antes del borrado de valores atípicos

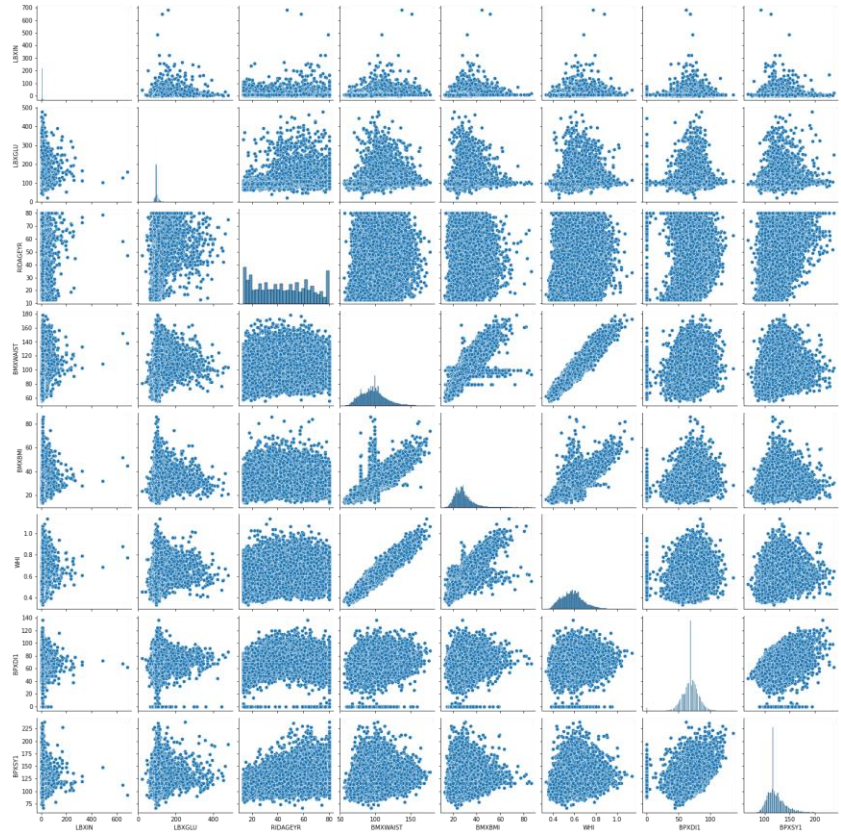

Figura 27: Diagrama de dispersión tras borrado de valores atípicos

4.2 RESULTADOS DEL ANÁLISIS ESTADÍSTICO DE LOS DATOS

A continuación, se muestra el estudio que se realiza de la muestra de participantes de los ciclos seleccionados.

4.2.1 DISTRIBUCIÓN DE PARTICIPANTES POR VALORES DEMOGRÁFICOS

En este apartado se procede a analizar el tipo de muestra de participantes por variables demográficas, así como a estudiar la influencia de diferentes variables sobre la probabilidad de sufrir o no Síndrome Metabólico.

Podemos ver que tenemos una muestra bastante homogénea en lo que a género se refiere. De todos los registros el 50.7% representa a mujeres y el 49.3% a hombres, en la Figura 28. En cuanto a la etnia, en la Figura 29, hay una mayor participación de personas blancas no hispanas (33%), seguido de no hispanos negros (24%), americanos mexicanos (16%), asiáticos (12%), otros hispanos (11%) y finalizando con un 5% de representación de otra raza. En la Figura 30 se observa la distribución por edad, que también resulta bastante homogénea.

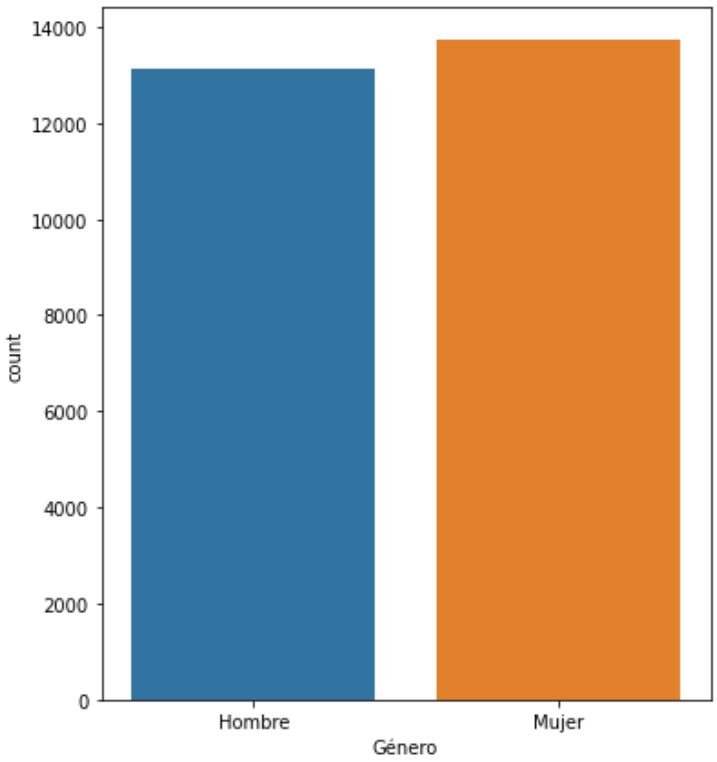

Figura 28: Distribución por género

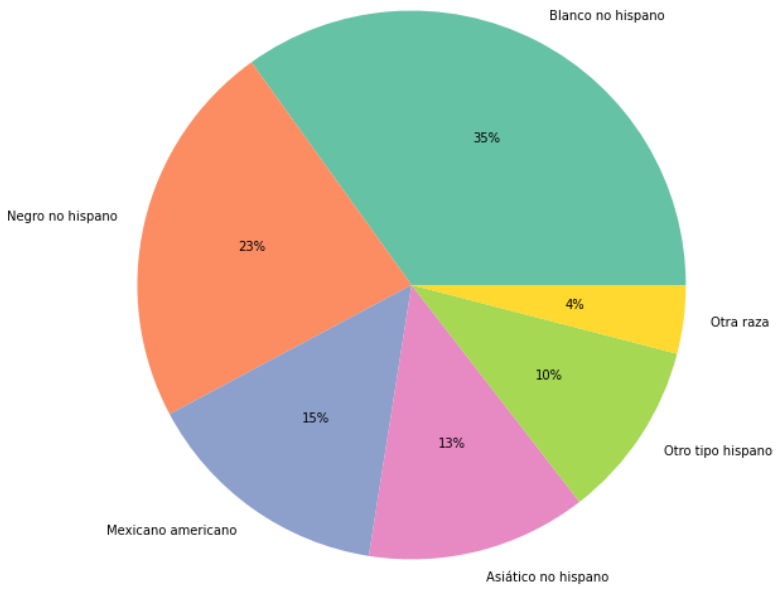

Figura 29: Distribución por etnia

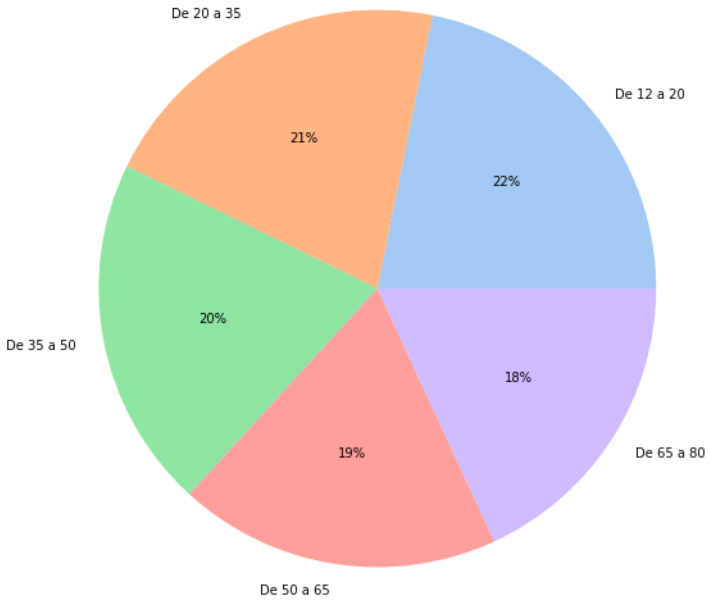

Figura 30: Distribución por grupo de edad

4.2.2 PREVALENCIA DE SÍNDROME METABÓLICO DE LOS PARTICIPANTES

De todos los participantes, el 29.9% presenta Síndrome Metabólico como se observa en la Figura 31. Siendo la barra naranja la representación de aquellos que sí lo sufren. Tanto la edad avanzada como el género masculino son factores de riesgo para padecer SM.

Hay que tener en cuenta que de 1 a 12 años no hay presencia de Síndrome Metabólico, lo cual coincide con los dos primeros grupos de edad. Este hecho, unido a que no se toman muestras de laboratorio de los menores de 12 años hace que se decida descartar los dos primeros grupos de edad (apartado 3.3.1.1.), tanto para el resto del análisis como para entrenar y probar los modelos de predicción. De esta forma, el conjunto de datos final cuenta con registros de **26.865** individuos, de los cuales el **41.47%** presenta Síndrome Metabólico. En un estudio que utiliza NHANES entre el 2001 y el 2012 [118] la prevalencia de SM en población adulta es del 35%, coincidiendo con otra investigación diferente de M. Vaduganathan et al. [119] que incorpora los años del 2011 hasta el 2016. Bastante evidencia afirma, además, que cada año este porcentaje se incrementa [120] [1]Por lo tanto, se puede considerar que los resultados son cifras razonables.

A continuación, se observa la prevalencia de SM por género en la Figura 32 y por grupo de edad en la Figura 33. Se confirma que los hombres sufren más esta patología, así como a los individuos dentro de los grupos de edad más avanzada.

Además, se muestra la prevalencia de SM en función del nivel educativo en la Figura 34. Aquellos con estudios superiores son los que más Síndrome Metabólico presentan. En cuanto a la etnia, en la Figura 35, hay mayor número de blancos no hispanos con SM, pero esto también es debido a que dentro de toda la muestra es la etnia más representada.

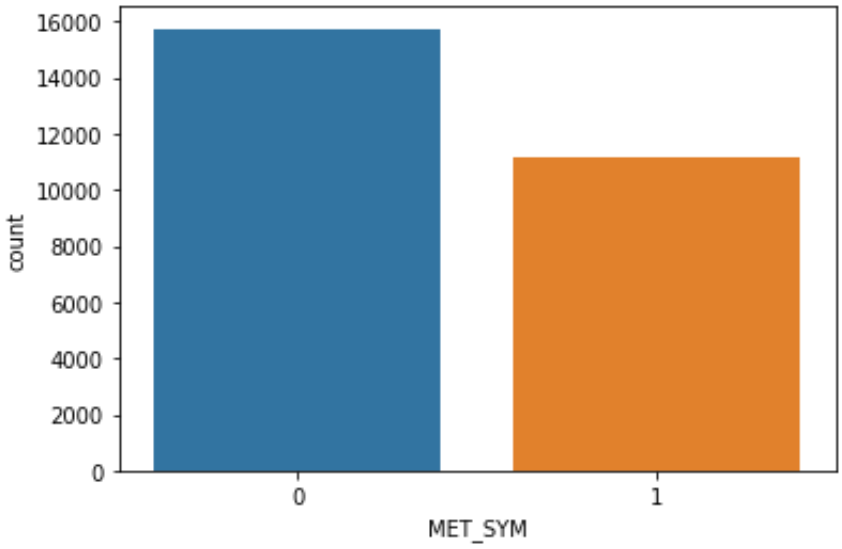

Figura 31: Prevalencia de SM de todos los participantes

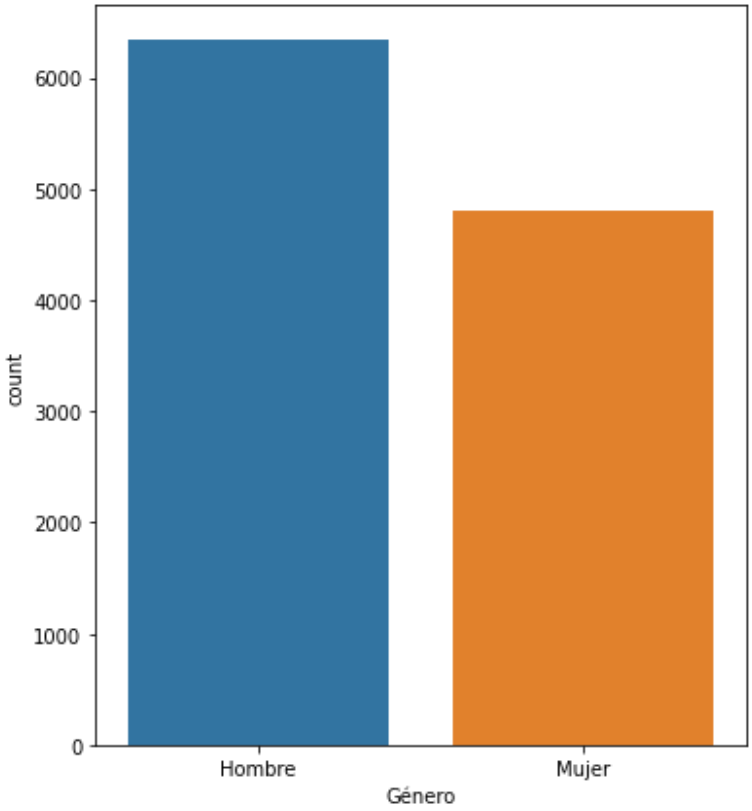

Figura 32: Prevalencia de SM por género

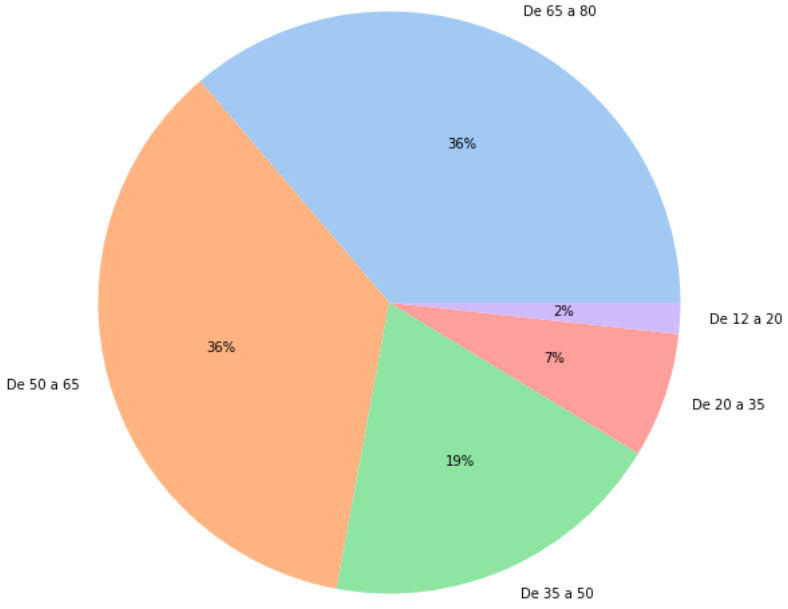

Figura 33: Prevalencia de SM por rango de edad

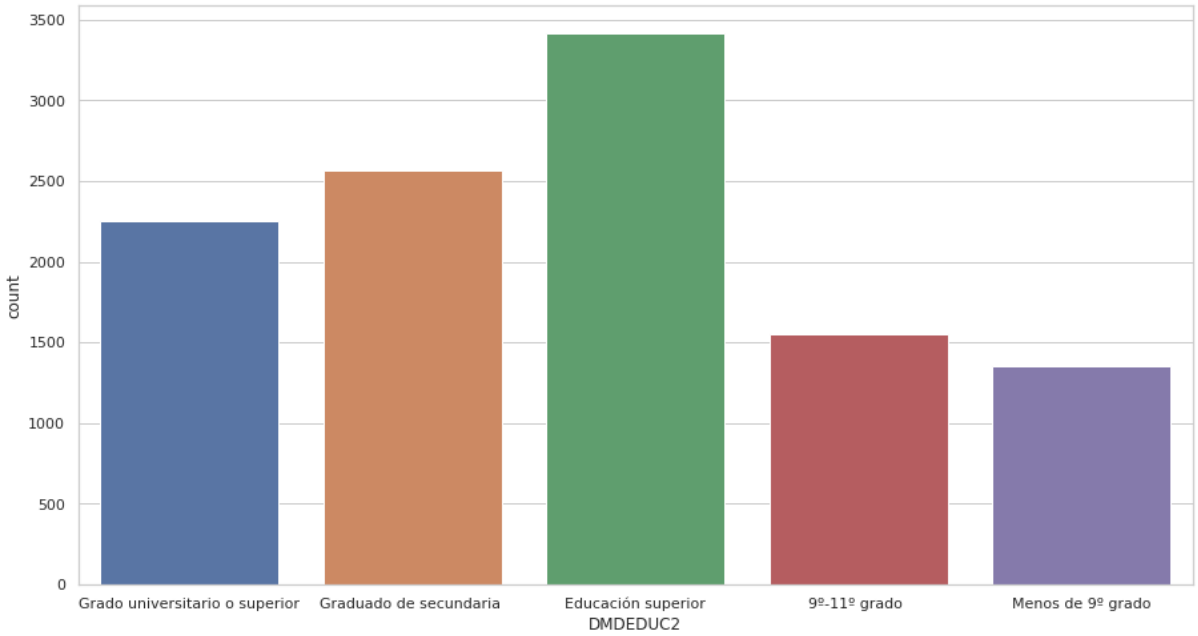

Figura 34: Prevalencia de SM por nivel educativo

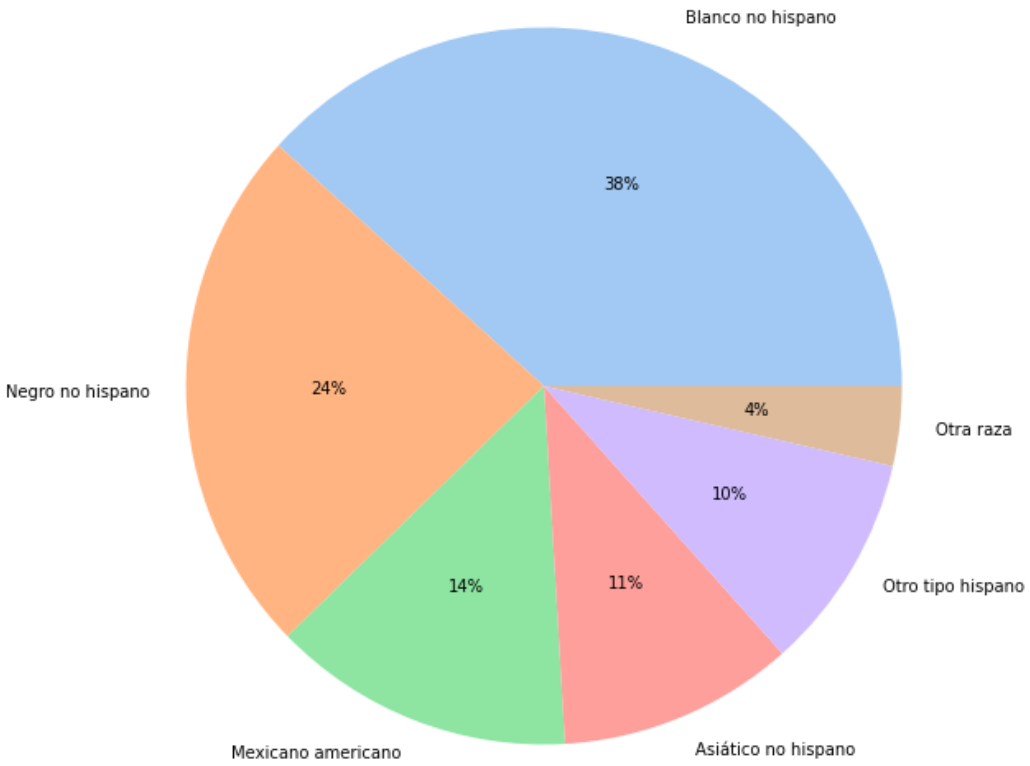

Figura 35: Prevalencia de SM por etnia

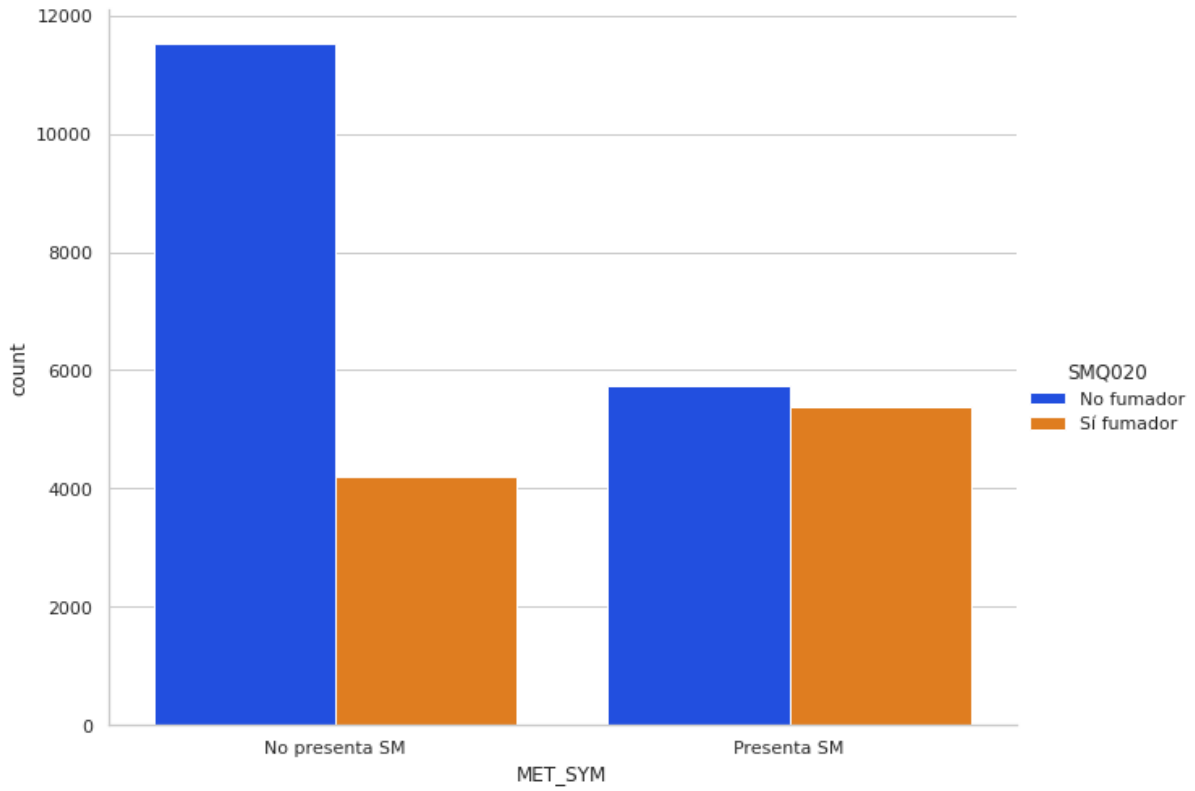

Figura 36: Prevalencia de SM por consumir tabaco

De todos los participantes con Síndrome Metabólico el 48% son fumadores, como se muestra en la Figura 36. Resulta evidente, atendiendo a la Figura 36 que ser fumador representa un factor de riesgo importante para padecer la enfermedad.

