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:** ……………….

**Autor:** D…………………….

**Tutor:** D. ……………….

**Ponente:** D. ………………

**Departamento:** …………

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 anormalidades metabólicas 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.

En primer lugar, se realizará un estudio sobre el estado del arte de las tecnologías existentes para el desarrollo de investigaciones de inteligencia artificial predictiva. A continuación, se procederá 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 efectuará 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 emplearán y optimizarán distintos modelos de Aprendizaje Automático que posibiliten la predicción de padecer Síndrome Metabólico.

En último lugar, se detallarán las conclusiones obtenidas y se planteará 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 is made up 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 gradually become more 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 in order to predict whether a person is susceptible to suffer from Metabolic Syndrome.

First, a study will be made 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 will be broken down, identifying those that could be of most interest for this study. At the same time, the data will be cleaned and preprocessed in order to carry out a statistical analysis of the chosen variables. Later, different Machine Learning models will be used and optimized to enable the prediction of suffering from Metabolic Syndrome.

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, python, jupyter, análisis de datos, correlación.*

KEYWORDS

*Artificial intelligence, metabolic syndrome, neural network, machine learning, predictive modelling, python, jupyter, data analytics, correlation.*

indice del contenido

[1. introducción y objetivos 1](#_Toc102296347)

[1.1. ANTECEDENTES: SÍNDROME METABÓLICO 1](#_Toc102296348)

[1.1.1. SELECCIÓN DE CRITERIOS 2](#_Toc102296349)

[1.2. Problemática 2](#_Toc102296350)

[1.3. Objetivos 3](#_Toc102296351)

[2. ESTADO DEL ARTE 4](#_Toc102296352)

[2.1. Estado del arte del análisis de datos 4](#_Toc102296353)

[2.1.1 Inteligencia artificial 5](#_Toc102296354)

[2.2. Aplicación del análisis de datos en el sector sanitario 7](#_Toc102296355)

[2.2.1 Support Vector Machine 8](#_Toc102296356)

[2.2.2 Decision Tree 8](#_Toc102296357)

[2.2.3 Random Forest 9](#_Toc102296358)

[2.2.4 Neural network 10](#_Toc102296359)

[2.3. Estado del arte del síndrome metabólico 11](#_Toc102296360)

[3. materiales y métodos 12](#_Toc102296361)

[3.1. Base de datos NHANES 12](#_Toc102296362)

[3.1.1 Extracción del dataframe 12](#_Toc102296363)

[3.2 Herramientas empleadas 16](#_Toc102296364)

[3.3 Metodología 16](#_Toc102296365)

[3.3.1 Limpieza y preprocesado de los datos 17](#_Toc102296366)

[3.3.2 Análisis estadístico de los datos 26](#_Toc102296367)

[Conclusiones 38](#_Toc102296368)

[Líneas futuras 38](#_Toc102296369)

[REFERENCIAS Error! Bookmark not defined.](#_Toc102296370)

[anexo a: aspectos ÉTICOS, económicos, sociales y ambientales 44](#_Toc102296371)

[A.1 Introducción 44](#_Toc102296372)

[A.2 DESCRIPCIÓN DE IMPACTOS RELEVANTES RELACIONADOS CON EL PROYECTO 45](#_Toc102296373)

[A.3 ANÁLISIS DETALLADO DE ALGUNO DE LOS PRINCIPALES IMPACTOS 45](#_Toc102296374)

[A.4 CONCLUSIONES 45](#_Toc102296375)

[anexo b: presupuesto económico 46](#_Toc102296376)

siglas y acrónimos

* **SM:** Síndrome metabólico
* **ML:** Machine Learning
* **IA:** Inteligencia Artificial
* **SVM:** Support vector machine
* **DL:** Deep learning
* **DT:** Decision tree
* **RF:** Random forest
* **NN:** Neural network
* **NHANES:** The National Health and Nutrition Examination Survey
* **BBDD:** Base de datos

# introducción y objetivos

## 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. En España, Gregorio Marañón ya resumió las evidencias sobre la asociación entre hipertensión, alteraciones del metabolismo glucídico y la obesidad. Propuso que la primera línea de prevención debía ser una dieta adecuada y suficiente combinada con ejercicio físico. [2]

Table

Description automatically generatedNo 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 acuña el término de Síndrome Metabólico como entidad diagnóstica con criterios definidos para referirse a estas alteraciones[3]. Aunque la definición más utilizada es la propuesta por el ATP III [4].