4.2.3 DISTRIBUCIÓN DE LOS PARTICIPANTES POR MEDIDAS CORPORALES

El primer requisito para padecer SM es el de tener un perímetro abdominal elevado. Entre los participantes del estudio hay un gran porcentaje con un índice cintura altura que indica sobrepeso y obesidad según la Tabla 4 en el apartado 3.1.1. A continuación, se muestran los diagramas de cajas separados por grupos de edad de los participantes sobre su cintura en centímetros en la Figura 37, su índice cintura altura en la Figura 38 y su IMC en la Figura 39. Los grupos de edad son los siguientes (atributo RIDAGEYR_Tramos):

- **3**: De 12 a 20 años.
- **4**: De 20 a 35 años.
- **5**: De 35 a 50 años.
- **6**: De 50 a 65 años.
- **7**: De 65 a 80 años.

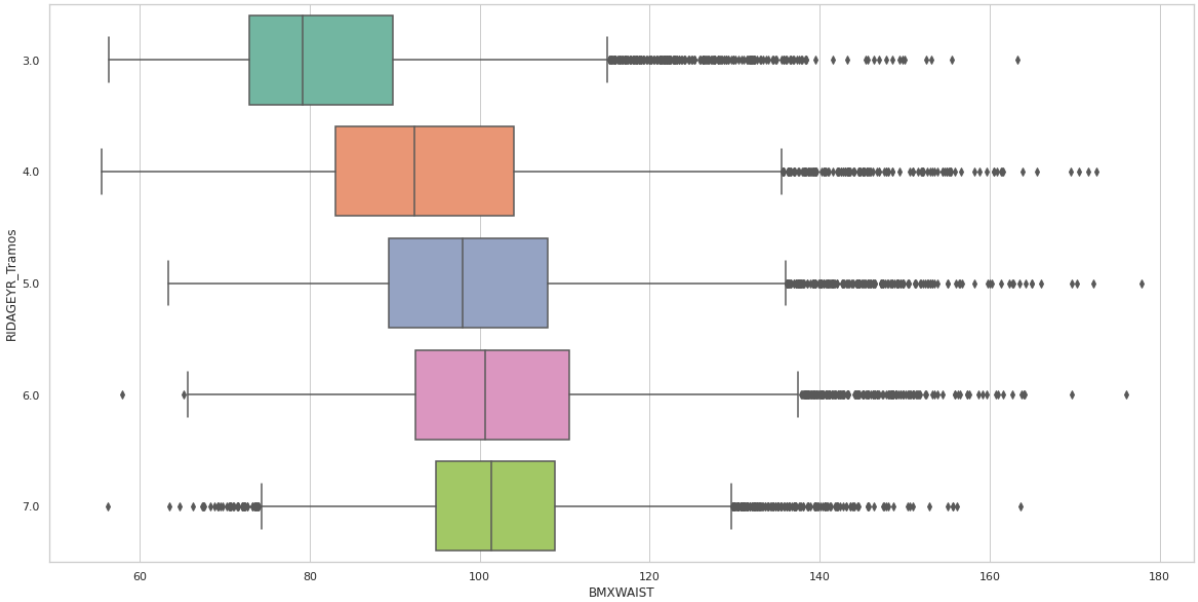

Figura 37: Cintura en cm de los participantes por grupo de edad

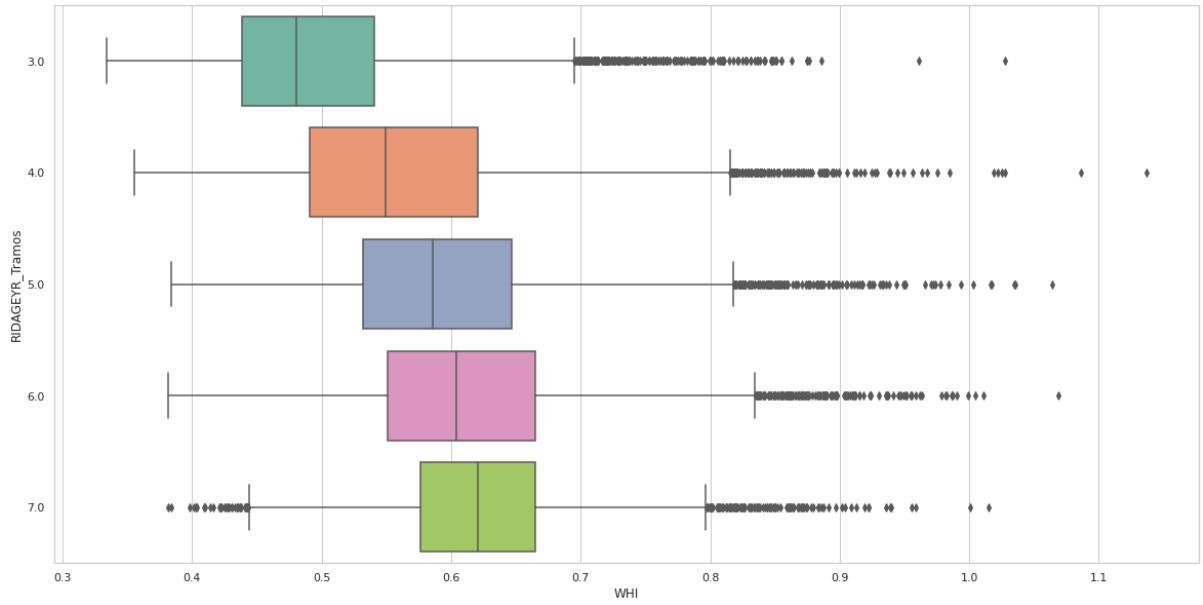

Figura 38: Índice cintura-altura por grupo de edad

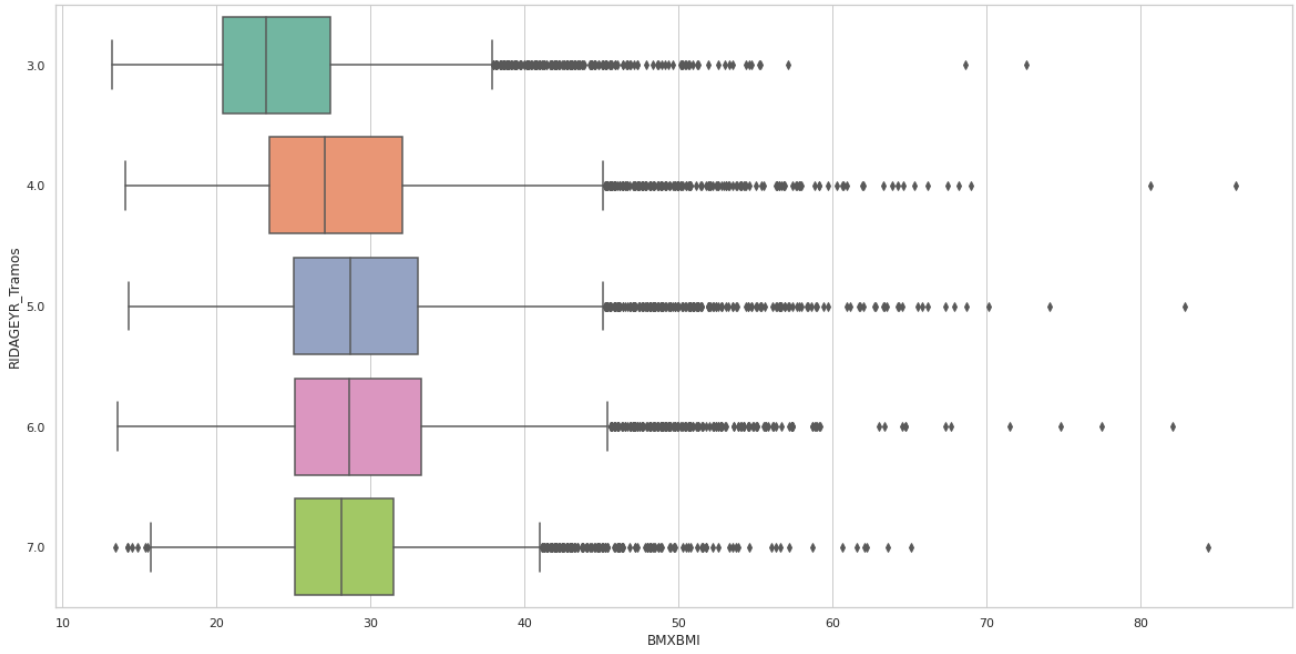

Figura 39: IMC por grupo de edad

Se observa que el IMC y el WIH se distribuye de manera similar entre los grupos de edad de los participantes.

Es interesante mencionar que en el caso de los hombres hay un reparto más homogéneo entre los que padecen sobrepeso, sobrepeso elevado y obesidad mientras que más de la mitad de las mujeres padecen obesidad, se observa en la Tabla 11. No obstante, los resultados cuadran con los estudios demográficos sobre el incremento de obesidad entre la población de EEUU, y en general, en los países del primer mundo [121], [122].

Tabla 11: Ratios de sobrepeso entre los participantes

	Sobrepeso	**Sobrepeso elevado**	**Obesidad**
Mujeres	12.33%	13.46%	54.47%
Hombres	20.74%	22.6%	25.37%

4.3 RESULTADO DE LAS MATRICES DE CORRELACIÓN

En este apartado se procede a estudiar la correlación que tienen las diferentes columnas sobre padecer o no Síndrome Metabólico. Como guía a la hora de interpretar los valores de las matrices de correlación se utiliza el siguiente criterio [123].

- Entre 0 y 0,10: correlación inexistente.
- Entre 0,10 y 0,29: correlación débil.
- Entre 0,30 y 0,50: correlación moderada.
- Entre 0,50 y 1,00: correlación fuerte.

4.3.1 CORRELACIÓN POR ETNIA

Tal y como se comentó anteriormente, diversos estudios apuntan a que parece haber mayor prevalencia de SM entre la población hispana [37]. A pesar de esto, en la matriz de correlación en la Figura 40 por etnia se observa más relación entre el Síndrome Metabólico (MET_SYM) y ser blanco no hispano.

- **Etnia_1.0:** Mexicano americano.
- **Etnia_2.0:** Hispano no mexicano.
- **Etnia_3.0:** Blanco no hispano.
- **Etnia_4.0:** Negro no hispano.
- **Etnia_6.0:** Asiático no hispano.
- **Etnia_7.0:** Otra raza.

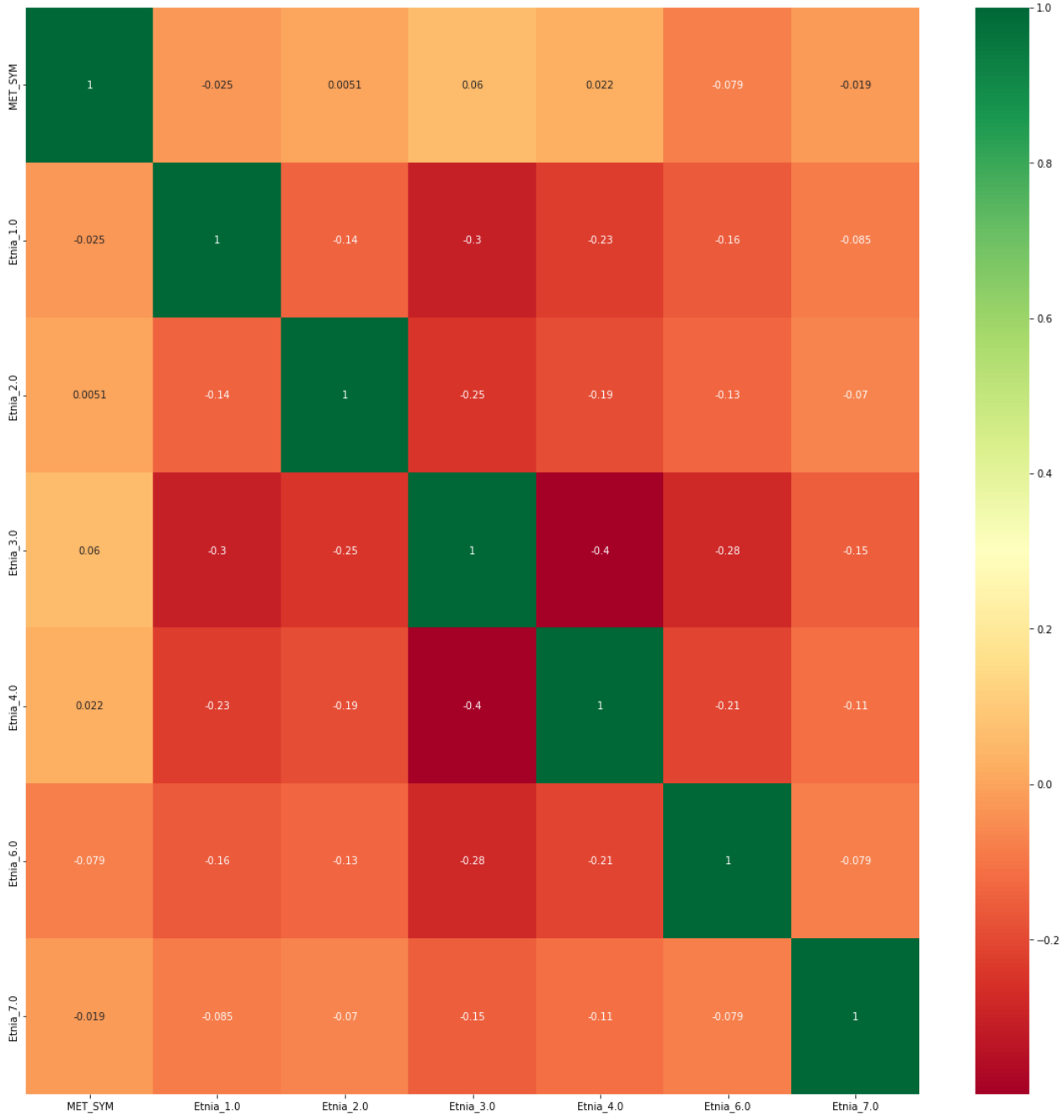

Figura 40: Correlación por etnia

4.3.2 CORRELACIÓN POR VARIABLES DEMOGRÁFICAS

En este punto, en la Figura 41, se visualiza la matriz de correlación por edad (RIDAGEYR y RIDAGEYR_Tramos), sexo (RIAGENDR), etnia (RIDRETH3) y nivel de educación (DMDEDUC2). Como es lógico, la edad es un factor muy influenciable sobre la probabilidad de padecer SM (MET_SYM). También hay una ligera correlación con el nivel educativo (a más nivel educativo menos probabilidad de presentar SM) y el género (como se vio anteriormente los hombres son más propensos).

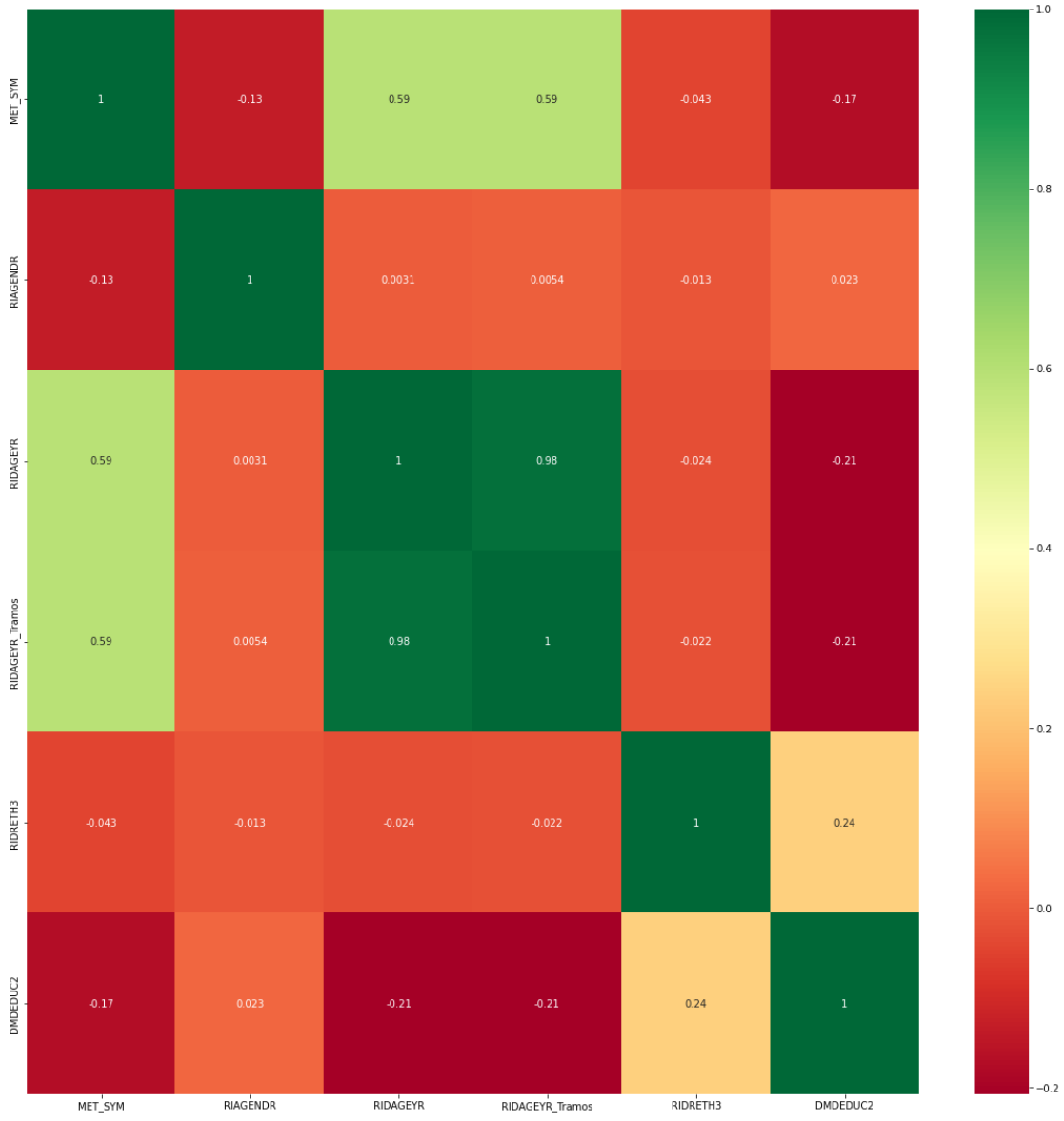

Figura 41: Correlación por variables demográficas

4.3.3 CORRELACIÓN POR EJERCICIO FÍSICO

A continuación, se observa en la Figura 42 como el parámetro PAQ635 que indica si el individuo suele caminar o montar en bicicleta y PAQ650 que indica si en sus planes de ocio el individuo suele realizar actividad física vigorosa, presenta una correlación inversa con padecer SM. PAQ605 que indica si el paciente realiza actividad física vigorosa en su día a día y PAQ620 que indica si el paciente realiza actividad física moderada normalmente no parecen guardar mucha relación con sufrir Síndrome Metabólico.

Hay que tener en cuenta la limitación que supone que este tipo de respuestas dependan de la percepción del individuo. Para objetivar estas cuestiones se debería realizar un seguimiento continuado en el tiempo y riguroso de la actividad física de los participantes. De esta forma, es bastante probable que los resultados pudieran diferir bastante en la matriz de correlación. Aunque NHANES recopile mucha información y sea una de las bases de datos más completas no nos ofrece ese seguimiento de ejercicio físico de los pacientes.

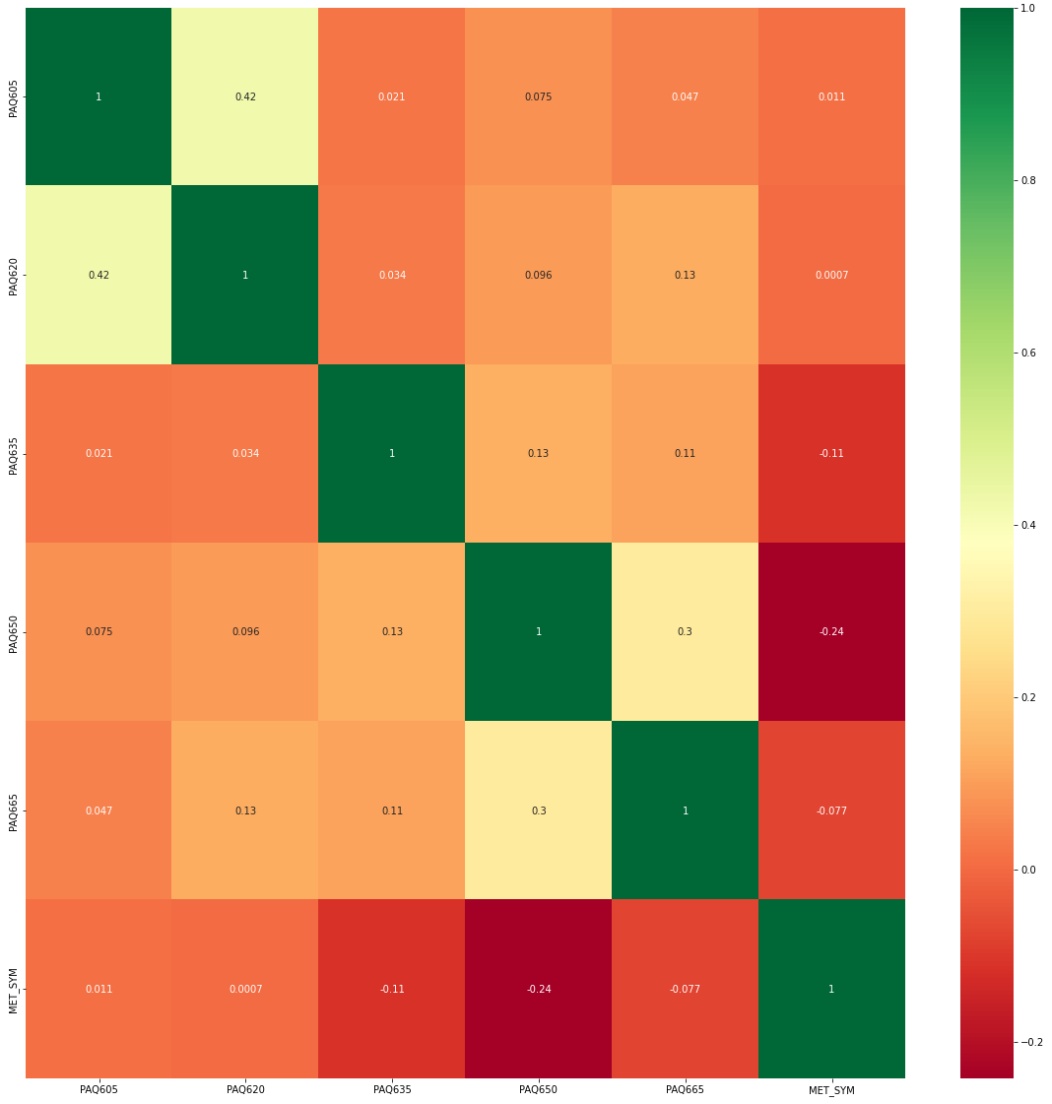

Figura 42: Correlación por ejercicio físico

4.3.4 CORRELACIÓN POR ALCOHOL Y TABAQUISMO

En este caso, se observa una clara relación entre fumar más de 100 cigarrillos a lo largo de la vida del encuestado (SMQ020) y presentar SM (MET_SYM). Mientras que el consumo de alcohol (ALQ130) no parece afectar demasiado. Es interesante señalar que el rango de bebidas alcohólicas de media que consume un individuo puede tratarse de un parámetro poco fiable, especialmente porque es posible que resulte complicado de recordar para el propio encuestado.

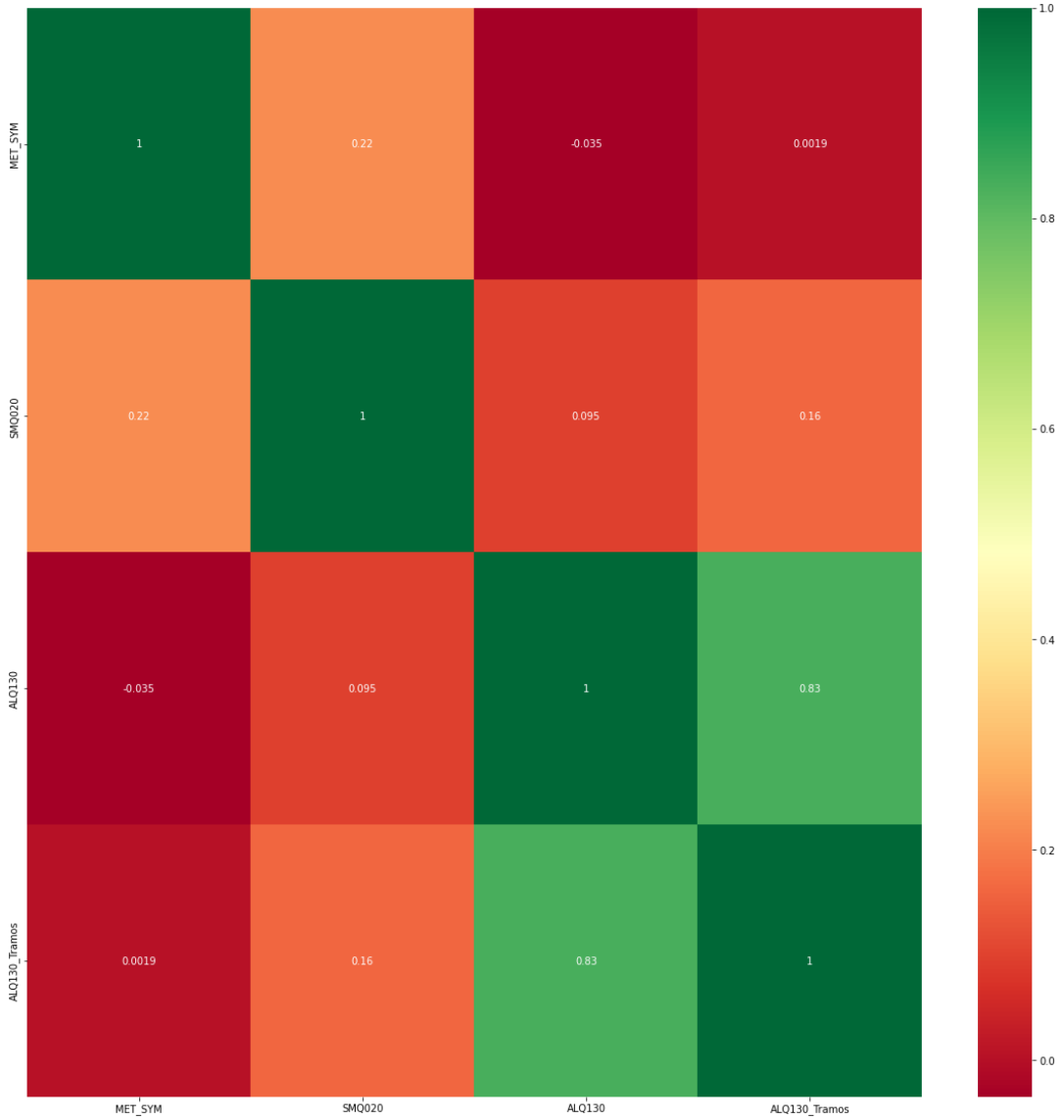

Figura 43: Correlación por alcohol y tabaquismo

4.3.5 CORRELACIÓN POR CALIDAD DIETÉTICA Y ESTADO GENERAL DE SALUD

La variable que indica el estado general de salud del paciente (HSD010) parece guardar bastante correlación con la probabilidad de presentar SM, como se muestra en la Figura 44, mientras que el que evalúa la calidad de su dieta (DBQ700) no. De la misma forma que en los casos previos, evaluar lo saludable que es la dieta de una persona es una tarea bastante laboriosa que requiere mucho seguimiento, no podemos considerar demasiado fiable lo que un paciente opina sobre la calidad de su propia alimentación. Pocas personas son capaces de objetivar la cantidad y la calidad de los alimentos que forman parte de su dieta, por otro lado, el desconocimiento sobre nutrición también supone una limitación para valorar esta cuestión.

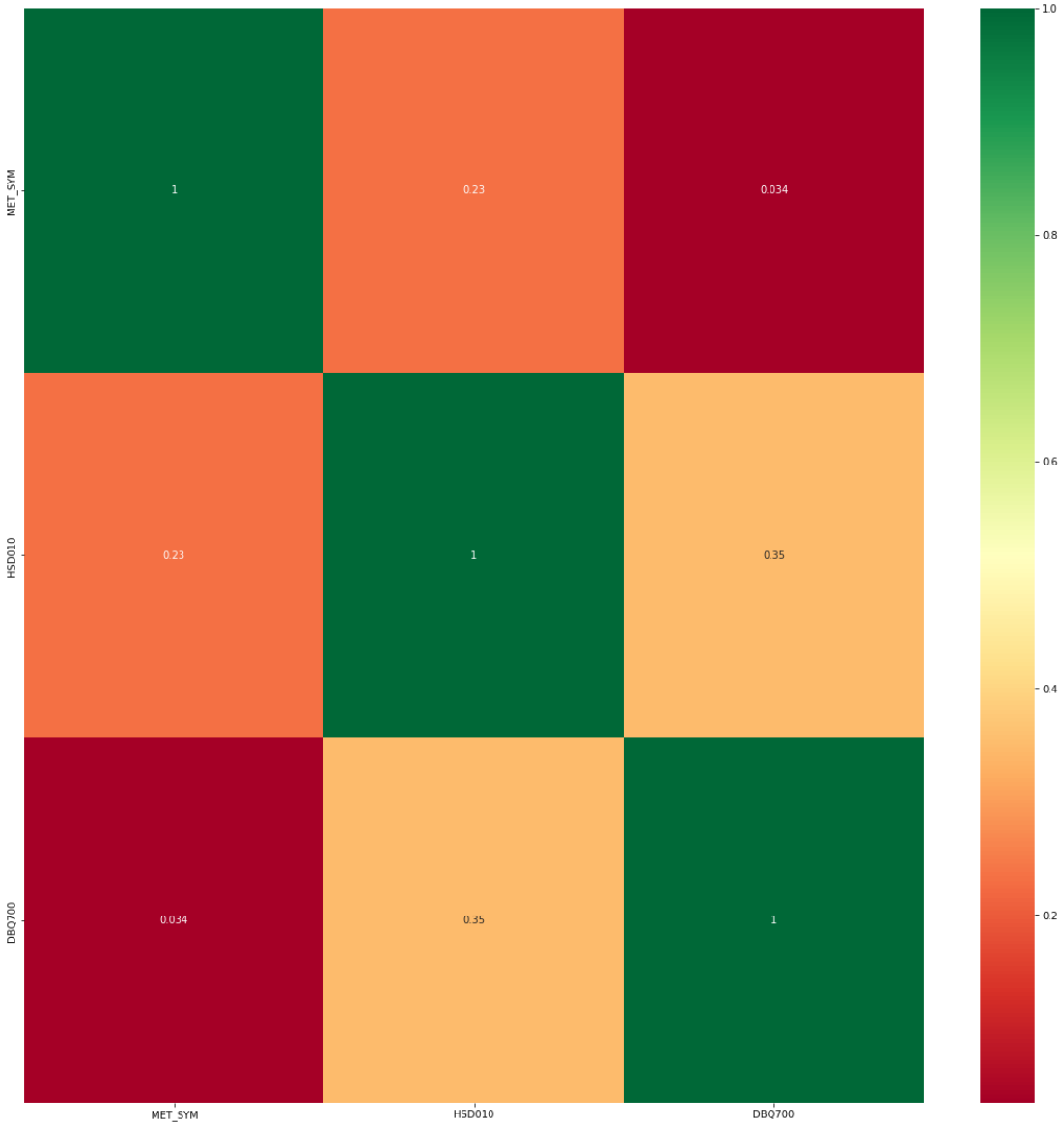

Figura 44: Correlación por calidad de dieta y estado general de salud

4.3.6 CORRELACIÓN POR INDICACIONES DEL DOCTOR SOBRE PRESENTAR SOBREPESO, ASMA, CÁNCER Y DIABETES

Se observa bastante correlación entre que el doctor haga saber al paciente su estado de sobrepeso (MCQ080), cáncer (MCQ220), diabetes (DIQ010) o prediabetes (MCQ300C) y tener Síndrome Metabólico (MET_SYM). Por otra parte, tener asma (MCQ010) no dice mucho sobre el SM.

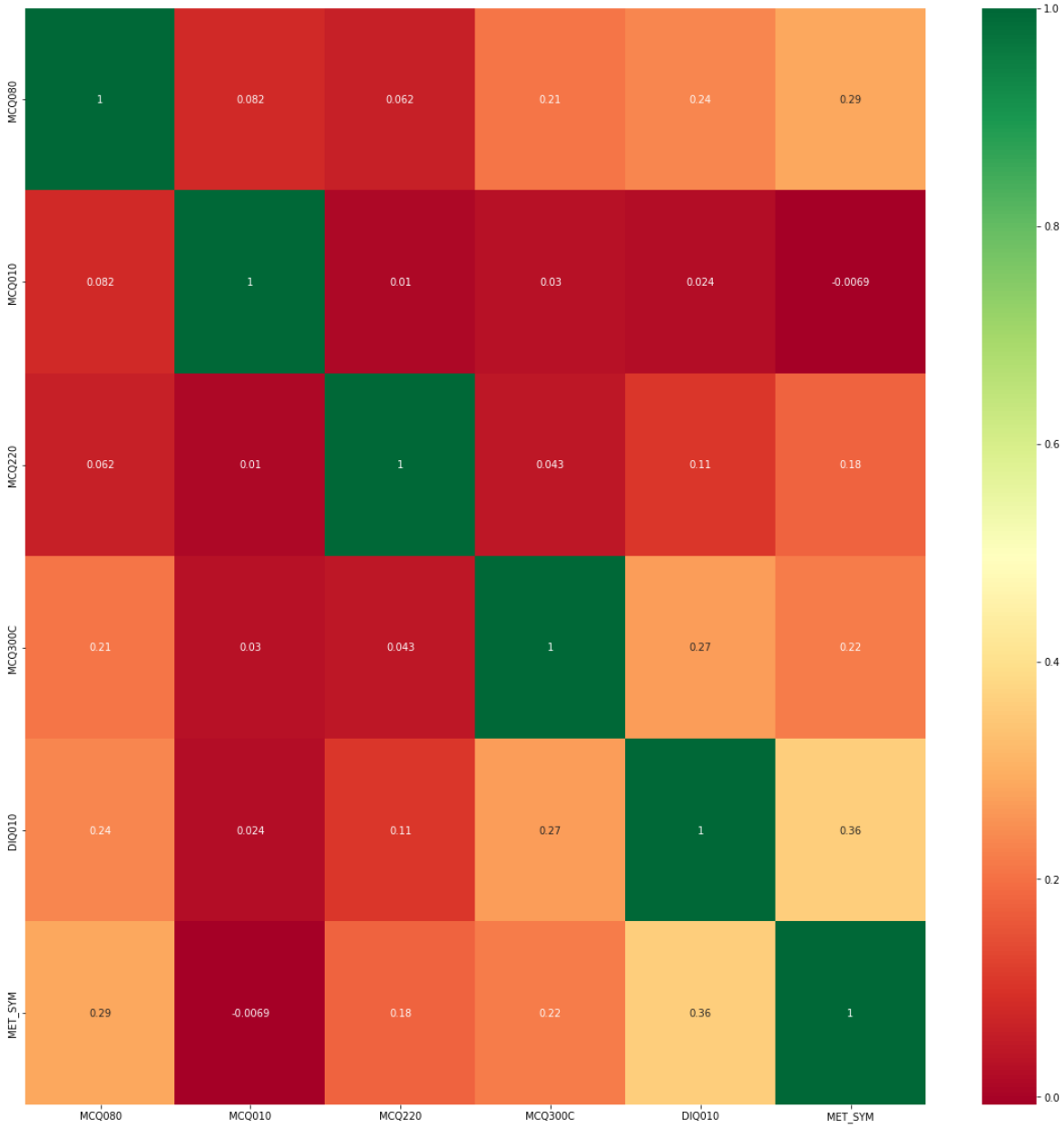

Figura 45: Correlación por sobrepeso, asma, cáncer y diabetes

4.3.7 CORRELACIÓN POR PROBLEMAS PARA DORMIR Y DEPRESIÓN

Se puede establecer, según la Figura 46, una correlación débil entre DPQ020 (sentirse triste, deprimido o desesperanzado), DPQ030 (problemas para dormir o dormir demasiado), DPQ040 (sentirse cansado o con poca energía) y SLQ050 (haber informado al doctor de problemas para conciliar el sueño). DPQ050 (poco apetito o comer en exceso) no parece guardar demasiada relación. Se debe volver a comentar el grado de subjetividad que suponen estas variables. Desde otra perspectiva, sufrir problemas de estrés, insomnio y salud mental puede ser ocasionados por varios factores, entre los que se encuentran enfermedades físicas, pero extrapolar esta variable es una tarea compleja.