Tabla 1: Comparativa de criterios del SM

### SELECCIÓN DE CRITERIOS

A día de hoy 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 [5]. 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.

## Problemática

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 [6].

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) con la ventaja añadida de que se encuentran siempre disponibles, sin necesidad de realizar prácticamente ningún esfuerzo.

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. 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.) [7]puntualizó en su momento que la muerte súbita es más común entre los que son naturalmente gordos.

La prevalencia global de Síndrome Metabólico según criterios de la Organización Mundial de la Salud es del 36,8% [8]. 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. Se estima que España suma cada día 257 personas con SM, hasta los 94000 casos por año[9].

Chart, bar chart

Description automatically generated

Ilustración 1: Prevalencia SM según edad e IMC

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 son avisados mediante mecanismos endógenos obvios, como por ejemplo el dolor. A pesar de esto, es una patología muy fácil de prevenir mediante alimentación adecuada, ejercicio físico, control del estrés, un descanso suficiente y evitar 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 [10].

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.

Por otro lado, de los estudios 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 [11].

La identificación de la obesidad como problema y desafío de las próximas décadas está empezando a reconocerse, tanto entre la sociedad como políticamente. España se encuentra en el comienzo del desarrollo de políticas para disminuir los efectos de la epidemia de obesidad. 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 [12].

## Objetivos

Este trabajo pretende identificar la probabilidad de que un individuo padezca SM en base a técnicas de Inteligencia Artificial.

La finalidad es implementar una serie de modelos predictivos basados en aprendizaje automático (ML) 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:

* 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.
* Obtención de la prevalencia de SM en la muestra recogida por la base de datos.
* Selección de las columnas de interés de la base de datos para inyectarlas en los modelos predictivos.
* Análisis, limpieza y preprocesado de los campos seleccionados de la base de datos para establecer su relación con el SM
* Estudio, selección, implementación y optimización de los diferentes algoritmos de aprendizaje automático para resolver el problema de predicción del SM.

# ESTADO DEL ARTE

En este apartado se procede a hacer un breve repaso de las bases teóricas del TFM. Se enfocará 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.

## 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 mercadotecnica, las ciencias sociales y puras o la medicina. Se podría establecer una primera división entre los tipos de análisis de datos [13] [14] .

* **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 númerica 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. [15]

* **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.

Timeline

Description automatically generated

Ilustración 2: Métodos de análisis de datos

### 2.1.1 Inteligencia artificial

Timeline

Description automatically generatedEl 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 [16] .

Ilustración 3: Historia de la IA

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 [17].

Diagram

Description automatically generatedEl machine learning es una rama de la inteligencia artificial que combina grandes vólumenes 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.[18]

Ilustración 4: Flujo en machine learning

Ilustración 5: Campos de aprendizaje en IA

Diagram

Description automatically generated

A picture containing calendar

Description automatically generatedEl deep learning 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 y requiere tiempos de aprendizaje más largos que su predecesor. [19], [20]

Ilustración 6: niveles del aprendizaje automático

## 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 [21] . 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.[22]

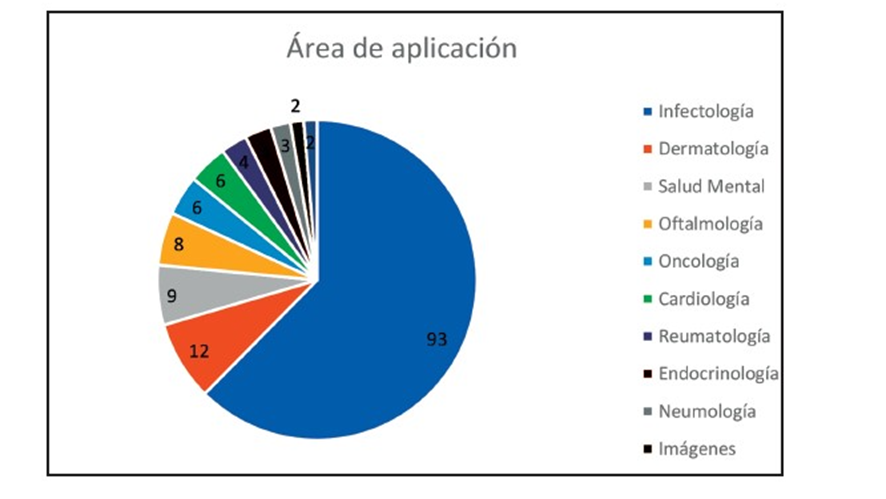
La implementación del análisis de datos en salud da lugar a diversas aplicaciones tales como: estudio de la transmisión de enfermedades infecciosas, seguridad farmacológica, gestión de recursos humanos y suministros, vigilancia de la salud, diagnóstico automatizado por imagen o predicción de factores de riesgo [23].