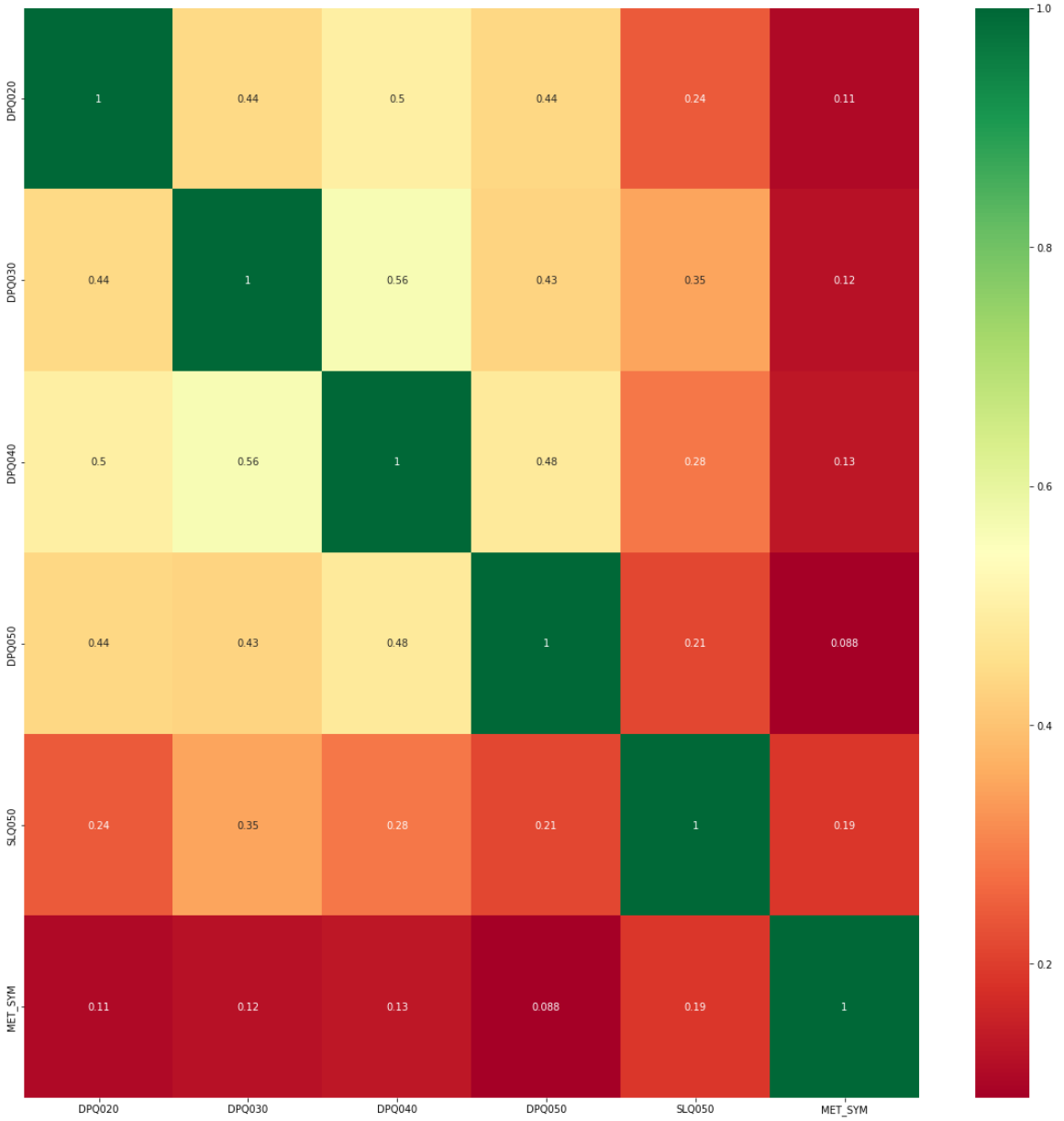

Figura 46: Correlación con problemas de depresión y sueño

4.3.8 CORRELACIÓN CON NIVEL ADQUISITIVO

Se muestra en la Figura 47 la matriz de correlación por nivel adquisitivo. INQ020 indica si el individuo percibe ingresos de su salario, INDFMMPI el índice de pobreza mensual familiar, INDFMMPC la categoría según el índice de pobreza y HIQ011 si el sujeto tiene seguro de salud. No parece haber mucha relación con ninguna de las variables, exceptuando INQ020.

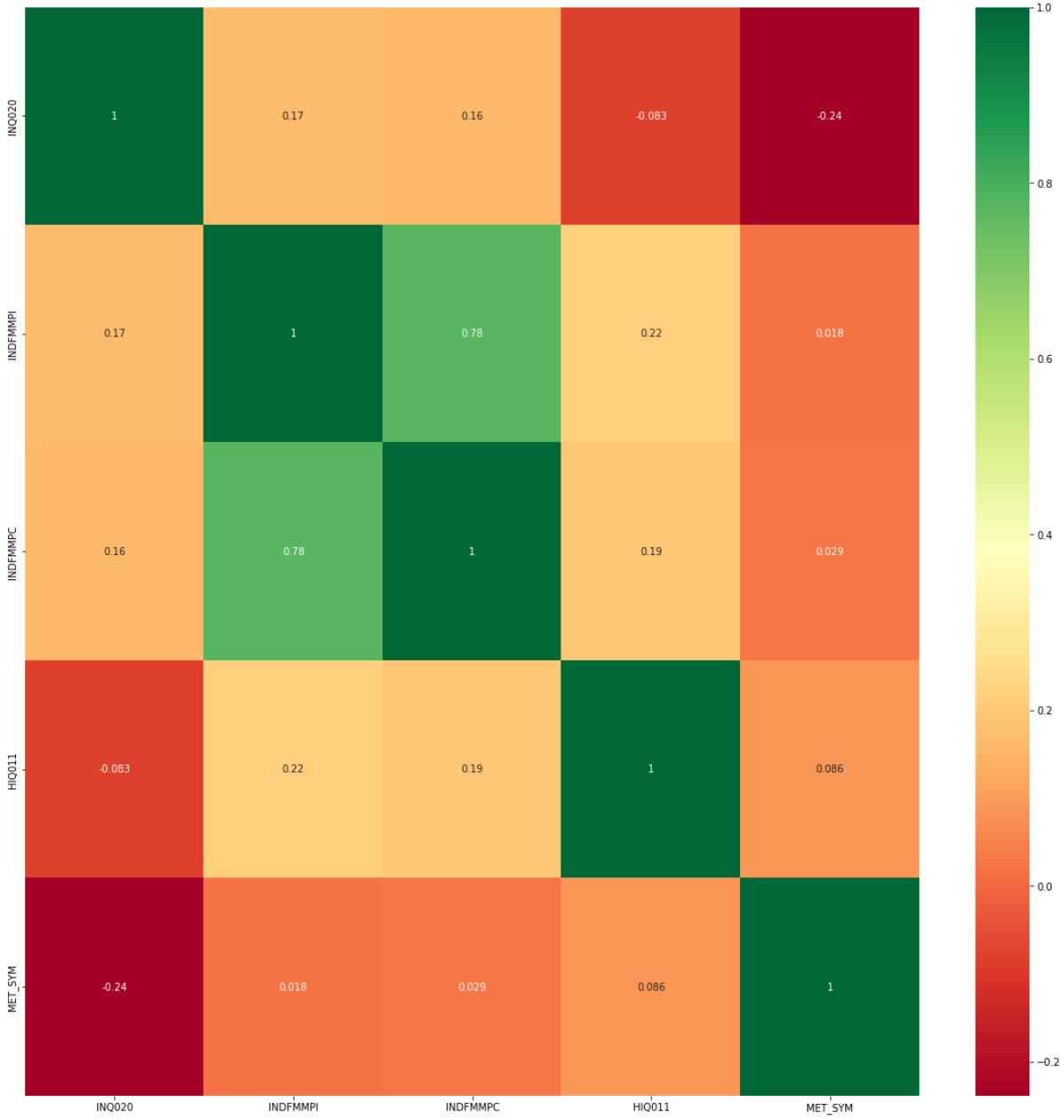

Figura 47: Correlación por nivel adquisitivo

4.3.9 DESCARTE DE VARIABLES

Tras realizar el análisis previo se decide descartar los atributos que menor grado de correlación mantienen con el Síndrome Metabólico: **DPQ050 (poco apetito o comer en exceso), MCQ010 (padecer asma), DBQ700 (calidad de la dieta), ALQ130 (media de bebidas alcohólicas), ALQ130_Tramos (tramos de la media de bebidas alcohólicas), PAQ665 (actividades físicas recreativas moderadas), PAQ620 (ejercicio físico moderado) y PAQ605 (actividad física vigorosa)**.

También se eliminan los atributos relativos al peso y la altura respectivamente (**BMXWT** y **BMXHT**), debido a que han servido solamente para calcular el índice cintura altura y el IMC

4.4 EVALUACIÓN DE LOS MODELOS

Para evaluar los modelos se utilizan las siguientes métricas [124].

- **Matriz de confusión (*confusion matrix*)**: representación matricial de los resultados de las predicciones de cualquier prueba binaria sobre un conjunto de datos de test cuyos valores reales se conocen (Ecuación 9).
- **Exactitud (*accuracy*):** porcentaje de predicciones correctas realizadas sobre el total (Ecuación 6).
- **Precisión (*precision*):** frecuencia con la que el modelo realiza predicciones positivas de forma correcta (Ecuación 4)
- **Exhaustividad (*recall*):** proporción de los verdaderos positivos entre todas las predicciones positivas(Ecuación 5).
- **Puntuación F1 (*F1-score*):** media armónica de la precisión y exhaustividad. Su rango de valores se encuentra entre 0 a 1. Dado que la media armónica de una lista de números se inclina fuertemente hacia últimos elementos de la lista, tiende (en comparación con la media aritmética) a mitigar el impacto de los grandes valores atípicos y a agravar el impacto de los pequeños. Se trata de una métrica de evaluación especialmente efectiva en los escenarios en los que los falsos positivos y los falsos negativos son costosos por igual, cuando añadir más datos no cambia significativamente el resultado y cuando la tasa de verdaderos negativos es alta (como en las predicciones de enfermedades).
- **Pérdidas (*loss*):** solo aplica a la red neuronal y evalúa la desviación entre las predicciones realizadas y los valores reales de las observaciones utilizadas durante el aprendizaje. Cuanto menor es el resultado, más eficiente es la red neuronal [125].
- **Curva ROC (*ROC curve*):** gráfico de rendimiento del modelo en el que se visualiza el ratio de verdaderos positivos (Ecuación 7) y el ratio de verdaderos negativos(Ecuación 8)
- **AUC (*area under the ROC curve*):** Proporciona una medida del rendimiento en todos los umbrales de clasificación posibles. Una interpretación es la de que el modelo clasifique mejor un ejemplo positivo aleatorio que uno negativo aleatorio [126].

$$\boldsymbol{Precisión = \frac{Positivos\ verdaderos}{Positivos\ verdaderos + Positivos\ falsos}}$$

Ecuación 4: Precisión

$$\boldsymbol{Exhaustividad = \frac{Positivos\ verdaderos}{Positivos\ verdaderos + Positivos\ falsos}}$$

Ecuación 5: Exhaustividad

$$\boldsymbol{Exactitud = \frac{Positivos\ verdaderos + Negativos\ verdaderos}{Total}}$$

Ecuación 6: Exactitud

$$\boldsymbol{Ratio\ de\ positivos\ reales} = \frac{\boldsymbol{Positivos\ verdaderos}}{\boldsymbol{Positivos\ verdaderos + Negativos\ falsos}}$$

Ecuación 7: Ratio de positivos reales

$$\boldsymbol{Ratio\ de\ negativos\ reales} = \frac{\boldsymbol{Positivos\ falsos}}{\boldsymbol{Positivos\ falsos + Negativos\ verdaderos}}$$

Ecuación 8: Ratio de negativos reales

$$\boldsymbol{Matriz\ de\ confusión} = \begin{bmatrix} \boldsymbol{Positivos\ verdaderos} & \boldsymbol{Positivos\ falsos} \\ \boldsymbol{Negativos\ falsos} & \boldsymbol{Negativos\ verdaderos} \end{bmatrix}$$

Ecuación 9: Matriz de confusión

4.4.1 EVALUACIÓN DEL MODELO DECISION TREE

Se muestran las métricas de evaluación del modelo en la Tabla 12, así como la curva ROC en la Figura 48. Por otro lado, Scikit-learn nos permite observar una tabla comparando los atributos que más ha tenido en cuenta el modelo para la predicción en la Tabla 13.

Es interesante observar que el atributo con más relevancia para la predicción es BPQ030 (el doctor ha hecho saber al paciente que tiene la presión elevada dos o más veces), seguido de la edad y el índice cintura altura (lo cual no resulta inesperado).

Tabla 12: Evaluación del Decision Tree

Precisión	Exhaustividad	Puntuación F1	Exactitud	AUC ROC
0.80	0.85	0.82	0.85	0.93

$$Matriz\ de\ confusión: \begin{bmatrix} 2742 & 473 \\ 323 & 1835 \end{bmatrix}$$

Ecuación 10: Matriz de confusión del Decision Tree

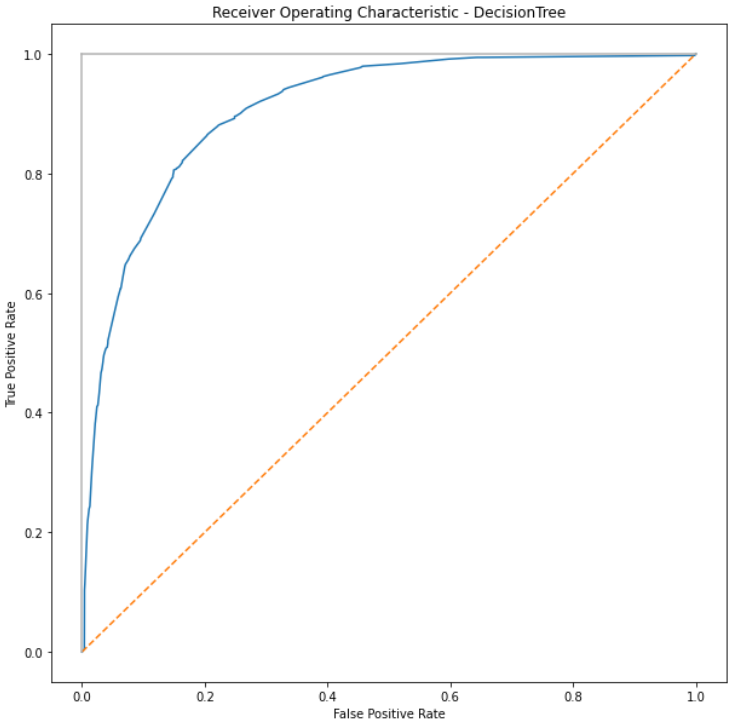

Figura 48: Curva ROC Decision Tree

Tabla 13: Importancia de atributos en Decision Tree

Atributos	Importancia	Atributos	Importancia
BPQ030	0,49872	**HSD010**	0,000325
RIDAGEYR_Tramos	0,212364	**HIQ011**	0,000313
WHI	0,087745	**PAQ635**	0,000252
LBXIN	0,071408	**PAQ650**	0,000181
BMXBMI	0,05328	**DPQ040**	9,64E-05
RIAGENDR	0,05319	**Etnia_2.0**	0
DIQ010	0,016833	**Etnia_6.0**	0
INDFMMPI	0,001661	**Etnia_4.0**	0
SMQ020	0,001035	**Etnia_3.0**	0
Etnia_7.0	0,000762	**MCQ080**	0
DMDEDUC2	0,000703	**Etnia_1.0**	0
MCQ300C	0,000413	**DPQ030**	0
SLQ050	0,00037	**MCQ220**	0
INDFMMPC	0,000351	**INQ020**	0
		DPQ020	0

4.4.2 EVALUACIÓN DEL MODELO RANDOM FOREST

Se muestran las métricas de evaluación en la Tabla 14, la curva ROC en la Figura 49 y los campos que más ha tenido en cuenta el Random Forest en la Tabla 15 a la hora de realizar las predicciones.

Se observa que tiene en consideración más atributos que su antecesor. En ambos casos encabezan la lista BPQ030 y la edad. En tercer lugar, se sitúa la insulina en ayunas seguida del índice cintura altura. En sexto lugar se encuentra el campo DIQ010 (indica si el médico ha avisado al paciente de que tiene diabetes), la diabetes está estrechamente relacionada con el Síndrome Metabólico, aunque no son patologías que necesariamente tengan que presentarse a la vez en un mismo individuo [127].

Tabla 14: Evaluación del Random Forest

Precisión	Exhaustividad	Puntuación F1	Exactitud	AUC ROC
0.82	**0.84**	**0.83**	**0.86**	**0.92**

$$Matriz\ de\ confusión: \begin{bmatrix} 2829 & 386 \\ 347 & 1811 \end{bmatrix}$$

Ecuación 11: Matriz de confusión del Random Forest

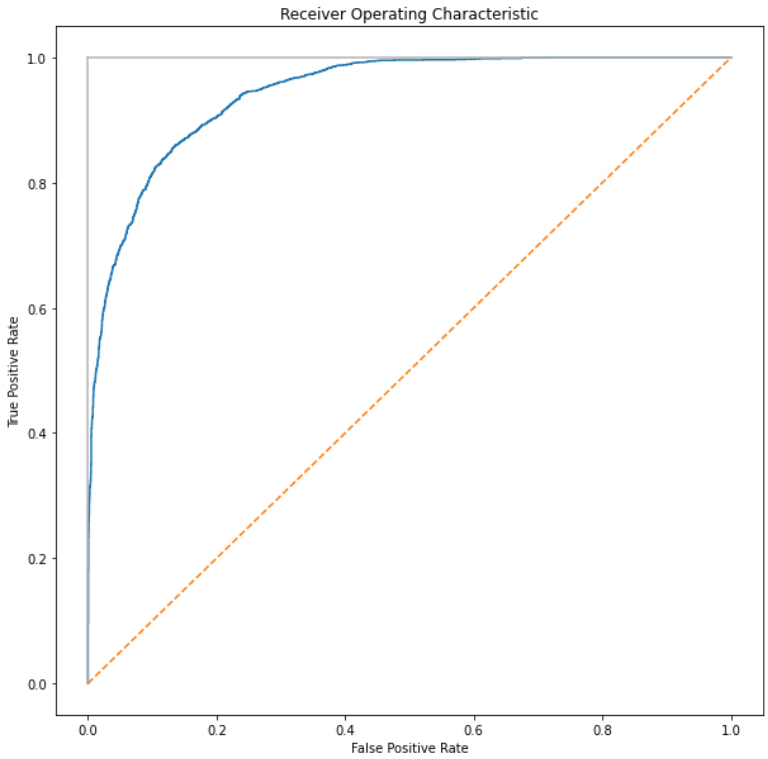

Figura 49: Curva ROC del Random Forest

Tabla 15: Importancia de atributos en Random Forest

Atributos	Importancia	Atributos	Importancia
BPQ030	0,246244	**HSD010**	0,007082
RIDAGEYR_Tramos	0,230107	**SLQ050**	0,005783
LBXIN	0,127377	**MCQ220**	0,003136
WHI	0,119183	**DPQ040**	0,003088
BMXBMI	0,079498	**DPQ030**	0,002944
DIQ010	0,054334	**HIQ011**	0,002923
RIAGENDR	0,030877	**DPQ020**	0,002422
MCQ080	0,01923	**INDFMMPC**	0,00189
INQ020	0,011025	**Etnia_4.0**	0,001662
SMQ020	0,010586	**Etnia_6.0**	0,001606
INDFMMPI	0,010122	**Etnia_3.0**	0,001574
DMDEDUC2	0,008297	**PAQ635**	0,001531
PAQ650	0,007649	**Etnia_1.0**	0,001038
MCQ300C	0,007114	**Etnia_2.0**	0,000877
		Etnia_7.0	0,000801

4.4.3 EVALUACIÓN DEL MODELO SVC

A continuación, se muestran las métricas de evaluación del modelo Support Vector Classifier en la Tabla 16. Al utilizar un kernel que no es lineal, los datos se proyectan a un espacio de características de mayor dimensión, por lo tanto, no se puede visualizar el grado de importancia que asigna el modelo a cada atributo para predecir el Síndrome Metabólico. Se observa que en este caso la ROC es bastante más deficiente en comparación con el resto de las modelos en la Figura 50.

Tabla 16: Evaluación del SVC

Precisión	Exhaustividad	Puntuación F1	Exactitud	AUC ROC
0.83	**0.82**	**0.82**	**0.86**	**0.85**

$$Matriz\ de\ confusión: \begin{bmatrix} 2844 & 371 \\ 387 & 1771 \end{bmatrix}$$

Ecuación 12: Matriz de confusión del SVC

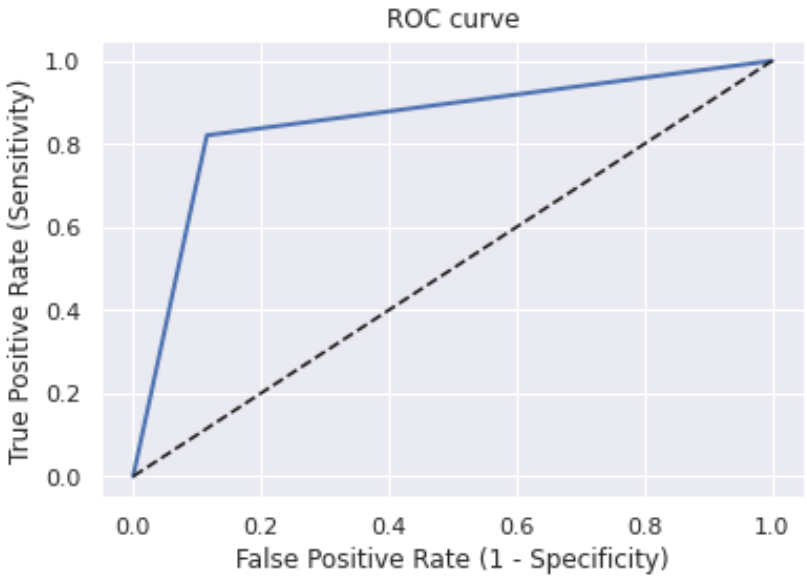

Figura 50: Curva ROC del SVC

4.4.4 EVALUACIÓN DEL MODELO XGBOOST

Se muestran a continuación la evaluación del modelo en la Tabla 17, así como la curva ROC en la Figura 51 y los atributos con más importancia para la predicción del SM en la Tabla 18.

BPQ030 y el tramo de edad siguen siendo los que más relevancia presencian. En tercer lugar, se encuentra el campo DIQ010 (indica si el médico ha avisado al paciente de que tiene diabetes), la diabetes está estrechamente relacionada con el Síndrome Metabólico, aunque no son patologías que necesariamente tengan que presentarse simultáneamente en el mismo individuo [127]. Los campos que preceden a DIQ010 son el índice cintura altura, el género y los niveles de insulina en sangre.

Tabla 17: Métricas del XGBoost

Precisión	Exhaustividad	Puntuación F1	Exactitud	AUC ROC
0.83	**0.85**	**0.84**	**0.87**	**0.95**

$$Matriz\ de\ confusión: \begin{bmatrix} 2833 & 382 \\ 319 & 1839 \end{bmatrix}$$

Ecuación 13: Matriz de confusión del XGBoost

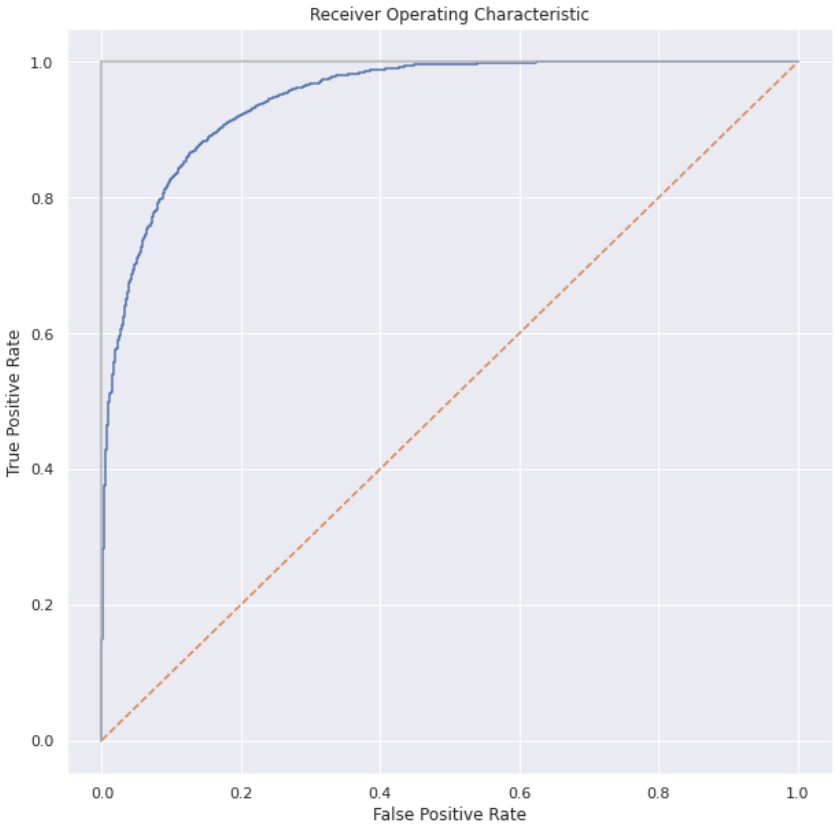

Figura 51: Curva ROC del XGBoost

Tabla 18: Importancia de atributos en XGBoost

Atributos	Importancia	Atributos	Importancia
BPQ030	0,598971	**HIQ011**	0,005652
RIDAGEYR_Tramos	0,127823	**MCQ220**	0,004819
DIQ010	0,039533	**HSD010**	0,004701
WHI	0,035625	**MCQ300C**	0,004625
RIAGENDR	0,031107	**Etnia_6.0**	0,004483
LBXIN	0,028894	**DPQ040**	0,004461
MCQ080	0,017853	**DPQ030**	0,004121
BMXBMI	0,01567	**INDFMMPI**	0,00408
SMQ020	0,012531	**DPQ020**	0,003805
SLQ050	0,00789	**Etnia_7.0**	0,003783
INQ020	0,006983	**PAQ635**	0,003298
DMDEDUC2	0,006166	**Etnia_2.0**	0,003025
PAQ650	0,006076	**Etnia_1.0**	0,002776
Etnia_4.0	0,005909	**Etnia_3.0**	0,002707
		INDFMMPC	0,002634

4.4.5 EVALUACIÓN DEL MODELO NEURAL NETWORK

A continuación, se muestran las métricas de evaluación del modelo Neural Network en la Tabla 19, las gráficas de precisión en relación a los ciclos completos a través del conjunto de datos para entrenar al modelo en las Figura 52 y 53 y la curva ROC en la Figura 55.

Tabla 19: Evaluación de la Red Neuronal

Precisión	Exhaustividad	Puntuación F1	Exactitud	AUC ROC
0.86	**0.94**	**0.86**	**0.87**	**0.94**

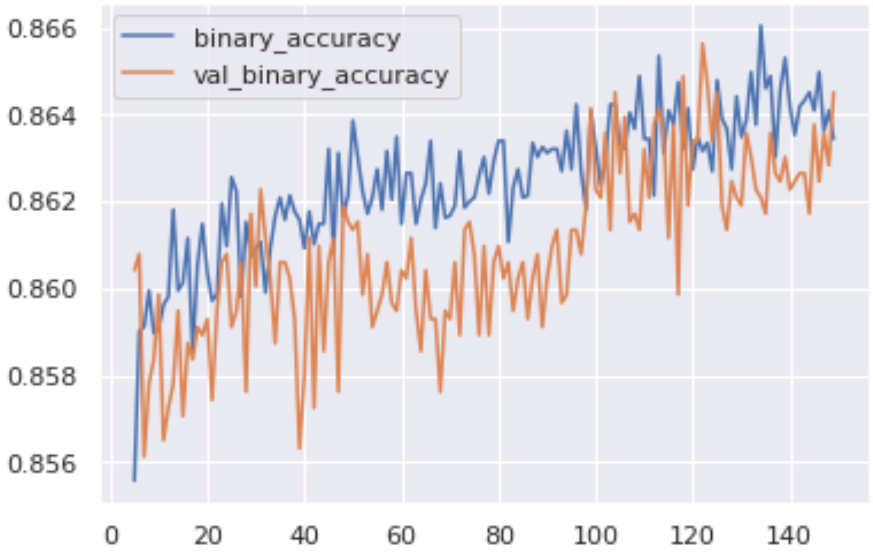

Figura 52*: Binary accuracy* vs *epochs*

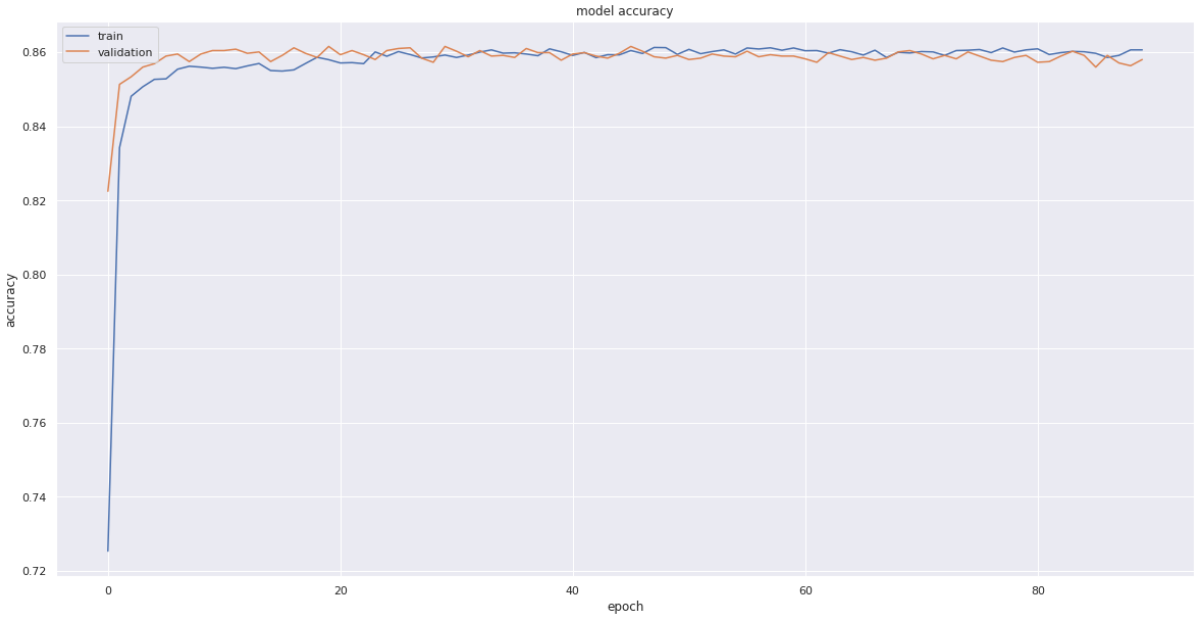

Figura 53: *Accuracy* vs *epochs*

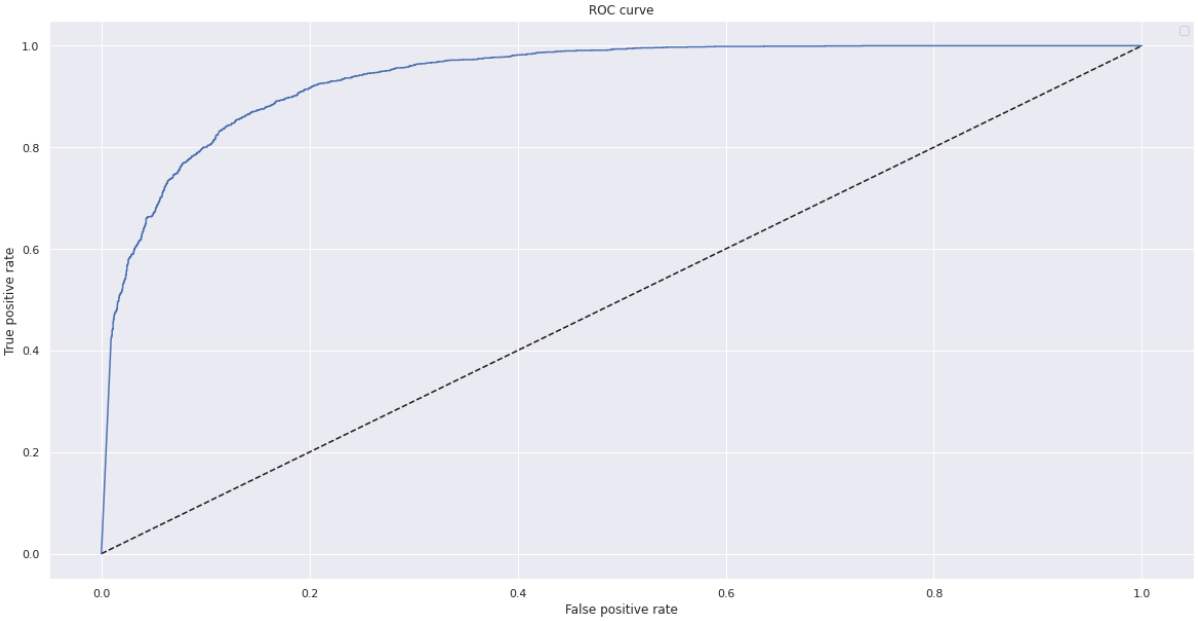

Figura 54: Curva ROC del Neural Network

4.4.6 COMPARATIVA DE LOS MODELOS EVALUADOS

Comparando las métricas de evaluación de todos los modelos en la Tabla 20, así como la curva ROC se puede deducir que, aproximadamente, todos tienen un rendimiento similar en las predicciones, siendo ligeramente mejor el XGBoost.

Tabla 20: Comparativa de las métricas de todos los modelos

	Decision Tree	Random Forest	SVC	XGBoost	Neural Network
Exactitud	0.8519	0.8636	0.8589	**0.8695**	0.8656
Precisión	0.7951	0.8243	0.8268	**0.8280**	0.8875
Exhaustividad	0.8503	0.8392	0.8207	**0.8522**	0.9394
Puntuación-F1	0.8217	0.8317	0.8237	**0.8400**	0.8645
AUC ROC	0.9297	0.9428	0.8526	**0.9472**	0.9396

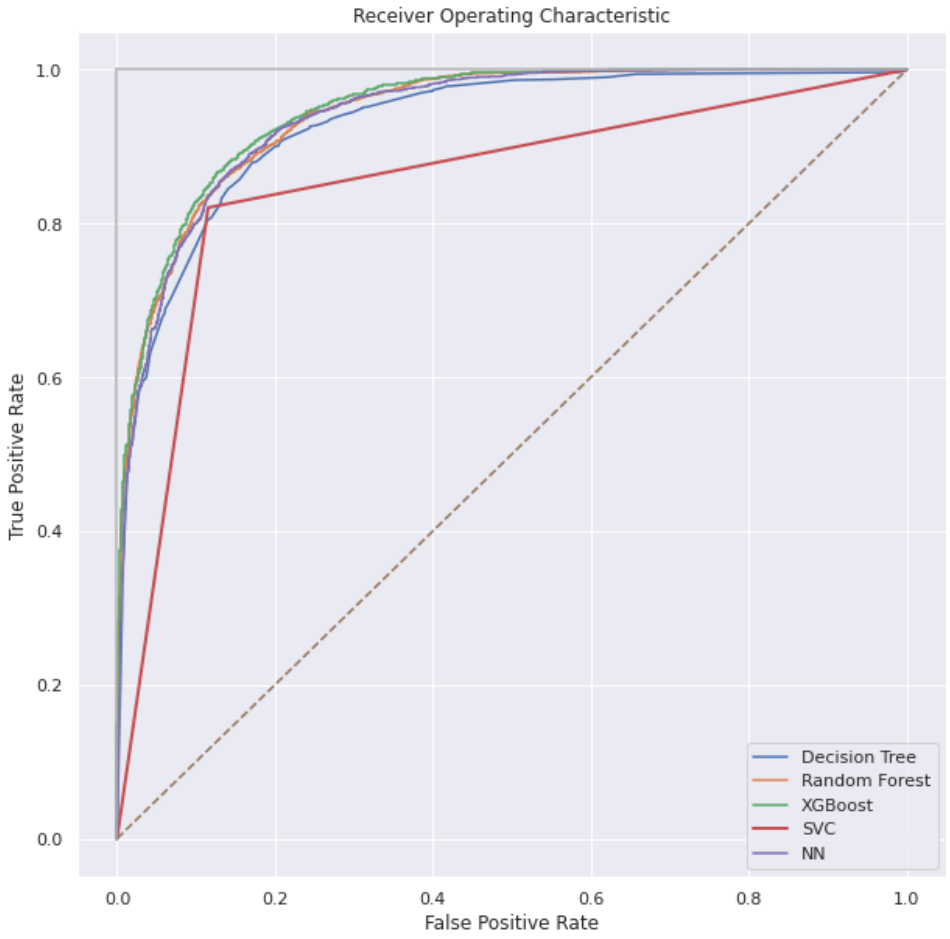

Figura 55: Comparación de las curvas ROC

4.5 VALIDACIÓN DE LOS RESULTADOS

En este punto se realiza una encuesta a profesionales de la salud para valorar el trabajo desarrollado en el presente documento.

Aunque este trabajo se enfoque en el Síndrome Metabólico, se ha contado con la participación de personal sanitario de diversos campos, como por ejemplo enfermería, odontología, fisioterapia y nutrición. Esto es debido a que el empleo de las técnicas de Inteligencia Artificial para la predicción de diferentes patologías puede resultar potencialmente beneficiosas en todos los campos de la salud. Por tanto, además de valorar la herramienta desarrollada en el presente documento, resulta interesante estudiar el concepto que tiene el personal sanitario sobre este tipo de tecnología.

Se recoge una muestra de **118** participantes. En la Figura 56 y la Figura 57 se plasma la distribución del conjunto de los participantes por edad y género, respectivamente.

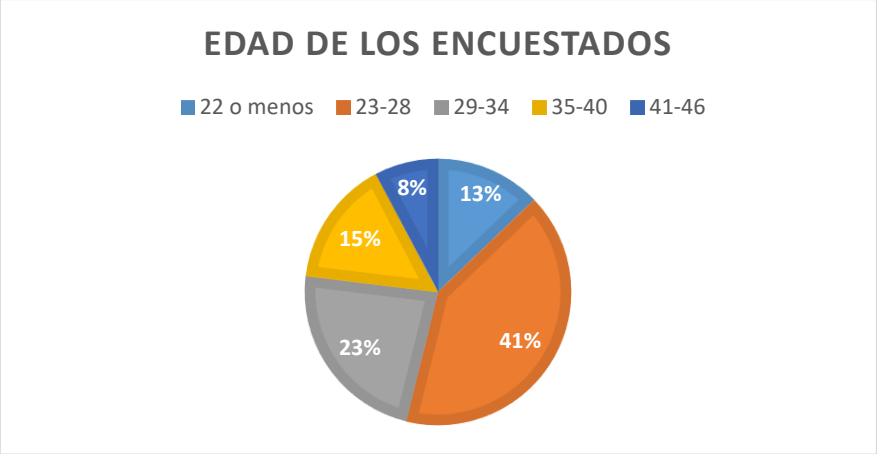

Figura 56: Edad de los encuestados

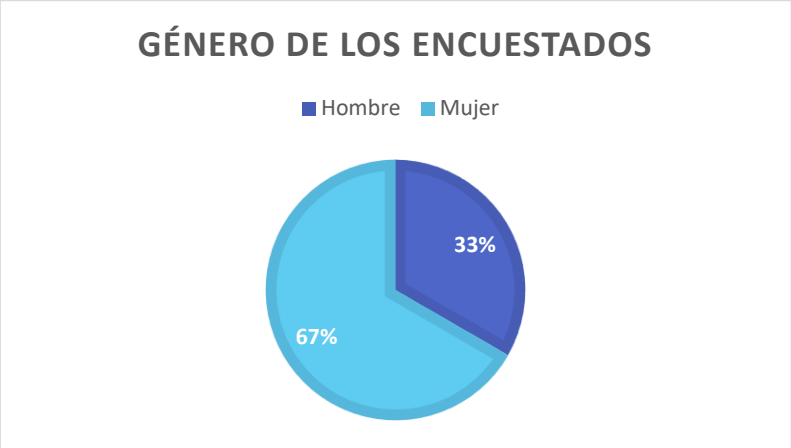

Figura 57: Género de los encuestados

A continuación, se visualiza un gráfico con la especialidad sanitaria de los encuestados en la Figura 58:

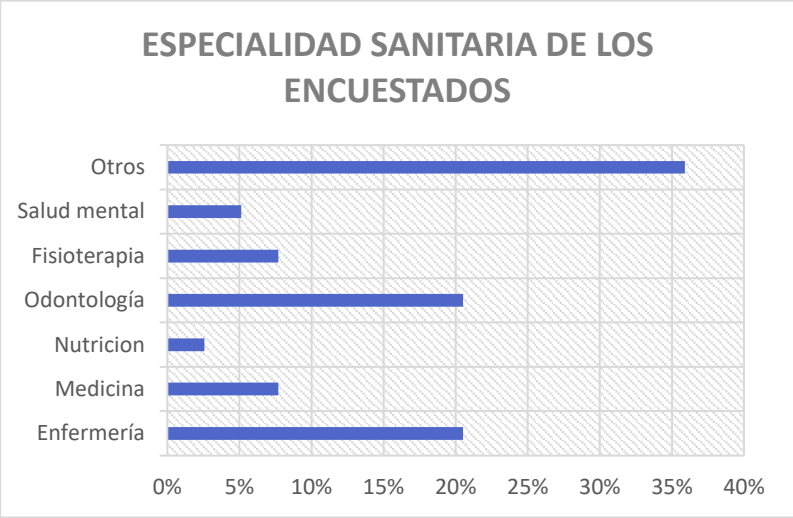

Figura 58: Especialidad sanitaria de los encuestados

Una vez caracterizada la muestra de los participantes del sondeo, se procede a mostrar los resultados de las preguntas del formulario. Para ello se emplea el "marcador de promoción neto" ("*net promoter score*, NPS" en su nombre original en inglés). Consiste en responder mediante una escala del 1 al 10 a la pregunta en cuestión, en función de la puntuación que otorgue, al individuo se le puede introducir en tres categorías. Finalmente, se calcula el marcador de promoción neto restando el número de detractores al de promotores, este marcador puede oscilar entre -100 y 100, si el resultado es positivo se considera un nivel bastante aceptable, si supera los 50 puntos se considera excelente [128]. Se muestra en qué rango de valores se sitúa cada perfil en la Figura 59.

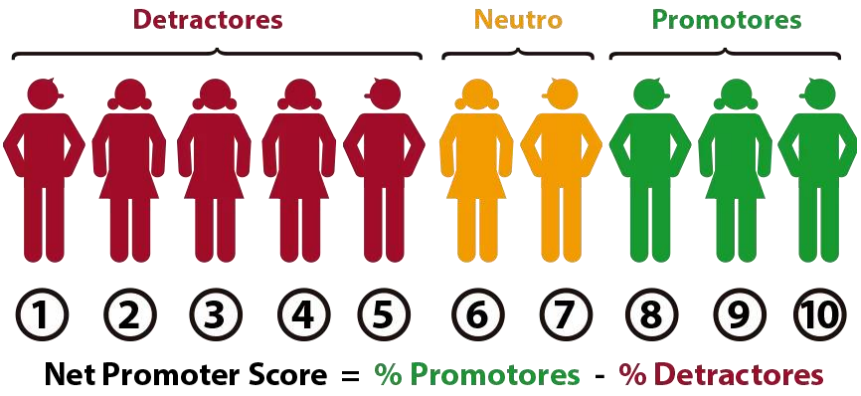

Figura 59: Marcador de promoción neto [129]

- **Pregunta 1:** ¿Cómo valorarías la utilidad en un entorno médico el empleo de estas técnicas? Se obtiene un marcador de promoción neto del **40%.**

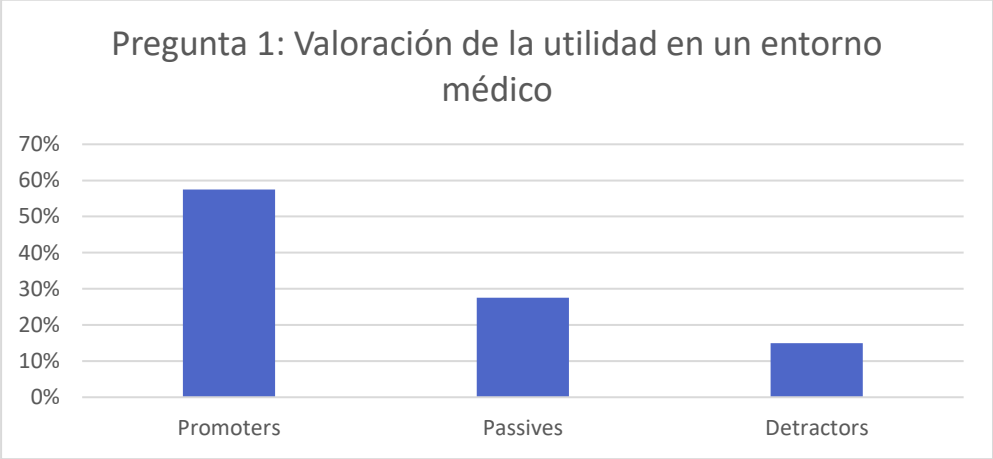

Figura 60: Resultados de la pregunta 1

- **Pregunta 2:** ¿Te parece que el empleo de estas técnicas podría ahorrar recursos y tiempo al personal sanitario y a los centros de salud? Se obtiene un marcador de promoción neto del **42.5%**

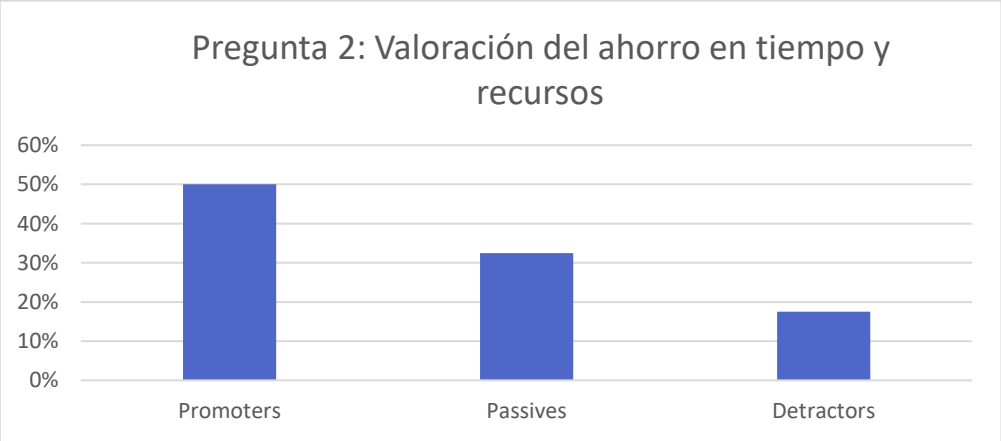

Figura 61: Resultados de la pregunta 2

- **Pregunta 3:** ¿Cuánta mejora piensa que se podría obtener a la hora de prevenir y tratar enfermedades si el empleo de estas técnicas se aplicara en los centros médicos? Se obtiene un marcador de promoción neto del **32.5%.**

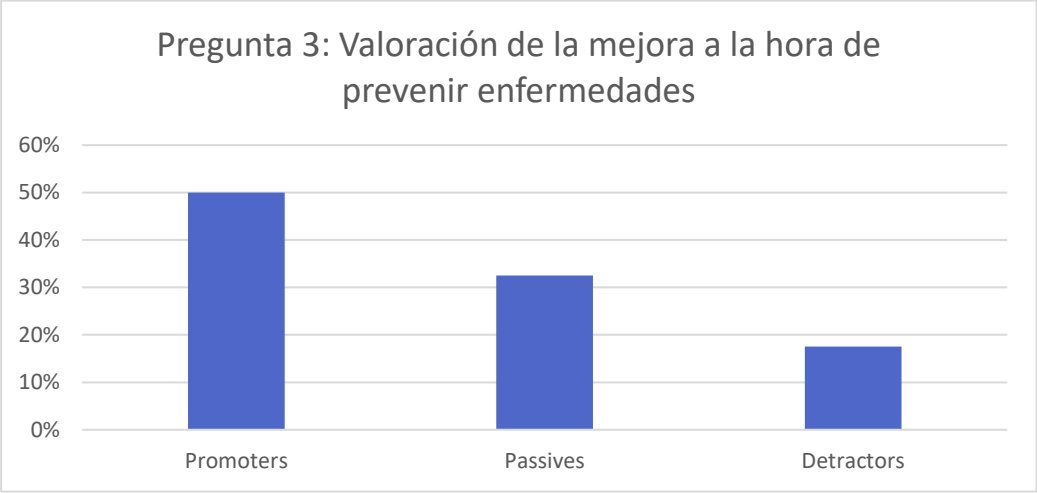

Figura 62: Resultados de la pregunta 3

- **Pregunta 4:** ¿Qué probabilidades habría de que quisiera utilizar esta herramienta en su entorno de trabajo (si se pusiera a su disposición) como ayuda para tomar Decisiones sobre el diagnóstico y tratamiento de un paciente? Se obtiene un marcador de promoción neto del **32.5%.**

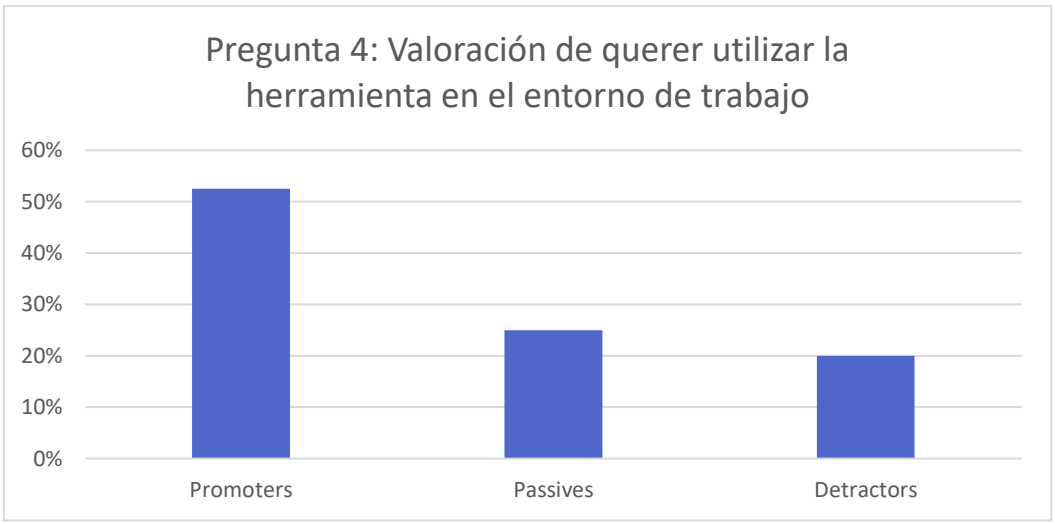

Figura 63: Resultados de la pregunta 4

- **Pregunta 5:** ¿Cuánta desconfianza le producen los modelos de inteligencia artificial para el diagnóstico del Síndrome Metabólico? Se obtiene un marcador de promoción neto del **-57.5%,** en este caso al ser un resultado negativo, nos indica que los encuestados apenas muestran grado de desconfianza en las técnicas expuestas.

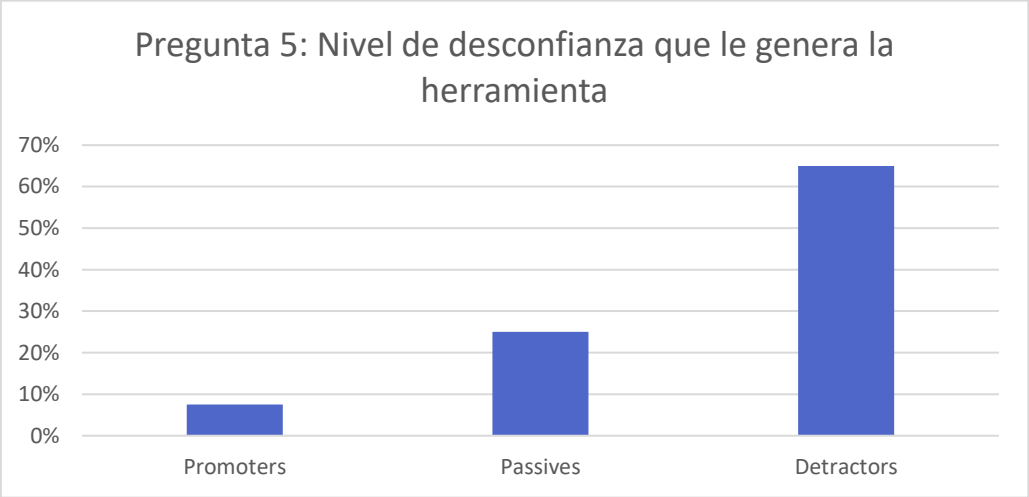

Figura 64: Resultados de la pregunta 5

Se puede extraer de los resultados de la encuesta que el personal de salud considera los modelos de inteligencia artificial como una tecnología bastante buena.

Finalmente, se plantean dos preguntas opcionales de respuesta libre.

- **Pregunta 6:** ¿Qué barreras identificas a la hora de adoptar este tipo de soluciones?
  Algunas de los comentarios que más se han repetido son:

1) Eliminar puestos de trabajo (al ser sustituidos por computadoras).
2) Dificultad para recopilar toda la información necesaria del paciente de manera fiable.
3) Predisposición de los pacientes al requerir su expediente médico para implementar los modelos. Potencial pérdida de criterio propio del profesional
4) Personal reticente para aprender a utilizar estas técnicas.

- **Pregunta 7:** ¿Mejorarías algo de esta herramienta? Algunos de los comentarios a señalar son:

1) Su difusión.
2) La mejora vendrá dada con el tiempo al identificar necesidades específicas.
3) Familiarizar al paciente con este tipo de estrategias, no solo al personal médico.

5. DISCUSIÓN DE LOS RESULTADOS

En este apartado se discuten e interpretan los resultados obtenidos a lo largo del presente TFM. El rendimiento de los modelos de aprendizaje automático es razonablemente bueno. Utilizar técnicas de inteligencia artificial avanzada en un entorno médico para detectar de forma precoz el Síndrome Metabólico podría suponer una herramienta muy útil, especialmente cuando la mayor parte de datos que se deben recopilar del paciente consisten en un formulario de hábitos de vida y unas cuantas mediciones físicas, es decir, que son muy sencillos de obtener.

Las técnicas de inteligencia artificial están irrumpiendo con fuerza en la medicina, provocando una disrupción del paradigma del sector desde diferentes perspectivas. Procesar de manera rápida muchos datos médicos ayuda a que se puedan detectar patologías con pequeños márgenes de error. De esta forma, este proyecto pretende agilizar la detección del Síndrome Metabólico con el consecuente alivio de la carga de trabajo a profesionales médicos.

Al no existir un criterio unificado para diagnosticar Síndrome Metabólico por las diferentes organizaciones que estudian esta enfermedad, tal y como se comentó en el apartado 1.1, a pesar de que las primeras investigaciones al respecto son publicadas hace 90 años por B. C. Hansen et al. [130] se vuelve necesario establecer nuevos mecanismos no convencionales para su diagnóstico. Una buena alternativa es la utilización de modelos de aprendizaje automático. En la literatura no se identifica ningún estudio que haga una comparativa de diferentes modelos para la predicción del SM debido a que la mayoría se limitan a examinar su prevalencia en distintas poblaciones según criterios demográficos.

No obstante, para patologías con menos prevalencia que el SM como por ejemplo la diabetes, sí existen multitud de investigaciones que aplican técnicas de inteligencia artificial avanzadas para predecir la enfermedad. Técnicas que dan unos buenos resultados y que son extensibles a otras enfermedades como el SM, al compartir varias de sus características principales.

Por otro lado, la base de datos NHANES se ha utilizado ampliamente en la investigación de factores de riesgo de patologías como la diabetes [131] u obesidad [132], [133], o en estudios acerca del impacto del tipo de alimentos sobre la ingesta total de comida [134] y sobre la relación del consumo de ultra procesados con las tasas de mortalidad [135]. A pesar de esto, los estudios relacionados con el SM en los que se emplea NHANES se limitan a estudiar su prevalencia a lo largo de los periodos de la base de datos [136]. Procede mencionar la existencia de un estudio que utiliza NHANES de Raúl Sánchez Temporal [137] en el que se aplican modelos de aprendizaje automático no supervisado en la agrupación de pacientes con características semejantes que presenten SM. En contraposición a toda la literatura mencionada, la principal aportación de este trabajo es el desarrollo de modelos de aprendizaje automático supervisado para la predicción del Síndrome Metabólico a partir de los atributos relacionados con los hábitos de vida de los sujetos.

También cabe destacar que, independientemente de la base de datos empleada, existe poca literatura sobre modelos de clasificación aplicados a predecir el Síndrome Metabólico en base al cuestionario sobre hábitos de vida de los individuos, la mayoría se limitan a desarrollar el estudio a partir de examinaciones médicas y de laboratorio. Al ser una enfermedad cuya causa está tan estrechamente relacionada con la dieta, el ejercicio físico y la presencia de un tejido adiposo disfuncional, se identifica la necesidad de tener en cuenta el cuestionario para su predicción.

Con el modelo XGBoost, que ha resultado ser el que mejores resultados obtiene, se puede predecir el SM con una exactitud del 87% y una precisión del 83% a base de datos tomados de un cuestionario de hábitos de vida y mediciones físicas. Esto indica que se pueden desarrollar modelos de inteligencia artificial para diagnosticar el Síndrome Metabólico con altos porcentajes de éxito. El objetivo final es evaluar las probabilidades de un individuo de padecer esta patología para tomar medidas al respecto lo más rápido posible.

En cuanto a los atributos que más parecen guardar relación con la presencia de SM en los individuos se destacan el de presión sanguínea elevada, la edad avanzada, el nivel de insulina en sangre en ayunas alto, presencia de diabetes y un porcentaje graso elevado. Cabe destacar que se ha visto que el índice cintura altura es un marcador más fiable que el índice de masa corporal para evaluar el estado de salud de un sujeto, esto también supone una novedad con respecto al resto de estudios comentados.

Finalmente, se cuenta con la validación de personal sanitario que considera esta herramienta con mucho potencial para ser aplicada en entornos médicos.

5.1 LIMITACIONES DEL ESTUDIO

A pesar de que los resultados de los modelos de predicción han sido bastante satisfactorios, es conveniente destacar algunas limitaciones.

En primer lugar, es necesario señalar que, en el ámbito de la salud, correlación no implica causalidad. Las patologías médicas suelen tener una causa multifactorial y aislar variables resulta complicado, especialmente en el marco de los hábitos de vida de una persona. Los factores genéticos, ambientales y sociales pueden actuar sinérgicamente para causar enfermedad en un sujeto.

Por otro lado, se debe considerar la recopilación de información de la base de datos. La mayor parte de atributos utilizados para el entrenamiento de los algoritmos son frutos de un cuestionario. Estos datos pueden no ser del todo fiables por el grado de subjetividad que suponen, a pesar de haber utilizado un formato que trata de no dar lugar a la ambigüedad en la medida de lo posible. Los datos autoinformados o basados en la percepción de un individuo no se pueden verificar de forma independiente y pueden contener varias fuentes potenciales de sesgo.

Otro factor que es una barrera a la hora de poner en práctica este típico de técnicas es la desconfianza de los pacientes y del personal sanitario, así como la reticencia que pueden experimentar algunos profesionales de la salud a la hora de utilizarlas. Ya que supone el tener que acostumbrarse a una herramienta completamente nueva para ellos que deben aprender desde cero.

Además, cabe mencionar las limitaciones computacionales del estudio, si se hubieran podido destinar más recursos a la optimización de los hiperparámetros los modelos habrían sido más precisos.

Finalmente, se debe señalar el efecto longitudinal, el tiempo disponible para investigar un tema tan extenso como el Síndrome Metabólico es reducido, por lo tanto, es probable que el enfoque de este trabajo no haya sido el más acertado posible.

6. CONCLUSIONES

A lo largo del presente proyecto se ha llevado a cabo una investigación del estado del arte del Síndrome Metabólico y de la aplicación de modelos de inteligencia artificial en la predicción de diversas patologías. Se ha realizado un exhaustivo estudio a través de la base de datos NHANES con el fin de evaluar qué características podían estar relacionadas con el Síndrome Metabólico y obtener la prevalencia del mismo en los sujetos que participaron en la encuesta, en los ciclos que comprenden desde el 2011 al 2018. A partir de ello, se ha podido realizar el análisis estadístico del conjunto de datos y el desarrollo e implementación de los algoritmos de predicción.

En este desarrollo se ha visto que los factores que más influyen en el Síndrome Metabólico son la edad, padecer un estado prediabético o diabético, los niveles elevados de insulina en ayunas, el exceso de porcentaje graso y ser fumador. Es interesante señalar que se ha incluido el índice cintura altura, resultando ser un marcador más fiable que el IMC a la hora de evaluar el estado de sobrepeso de los sujetos. Esto es una novedad, debido a que en prácticamente toda la literatura existente se limitan a utilizar el IMC. En menor grado, la ausencia de actividad física y el índice de pobreza de los sujetos también guarda relación con la presencia del SM.

Por otro lado, la mayor aportación de este trabajo consiste en la implementación de modelos de aprendizaje automático para predecir el Síndrome Metabólico. Esto es algo en lo que se ha indagado muy poco. Tal y como se ha comentado en el estado del arte, la mayor parte de investigaciones en las que aparece una intersección entre el Síndrome Metabólico y la inteligencia artificial se enmarcan en el estudio de su prevalencia. Uno de los intereses principales de este trabajo es ser capaces de predecir una patología que afecta a tantos individuos mediante la información que se puede recopilar fácilmente mediante un formulario. Tales como el grado de actividad física, el nivel de educación, la clase social o la calidad del descanso.

Este trabajo contribuye a dos de los diecisiete Objetivos de Desarrollo Sostenible de la Agenda 2030, los cuales constituyen un llamamiento universal a la acción para mejorar la vida de las personas en todo el mundo. Estos objetivos son el de fomentar la innovación y promover una mejora de la salud global.

Uno de los puntos que más me llama la atención es no haber hallado correlación entre ser de etnia hispana y padecer SM, debido a la multitud de estudios mencionados en el apartado 2.3. Lo cual puede ser debido al grupo muestral del que se han recopilado los datos. Otra consideración que me gustaría remarcar es la confirmación de que el índice cintura altura parece suponer un mejor predictor que el IMC.

Concluyendo este apartado, he alcanzado unos resultados con los que puedo estar satisfecha. Además, el desarrollo de este TFM me ha permitido aprender sobre materias con las que no tenía previa experiencia, la ciencia de datos y la enfermedad del Síndrome Metabólico. Me ha resultado especialmente útil para aprender a enfrentarme a esta clase de problemas en los que se trabaja con una cantidad de datos inmensa y en los que, además, tienen cabida distintas perspectivas y tipos de solución. Todo ello enmarcado en un campo que me apasiona, la salud.

6.1 LÍNEAS FUTURAS

A continuación, se proponen una serie de líneas futuras que se han identificado como mejoras o continuaciones lógicas de este proyecto.

- Tratar de acceder a otras bases de datos similares al NHANES para probar los modelos de aprendizaje automático en otro grupo muestral.
- Desarrollar otros modelos hallados en la literatura, como el de k vecinos más próximos.
- Intentar dar una aplicación de los modelos desarrollados en la predicción de otras enfermedades análogas al SM, como la diabetes o la obesidad.
- Realizar un refinamiento de parámetros más exhaustivo destinando más recursos computacionales.
- Añadir y evaluar otro tipo de atributos de la base de datos que se descartaron en primera instancia para el entrenamiento de los algoritmos.

REFERENCIAS

[1] M. Para Optar Al Grado, "UNIVERSITAT DE LES ILLES BALEARS SÍNDROME METABÓLICO, DIETA Y MARCADORES DE INFLAMACIÓN."

[2] V. Huggo Córdova-Pluma, G. Castro-Martínez, A. Rubio-Guerra, M. E. Hegewisch, and L. Salle, "Breve crónica de la definición del síndrome metabólico," *Med Int Méx*, vol. 30, pp. 312–328, 2014.

[3] K. Shiwaku *et al.*, "Prevalence of the Metabolic Syndrome using the Modified ATP III Definitions for Workers in Japan, Korea and Mongolia," *Journal of Occupational Health*, vol. 47, no. 2, pp. 126–135, Mar. 2005, doi: 10.1539/JOH.47.126.

[4] "Síndrome metabólico. Declaración conjunta, octubre 2009." https://www.elsevier.es/es-revista-endocrinologia-nutricion-12-pdf-S1575092209735247 (accessed May 05, 2022).

[5] A. Jover *et al.*, "Prevalencia del síndrome metabólico y de sus componentes en pacientes con síndrome coronario agudo," *Revista Española de Cardiología*, vol. 64, no. 7, pp. 579–586, Jul. 2011, doi: 10.1016/J.RECESP.2011.03.010.

[6] "Los circuitos de recompensa del cerebro responden de manera diferente a dos tipos de azúcar - El médico interactivo." https://elmedicointeractivo.com/circuitos-recompensa-cerebro-responden-manera-diferente-tipos-azucar-20141212111956056477/ (accessed Jun. 07, 2022).

[7] J. Verdejo-Román, R. Vilar-López, J. F. Navas, C. Soriano-Mas, and A. Verdejo-García, "Brain reward system's alterations in response to food and monetary stimuli in overweight and obese individuals," *Human Brain Mapping*, vol. 38, no. 2, pp. 666–677, Feb. 2017, doi: 10.1002/HBM.23407/ABSTRACT.

[8] "Historia de la obesidad en el mundo (página 2)." https://www.monografias.com/trabajos65/historia-obesidad/historia-obesidad2 (accessed Jun. 07, 2022).

[9] "Visión epistemológica de la obesidad a través de la historia." http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1690-32932010000200011 (accessed May 05, 2022).

[10] D. Fernández-Bergés *et al.*, "Prevalencia de síndrome metabólico según las nuevas recomendaciones de la OMS. Estudio HERMEX Prevalence of metabolic syndrome estimated with the new World Health Organization recommendations. The HERMEX study," *Gac Sanit*, vol. 25, no. 6, pp. 519–524, 2011, doi: 10.1016/j.gaceta.2011.05.009.

[11] C. L. Scott, "Diagnosis, Prevention, and Intervention for the Metabolic Syndrome", doi: 10.1016/S0002-9149(03)00507-1.

[12] C. Arauz-Pacheco, M. A. Parrott, and P. Raskin, "The Treatment of Hypertension in Adult Patients With Diabetes," *Diabetes Care*, vol. 25, no. 1, pp. 134–147, Jan. 2002, doi: 10.2337/DIACARE.25.1.134.

[13] Y. C. Yang, T. Li, and Y. Ji, "Impact of social integration on metabolic functions: Evidence from a nationally representative longitudinal study of US older adults," *BMC Public Health*, vol. 13, no. 1, pp. 1–11, Dec. 2013, doi: 10.1186/1471-2458-13-1210/TABLES/2.

[14] G. E. Miller, M. E. Lachman, E. Chen, T. L. Gruenewald, A. S. Karlamangla, and T. E. Seeman, "Pathways to resilience: maternal nurturance as a buffer against the effects of childhood poverty on metabolic syndrome at midlife.," *Psychol Sci*, vol. 22, no. 12, pp. 1591–9, Dec. 2011, doi: 10.1177/0956797611419170.

[15] M. D. C. Navarro, P. Saavedra, E. Jõdar, M. J. Gõmez De Tejada, A. Mirallave, and M. Sosa, "Osteoporosis and metabolic syndrome according to socio-economic status, contribution of PTH, vitamin D and body weight: The Canarian Osteoporosis Poverty Study (COPS)," *Clinical Endocrinology*, vol. 78, no. 5, pp. 681–686, May 2013, doi: 10.1111/CEN.12051.

[16] S. H. Woolf, "The Power of Prevention and What It Requires," *JAMA*, vol. 299, no. 20, pp. 2437–2439, May 2008, doi: 10.1001/JAMA.299.20.2437.

[17] A. R. Oddo, "Health Insurance: Economic and Ethical Issues," *Research in Ethical Issues in Organizations*, vol. 7, pp. 161–168, 2006, doi: 10.1016/S1529-2096(06)07008-8/FULL/HTML.

[18] K. W. Davidson *et al.*, "Collaboration and Shared Decision-Making between Patients and Clinicians in Preventive Health Care Decisions and US Preventive Services Task Force Recommendations," *JAMA - Journal of the American Medical Association*, vol. 327, no. 12, pp. 1171–1176, Mar. 2022, doi: 10.1001/JAMA.2022.3267.

[19] "Obesidad: un desafío para las políticas públicas." http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-45572013000100007 (accessed May 05, 2022).

[20] M. JJ, "Options for slowing the growth of health care costs.," *N Engl J Med*, vol. 358, no. 14, pp. 1509–1514.

[21] "Objetivos y metas de desarrollo sostenible - Desarrollo Sostenible." https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/ (accessed Jun. 19, 2022).

[22] "Salud - Desarrollo Sostenible." https://www.un.org/sustainabledevelopment/es/health/ (accessed Jun. 19, 2022).

[23] "Infraestructura - Desarrollo Sostenible." https://www.un.org/sustainabledevelopment/es/infrastructure/ (accessed Jun. 19, 2022).

[24] "Los datos son el nuevo petróleo • Red Forbes • Forbes México." https://www.forbes.com.mx/red-forbes-los-datos-son-el-nuevo-petroleo/ (accessed Jun. 10, 2022).

[25] "Los datos, ¿el nuevo petróleo? - deepsense | Inteligencia Artificial a medida de la empresa." https://deepsense.es/es/blog/los-datos-el-nuevo-petroleo/ (accessed Jun. 10, 2022).

[26] "3 tipos de análisis de datos para mejorar la toma de decisiones." https://www.pragma.com.co/blog/3-tipos-de-analisis-de-datos-para-mejorar-la-toma-de-decisiones (accessed May 05, 2022).

[27] "Análisis de datos: tipos, herramientas y fases para llevarlo a cabo." https://tutfg.es/analisis-de-datos (accessed May 05, 2022).

[28] "Análisis Descriptivo, Predictivo y Prescriptivo de datos - IArtificial.net." https://www.iartificial.net/analisis-predictivo-y-prescriptivo-con-machine-learning/ (accessed May 05, 2022).

[29] "En 1956, John McCarthy acuñó la expresión «inteligencia artificial», y la definió como «la ciencia e ingenio de hacer máquinas inteligentes» – Empleo UGR. Centro de Promoción de Empleo y Prácticas." https://empleo.ugr.es/blog-empleo/en-1956-john-mccarthy-acuno-la-expresion-inteligencia-artificial-y-la-definio-como-la-ciencia-e-ingenio-de-hacer-maquinas-inteligentes/ (accessed May 05, 2022).

[30] I. Castiglioni *et al.*, "AI applications to medical images: From machine learning to deep learning," *Physica Medica*, vol. 83, pp. 1120–1797, 2021, doi: 10.1016/j.ejmp.2021.02.006.

[31] I. el Naqa and M. J. Murphy, "What Is Machine Learning?," *Machine Learning in Radiation Oncology*, pp. 3–11, 2015, doi: 10.1007/978-3-319-18305-3_1.

[32] M. van Otterlo and M. Wiering, *Reinforcement learning and markov decision processes*, vol. 12. Springer Verlag, 2012. doi: 10.1007/978-3-642-27645-3_1.

[33] P. Langley, "The changing science of machine learning," *Machine Learning*, vol. 82, no. 3, pp. 275–279, Mar. 2011, doi: 10.1007/s10994-011-5242-y.

[34] J. Cerda, G. Valdivia, C. J. Snow, and J. Cerda Lorca, "John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna," *Revista chilena de infectología*, vol. 24, no. 4, pp. 331–334, Aug. 2007, doi: 10.4067/S0716-10182007000400014.

[35] "Vista de Las TIC en el sector salud, machine learning para el diagnóstico y prevención de enfermedades." https://www.revistacuantica.com/index.php/rcq/article/view/27/29 (accessed May 05, 2022).

[36] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective."

[37] N. H. Quiroz, M. Lourdes Posadas-Martínez, E. Rossi, D. H. Giunta, and M. R. Risk, "Aprendizaje automático aplicado en área de la salud. Parte 2," *Revista del Hospital Italiano de Buenos Aires*, vol. 42, no. 1, pp. 56–58, Mar. 2022, doi: 10.51987/REVHOSPITALBAIRES.V42I1.152.

[38] "El análisis de datos en el sector de la salud | Blog | España | Merkle." https://www.merkleinc.com/es/es/blog/analisis-datos-sector-salud (accessed Jun. 10, 2022).

[39] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput Struct Biotechnol J*, vol. 15, pp. 104–116, 2017, doi: 10.1016/J.CSBJ.2016.12.005.

[40] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Frontiers in Genetics*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/FGENE.2018.00515/BIBTEX.

[41] "Máquinas de vectores de soporte - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte (accessed May 05, 2022).

[42] "1.10. Decision Trees — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/modules/tree.html (accessed May 05, 2022).

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[44] "sklearn.ensemble.RandomForestClassifier — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed May 05, 2022).

[45] "(PDF) Métodos de Ensembles para Machine Learning." https://www.researchgate.net/publication/294582534_Metodos_de_Ensembles_para_Machine_Learning (accessed Jun. 08, 2022).

[46] "Machine Learning with XGBoost and Scikit-learn | Engineering Education (EngEd) Program | Section." https://www.section.io/engineering-education/machine-learning-with-xgboost-and-scikit-learn/ (accessed May 11, 2022).

[47] "Gradient boosting - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/Gradient_boosting (accessed May 11, 2022).

[48] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[49] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, Sep. 2010, doi: 10.1162/neco_a_00052.

[50] E. A. Petrova, A. E. Kalinina, and P. v. Bondarenko, "Neural Network Prediction of Economic Structural Changes in the Context of Industry 4.0," *Smart Innovation, Systems and Technologies*, vol. 287, pp. 151–162, 2022, doi: 10.1007/978-981-16-9804-0_13.

[51] M. Keilbach, K. Obermayer, R. Herbrich, T. Graepel, and P. Bollmann-Sdorra, "Particle Tracking Velocimetry View project state-space Ising model View project NEURAL NETWORKS IN ECONOMICS Background, Applications and New Developments", doi: 10.1007/978-1-4615-5029-7_7.

[52] Z. H. Khan, S. K. Mohapatra, P. K. Khodiar, and S. N. R. Kumar, "Artificial neural network and medicine.," *Indian Journal of Physiology and Pharmacology*, vol. 42, no. 3, pp. 321–342, Jul. 1998, Accessed: Jun. 08, 2022. [Online]. Available: https://europepmc.org/article/med/9741647

[53] "YouTube - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/YouTube (accessed Jun. 10, 2022).

[54] "Amazon.es: compra online de electrónica, libros, deporte, hogar, moda y mucho más." https://www.amazon.es/ (accessed Jun. 17, 2022).

[55] "8 Applications of Neural Networks | Analytics Steps." https://www.analyticssteps.com/blogs/8-applications-neural-networks (accessed May 05, 2022).

[56] "La diabetes entra en el 'top ten' de enfermedades con más mortalidad – Solidaridad Intergeneracional." https://solidaridadintergeneracional.es/wp/la-diabetes-entra-en-el-top-ten-de-enfermedades-con-mas-mortalidad/ (accessed Jun. 08, 2022).

[57] M. G. Saklayen, "HYPERTENSION AND OBESITY (E REISIN, SECTION EDITOR) The Global Epidemic of the Metabolic Syndrome," 1906, doi: 10.1007/s11906-018-0812-z.

[58] "Framingham Study | Boston Medical Center." https://www.bmc.org/stroke-and-cerebrovascular-center/research/framingham-study (accessed May 05, 2022).

[59] "(PDF) Síndrome metabólico: definición, historia, criterios." https://www.researchgate.net/publication/26508500_Sindrome_metabolico_definicion_historia_criterios (accessed May 05, 2022).

[60] E. S. Ford, W. H. Giles, and W. H. Dietz, "Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey," *JAMA*, vol. 287, no. 3, pp. 356–359, Jan. 2002, doi: 10.1001/JAMA.287.3.356.

[61] J. Martínez Candela, J. Franch Nadal, J. Romero Ortiz, C. Cánovas Domínguez, A. Gallardo Martín, and M. Páez Pérez, "Prevalencia del síndrome metabólico en la población adulta de Yecla (Murcia). Grado de acuerdo entre tres definiciones," *Atención Primaria*, vol. 38, no. 2, pp. 72–79, Jun. 2006, doi: 10.1157/13090435.

[62] J. X. Moore, N. Chaudhary, and T. Akinyemiju, "Metabolic Syndrome Prevalence by Race/Ethnicity and Sex in the United States, National Health and Nutrition Examination Survey, 1988–2012," *Preventing Chronic Disease*, vol. 14, no. 3, 2019, doi: 10.5888/PCD14.160287.

[63] G. O. Gutiérrez-Esparza, O. I. Vázquez, M. Vallejo, and J. Hernández-Torruco, "Prediction of Metabolic Syndrome in a Mexican Population Applying Machine Learning Algorithms," *Symmetry 2020, Vol. 12, Page 581*, vol. 12, no. 4, p. 581, Apr. 2020, doi: 10.3390/SYM12040581.

[64] C. S. Yu *et al.*, "Predicting Metabolic Syndrome With Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study," *JMIR Med Inform 2020;8(3):e17110 https://medinform.jmir.org/2020/3/e17110*, vol. 8, no. 3, p. e17110, Mar. 2020, doi: 10.2196/17110.

[65] "(PDF) Predicting metabolic syndrome using decision tree and support vector machine methods." https://www.researchgate.net/publication/309630966_Predicting_metabolic_syndrome_using_decision_tree_and_support_vector_machine_methods (accessed May 05, 2022).

[66] E. K. Choe *et al.*, "Metabolic Syndrome Prediction Using Machine Learning Models with Genetic and Clinical Information from a Nonobese Healthy Population," *Genomics & Informatics*, vol. 16, no. 4, p. e31, Dec. 2018, doi: 10.5808/GI.2018.16.4.E31.

[67] H. A. Kakudi, C. K. Loo, and F. M. Moy, "Diagnosis of Metabolic Syndrome using Machine Learning, Statistical and Risk Quantification Techniques: A Systematic Literature Review," *medRxiv*, p. 2020.06.01.20119339, Jun. 2020, doi: 10.1101/2020.06.01.20119339.

[68] "Cardiorrenal.es | Diabetes Mellitus tipo 2 - Prevalencia." https://www.cardiorrenal.es/patologia-DM2-prevalencia (accessed Jun. 15, 2022).

[69] "Estadísticas del cáncer - NCI." https://www.cancer.gov/espanol/cancer/naturaleza/estadisticas (accessed Jun. 15, 2022).

[70] "Prevalencia del síndrome metabólico en población de 15 a 74 años del municipio Guantánamo | Gómez Torres | Revista Información Científica." http://www.revinfcientifica.sld.cu/index.php/ric/article/view/290/2819 (accessed Jun. 15, 2022).

[71] J. Salas *et al.*, "PUESTA AL DÍA SOBRE EL PACIENT … SPA A UT S PUESTA AL DIA SOBRE EL PACIENT E DIABET ICO Y SINDROME METABOLICO," *Nut rición Hospit alaria SUPLEMENT OS*, vol. 3, no. 1, 2010.

[72] "NHANES - National Health and Nutrition Examination Survey Homepage." https://www.cdc.gov/nchs/nhanes/index.htm (accessed May 05, 2022).

[73] C. H. H. Le, "The Prevalence of Anemia and Moderate-Severe Anemia in the US Population (NHANES 2003-2012)," *PLOS ONE*, vol. 11, no. 11, p. e0166635, Nov. 2016, doi: 10.1371/JOURNAL.PONE.0166635.

[74] Y. Ostchega, J. P. Hughes, A. Terry, T. H. I. Fakhouri, and I. Miller, "Abdominal obesity, body mass index, and hypertension in US adults: NHANES 2007-2010," *American Journal of Hypertension*, vol. 25, no. 12, pp. 1271–1278, Dec. 2012, doi: 10.1038/AJH.2012.120/2/AJH.1271.F1.JPEG.

[75] B. Dawson-Hughes *et al.*, "The potential impact of the National Osteoporosis Foundation guidance on treatment eligibility in the USA: an update in NHANES 2005-2008," *Osteoporos Int*, vol. 23, pp. 811–820, 2012, doi: 10.1007/s00198-011-1694-y.

[76] "NHANES Questionnaires, Datasets, and Related Documentation." https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011 (accessed May 05, 2022).

[77] "NHANES Questionnaires, Datasets, and Related Documentation." https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013 (accessed May 05, 2022).

[78] "NHANES Questionnaires, Datasets, and Related Documentation." https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015 (accessed May 05, 2022).

[79] "NHANES Questionnaires, Datasets, and Related Documentation." https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017 (accessed May 05, 2022).

[80] "DEMO_D." https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm (accessed Jun. 16, 2022).

[81] M. Ashwell, L. Mayhew, J. Richardson, and B. Rickayzen, "Waist-to-Height Ratio Is More Predictive of Years of Life Lost than Body Mass Index," 2014, doi: 10.1371/journal.pone.0103483.

[82] C. M. Y. Lee, R. R. Huxley, R. P. Wildman, and M. Woodward, "Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis'," *Journal of Clinical Epidemiology*, vol. 61, no. 7, pp. 646–653, Jul. 2008, doi: 10.1016/j.jclinepi.2007.08.012.

[83] "Índice cintura-altura - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/%C3%8Dndice_cintura-altura (accessed May 08, 2022).

[84] "Welcome to Python.org." https://www.python.org/ (accessed May 05, 2022).

[85] "NumPy." https://numpy.org/ (accessed May 05, 2022).

[86] "pandas - Python Data Analysis Library." https://pandas.pydata.org/ (accessed May 05, 2022).

[87] "Matplotlib — Visualization with Python." https://matplotlib.org/ (accessed May 05, 2022).

[88] M. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/JOSS.03021.

[89] "SciPy." https://scipy.org/ (accessed May 05, 2022).

[90] "scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/ (accessed May 05, 2022).

[91] "imbalanced-learn documentation — Version 0.9.0." https://imbalanced-learn.org/stable/ (accessed May 05, 2022).

[92] "Keras: the Python deep learning API." https://keras.io/ (accessed May 08, 2022).

[93] "Project Jupyter | Home." https://jupyter.org/ (accessed May 05, 2022).

[94] “TensorFlow.” https://www.tensorflow.org/?hl=es-419 (accessed May 08, 2022).

[95] “Te damos la bienvenida a Colaboratory - Colaboratory.” https://colab.research.google.com/ (accessed May 05, 2022).

[96] “Detección de outliers en Python | Aprende Machine Learning.” https://www.aprendemachinelearning.com/deteccion-de-outliers-en-python-anomalia/ (accessed May 05, 2022).

[97] “PCA con Python.” https://www.cienciadedatos.net/documentos/py19-pca-python.html (accessed May 08, 2022).

[98] “sklearn.preprocessing.StandardScaler — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed May 08, 2022).

[99] “sklearn.decomposition.PCA — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (accessed May 08, 2022).

[100] “sklearn.model_selection.train_test_split — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed May 08, 2022).

[101] “A Gentle Introduction to Imbalanced Classification.” https://machinelearningmastery.com/what-is-imbalanced-classification/ (accessed May 08, 2022).

[102] “El problema de desequilibrio de clases en conjuntos de datos de entrenamiento - Analytics Lane.” https://www.analyticslane.com/2018/07/04/el-problema-de-desequilibrio-de-clases-en-conjuntos-de-datos-de-entrenamiento/ (accessed May 08, 2022).

[103] “Interpretar todos los estadísticos y gráficas para Análisis de elementos - Minitab.” https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/item-analysis/interpret-the-results/all-statistics-and-graphs/ (accessed Jun. 16, 2022).

[104] “sklearn.tree.DecisionTreeClassifier — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (accessed May 10, 2022).

[105] “sklearn.model_selection.GridSearchCV — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed May 10, 2022).

[106] “Impureza de Gini | División de árboles de decisión con impureza de Gini | Datapeaker.” https://datapeaker.com/big-data/impureza-de-gini-division-de-arboles-de-decision-con-impureza-de-gini/ (accessed May 10, 2022).

[107] “Decision Trees: Gini vs Entropy | Quantdare.” https://quantdare.com/decision-trees-gini-vs-entropy/ (accessed May 10, 2022).

[108] “sklearn.model_selection.KFold — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html (accessed May 11, 2022).

[109] "Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping." https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap (accessed May 10, 2022).

[110] "3 Techniques to Avoid Overfitting of Decision Trees | by Satyam Kumar | Towards Data Science." https://towardsdatascience.com/3-techniques-to-avoid-overfitting-of-decision-trees-1e7d3d985a09 (accessed May 10, 2022).

[111] "sklearn.ensemble.RandomForestClassifier — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed May 11, 2022).

[112] "Máquinas de vectores de soporte - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte (accessed Jun. 17, 2022).

[113] "sklearn.ensemble.GradientBoostingClassifier — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (accessed May 11, 2022).

[114] "The Sequential class." https://keras.io/api/models/sequential/ (accessed May 20, 2022).

[115] "Cómo desarrollar modelos de Deep Learning con Keras - ▷ Cursos de Programación de 0 a Experto © Garantizados." https://unipython.com/como-desarrollar-modelos-de-deep-learning-con-keras/ (accessed May 20, 2022).

[116] "Dropout y Batch Normalization." https://vincentblog.xyz/posts/dropout-y-batch-normalization (accessed May 20, 2022).

[117] J. Gabriel and M. Castellanos, "M ´ ETODO DE DETECCI´ONDETECCI´ DETECCI´ON TEMPRANA DE OUTLIERS".

[118] M. Aguilar, T. Bhuket, S. Torres, B. Liu, and R. J. Wong, "Prevalence of the Metabolic Syndrome in the United States, 2003-2012," *JAMA*, vol. 313, no. 19, pp. 1973–1974, May 2015, doi: 10.1001/JAMA.2015.4260.

[119] M. Vaduganathan, J. van Meijgaard, M. R. Mehra, J. Joseph, C. J. O'donnell, and H. J. Warraich, "Trends in the Prevalence of Metabolic Syndrome in the United States, 2011-2016," *JAMA*, vol. 323, no. 24, pp. 2526–2528, Jun. 2020, doi: 10.1001/JAMA.2020.4501.

[120] E. S. Ford, W. H. Giles, and A. H. Mokdad, "Increasing Prevalence of the Metabolic Syndrome Among U.S. Adults," *Diabetes Care*, vol. 27, no. 10, pp. 2444–2449, Oct. 2004, doi: 10.2337/DIACARE.27.10.2444.

[121] "Overweight & Obesity Statistics | NIDDK." https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity (accessed May 06, 2022).

[122] "The State of Obesity 2020: Better Policies for a Healthier America - tfah." https://www.tfah.org/report-details/state-of-obesity-2020/ (accessed May 06, 2022).

[123] X. Liu, X. H. Zhu, P. Qiu, and W. Chen, "A correlation-matrix-based hierarchical clustering method for functional connectivity analysis," *Journal of Neuroscience Methods*, vol. 211, no. 1, pp. 94–102, Oct. 2012, doi: 10.1016/J.JNEUMETH.2012.08.016.

[124] "Métricas De Evaluación De Modelos En El Aprendizaje Automático." https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico (accessed May 10, 2022).

[125] "Inteligencia artificial fácil - Machine Learning y Deep Learning prácticos - Funciones de pérdida (Loss function) | Editiones ENI." https://www.ediciones-eni.com/open/mediabook.aspx?idR=8dd2ca32769cb24b49648b15ef8e777e (accessed May 22, 2022).

[126] "Classification: ROC Curve and AUC | Machine Learning Crash Course | Google Developers." https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es_419 (accessed May 10, 2022).

[127] A. Tiengo, G. P. Fadini, and A. Avogaro, "The metabolic syndrome, diabetes and lung dysfunction," *Diabetes & Metabolism*, vol. 34, no. 5, pp. 447–454, Nov. 2008, doi: 10.1016/J.DIABET.2008.08.001.

[128] "Net Promoter Score: todo lo que necesitas saber para medir la fidelidad de tus clientes - Cuaderno de Marketing." https://cuadernodemarketing.com/net-promoter-score/ (accessed Jun. 17, 2022).

[129] "Net Promoter Score: ¿Mides la satisfacción de tus clientes?" https://digitis.com/net-promoter-score-marketing-nps/ (accessed Jun. 17, 2022).

[130] B. C. Hansen and N. L. Bodkin, "Primary Prevention of Diabetes Mellitus by Prevention of Obesity in Monkeys," *Diabetes*, vol. 42, no. 12, pp. 1809–1814, Dec. 1993, doi: 10.2337/DIAB.42.12.1809.

[131] R. R. Kalyani, C. D. Saudek, F. L. Brancati, and E. Selvin, "Association of Diabetes, Comorbidities, and A1C With Functional Disability in Older AdultsResults from the National Health and Nutrition Examination Survey (NHANES), 1999–2006," *Diabetes Care*, vol. 33, no. 5, pp. 1055–1060, May 2010, doi: 10.2337/DC09-1597.

[132] J. E. Gangwisch, D. Malaspina, B. Boden-Albala, and S. B. Heymsfield, "Inadequate Sleep as a Risk Factor for Obesity: Analyses of the NHANES I," *Sleep*, vol. 28, no. 10, pp. 1289–1296, Oct. 2005, doi: 10.1093/SLEEP/28.10.1289.

[133] C. A. Befort, N. Nazir, and M. G. Perri, "Prevalence of Obesity Among Adults From Rural and Urban Areas of the United States: Findings From NHANES (2005-2008)," *The Journal of Rural Health*, vol. 28, no. 4, pp. 392–397, Sep. 2012, doi: 10.1111/J.1748-0361.2012.00411.X.

[134] G. Block, "Foods contributing to energy intake in the US: data from NHANES III and NHANES 1999–2000," *Journal of Food Composition and Analysis*, vol. 17, no. 3–4, pp. 439–447, Jun. 2004, doi: 10.1016/J.JFCA.2004.02.007.

[135] H. Kim, E. A. Hu, and C. M. Rebholz, "Ultra-processed food intake and mortality in the USA: results from the Third National Health and Nutrition Examination Survey (NHANES III, 1988–1994)," *Public Health Nutrition*, vol. 22, no. 10, pp. 1777–1785, Jul. 2019, doi: 10.1017/S1368980018003890.

[136] A. Mozumdar and G. Liguori, "Persistent Increase of Prevalence of Metabolic Syndrome Among U.S. Adults: NHANES III to NHANES 1999–2006," *Diabetes Care*, vol. 34, no. 1, pp. 216–219, Jan. 2011, doi: 10.2337/DC10-0879.

[137] "Estudio del conjunto de datos NHANES mediante el empleo de técnicas de aprendizaje no supervisado."

[138] E. G. García *et al.*, “LOS PROBLEMAS DE LOS GRUPOS VULNERABLES La obesidad y el síndrome metabólico como problema de salud pública. Una reflexión. Segunda parte*,” *Salud Mental*, vol. 32, no. 79, pp. 79–87, 2009.

[139] “La obesidad se lleva el 7% del gasto sanitario de España.” https://isanidad.com/156466/obesidad-se-lleva-7-por-ciento-gasto-sanitario-espana/ (accessed May 24, 2022).

[140] “Intelligent Techniques Introduction”.

ANEXO 1: SOBREAJUSTE DE LOS MODELOS

En este apartado se detalla el análisis llevado a cabo para confirmar que los modelos no están sobreajustados.

ANEXO 1.1: SOBREAJUSTE DEL DECISION TREE

Para comprobar que no haya sobreajuste en este modelo, podemos apreciar que a partir de max_depth = 7 en la Figura 66 el error de test se mantiene constante, este valor coincide con la optimización del *GridSearchCV*. El enlace más débil se caracteriza por un alpha efectivo, donde los nodos con el menor alpha se podan antes, se muestra la curva ROC sobre alpha en la Figura 65.

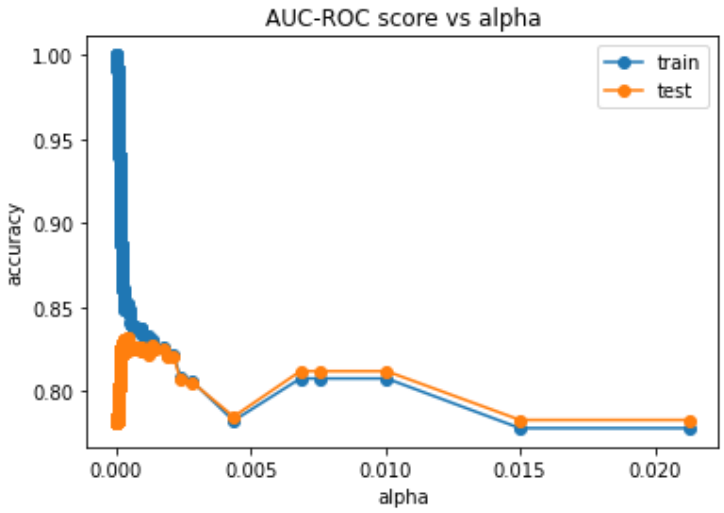

Figura 65: AUC-ROC vs alpha

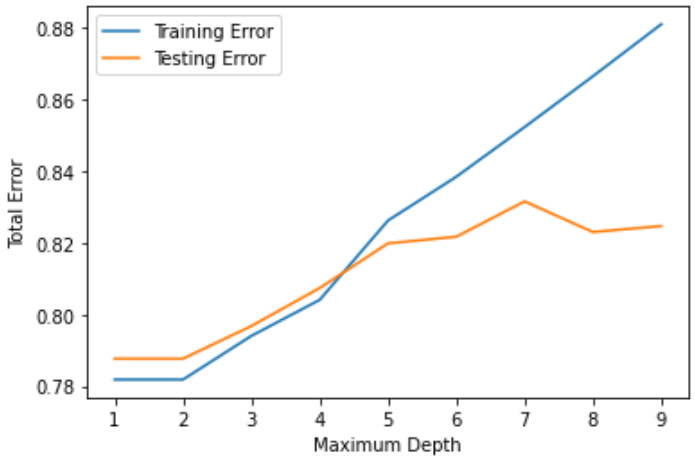

Figura 66: Máxima profundidad del árbol

ANEXO 1.2: SOBREAJUSTE DEL RANDOM FOREST

Para comprobar que no haya sobreajuste en este modelo, podemos apreciar en la Figura 67, que a partir de max_depth = 9 el error de test se mantiene aproximadamente constante, este valor coincide con el valor fijado.

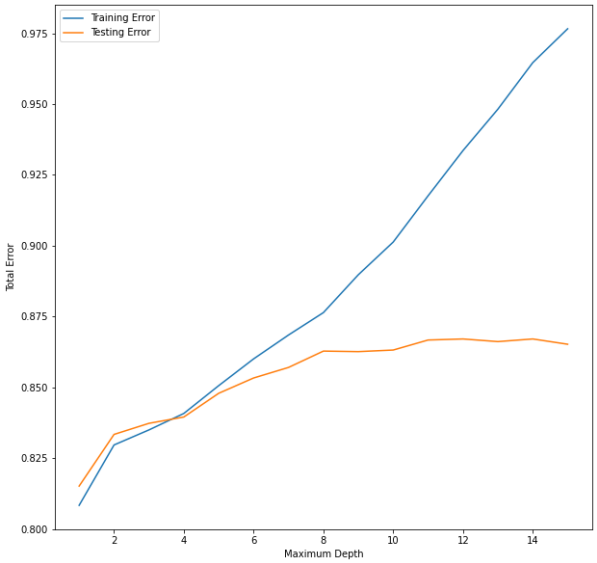

Figura 67: Máxima profundidad de los árboles en el Random Forest

ANEXO 1.3: SOBREAJUSTE DEL SVC

Con objetivo de comprobar que no existe sobreajuste en el modelo se compara la precisión de los datos del entrenamiento, 0.88, con la de los datos de test, 0.86 (se detalla con más profundidad en el apartado 4.4.3 Evaluación del modelo SVC). Están muy próximas por lo tanto podemos concluir que no existe *overfitting*.

ANEXO 1.4: SOBREAJUSTE DEL NEURAL NETWORK

Para comprobar que no haya sobreajuste en este modelo, se obtiene la gráfica de las pérdidas conforme aumenta los ciclos de entrenamiento. Se observa en la Figura 68 que las pérdidas de entrenamiento y las de validación descienden de forma bastante pareja. Por tanto, se puede considerar que el modelo no presenta sobreajuste.

Figura 68: ***Loss*** **vs** ***epochs***

ANEXO A: ASPECTOS ÉTICOS, ECONÓMICOS, SOCIALES Y AMBIENTALES

A.1 INTRODUCCIÓN

El análisis de datos y la Inteligencia artificial son técnicas que están demostrando tener una gran efectividad en el sector sanitario para predecir y prevenir enfermedades. Por otro lado, el ambiente obesogénico y los hábitos de vida cada vez menos saludables de la población, están provocando una epidemia de enfermedades metabólicas que merman sobre manera la calidad de vida de los que las padecen. Estas enfermedades son de los principales factores de riesgo para sufrir diabetes, enfermedad cardiovascular y otras patologías que representan las principales causas de muerte en los países desarrollados. Cuanto antes se diagnostique el Síndrome Metabólico en un individuo, antes puede realizarse un abordaje médico para revertir la enfermedad.

A.2 DESCRIPCIÓN DE IMPACTOS RELEVANTES RELACIONADOS CON EL PROYECTO

- **Impacto social:** Este Trabajo de Fin de Máster tiene un impacto social directo debido a que el Síndrome Metabólico es una patología muy extendida en los países del primer mundo. Un diagnóstico temprano incrementa las posibilidades del paciente de no agravar su enfermedad o de que esta no acabe desembocando en problemas de salud más graves. Por otro lado, uno de los objetivos que se pretenden conseguir es fomentar que los sistemas de salud sean más eficientes al colaborar con los profesionales de la medicina a la hora de tener que evaluar la situación de un paciente. Contribuir en la mejora del sistema de salud es algo que beneficia al conjunto de la sociedad, tal y como ha quedado presente al sucederse los acontecimientos de la última pandemia de COVID-19. Por otro lado, la concienciación de prevenir el Síndrome Metabólico pasa por una reestructuración cultural importante, en la que se deben sustituir ciertos hábitos de vida por otros más saludables. Para ello es fundamental identificar esta serie de costumbres nocivas para la salud de las personas. No obstante, se debe recalcar que algo que impacta en el bienestar de los sujetos es el reparto de riqueza, la pobreza es un factor de riesgo para padecer múltiples enfermedades, desde las relacionadas con el sobrepeso hasta las que guardan relación con la salud mental.
- **Impacto económico:** La obesidad y el Síndrome Metabólico son problemas de salud pública que directa o indirectamente suponen un coste relevante del presupuesto anual destinado a salud [138]. Presupuesto que con el paso de los años sufre un aumento debido al incremento de prevalencia de enfermedades relacionadas con el sobrepeso y la obesidad [139]. La prevención más efectiva para esta clase de enfermedades pasa por un cambio en los hábitos de vida de la persona, lo que supone un gasto cero para el Estado [11]. Además del dinero de los contribuyentes, para la economía individual también supone un ahorro el hecho de prevenir el Síndrome Metabólico, en lugar de enfrentar el coste que supone su tratamiento una vez se vuelve una enfermedad grave. Al mismo tiempo, una sociedad saludable es una sociedad más feliz, esto repercute de forma directa en la productividad de los individuos que la conforman.

- **Impacto ético:** En el desarrollo de este TFM no se han incluido muestras identificables ni identidad de los pacientes que han formado parte del estudio. Todos los datos recopilados son anónimos. Además, todos los recursos empleados son públicos y de libre acceso. Por lo tanto, las consideraciones éticas no aplican en el presente trabajo.
- **Impacto medioambiental:** El impacto medioambiental que se detecta en este proyecto se deriva de la utilización de la electricidad del ordenador portátil con el que se ha llevado a cabo. Lo cual supone un impacto mínimo.

A.3 CONCLUSIONES

En conclusión, la implementación de técnicas de inteligencia artificial aplicadas a la detección de enfermedades, entre ellas el Síndrome Metabólico, puede suponer un aumento en la calidad de vida de muchos sujetos y en un ahorro relevante para el sistema sanitario.

Los impactos más relevantes detectados son el económico y el social. Debido a que el desarrollo de tecnología para el diagnóstico precoz de enfermedades es beneficioso, no solo para la salud a nivel individual, también para los medios del sistema sanitario. Además, puede implicar que se destinen recursos en diferentes cuestiones que contribuyan a fomentar la evolución de la sociedad en su conjunto.

Es fundamental continuar haciendo hincapié en que el papel de los hábitos de vida saludable juega un papel necesario en la salud global e intentar establecerlos como costumbres culturales. De esta forma se puede obtener una sociedad más feliz, más saludable y más productiva.

ANEXO B: PRESUPUESTO ECONÓMICO

Costes de mano de obra (coste directo)		
Horas	Precio/hora	Total
750	15	11.250€

Costes de recursos materiales (coste directo)			
Precio de compra	Uso en meses	Amortización en años	Total
1.000€	4	5	66,66€

Material Fungible	
Impresión	100€
Encuadernación	300€
Total	400€

Costes totales	
Costes de mano de obra	11.250€
Costes de recursos materiales	66,66€
Costes generales (15% de los costes directos) (coste indirecto)	1.697,50 €
Beneficio industrial (6% de costes directos e indirectos)	780,65€
Material fungible	400,00€
Subtotal presupuesto	14.195€
IVA (21%)	2.980,91€
PRESUPUESTO TOTAL	**17.175,72€**