Ilustración 7: área de aplicación del análisis de datos en salud

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 [24] .

Actualmente, la utilización de modelos predictivos en enfermedades como la diabetes [25] 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[26]. A continuación se realiza un sucinto resumen sobre estos modelos:

### Support Vector Machine

Las máquinas de vectores de soporte (o support vector machines en su nombre original en inglés) [27] [28]son un conjunto de métodos supervisados de aprendizaje automático dedicados a tareas de regresión y clasificación. 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 cada uno de ellos pertenece a una categorías, un algoritmo basado en SVM construye un modelo capaz de predecir a qué categoría pertenece un punto nuevo. 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.

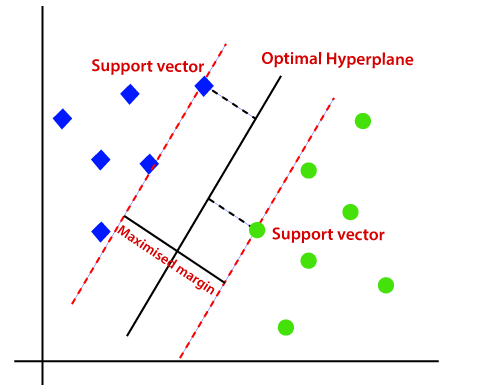
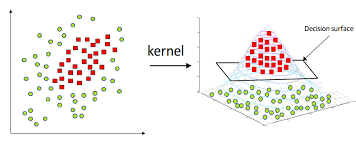


Ilustración 8: Función kernel en SVM

Ilustración 9: Gráfico de clasificación con SVM



### Decision Tree

Los árboles de decisión (o decision trees en su nombre original en inglés) son 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. Un árbol puede verse como una aproximación constante a trozos [29]

* 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)
* 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.

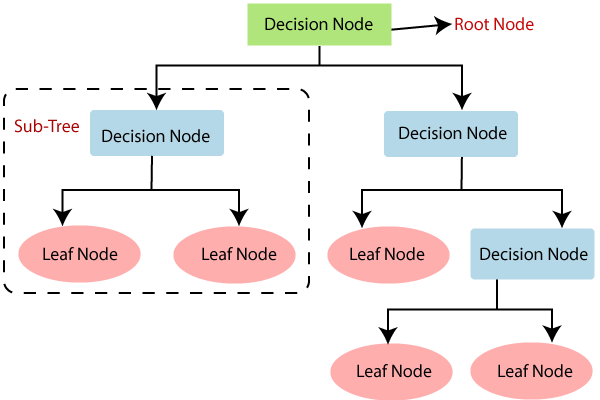


Ilustración 10: Árbol de decisión esquematizado

### Random Forest

El bosque aleatorio (random forest en su nombre original en inglés) [30], [31] 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. Un random forest es un conjunto de árboles de decisión combinados con bagging . 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.

Para problemas de [clasificación](https://www.iartificial.net/clasificacion-o-regresion/#Clasificacion), se suelen combinar los resultados de los árboles de decisión usando soft-voting (voto suave). En el voto suave, se le da más importancia a los resultados en los que los árboles estén muy seguros.

Para problemas de [regresión](https://www.iartificial.net/clasificacion-o-regresion/#Regresion), la forma más habitual de combinar los resultados de los árboles de decisión, es tomando su media aritmética.

Diagram

Description automatically generated

Ilustración 11: esquema random forest

### Neural network

Las redes neuronales artificiales se engloban dentro del deep learning 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. 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. [32], [33]

El alcance de las funciones de las redes neuronales es muy amplio, 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, 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, o en medicina 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, los precios dinámicos de Amazon, la identificación de riesgos en banca o en la personalización de estrategias de marketing. [34]

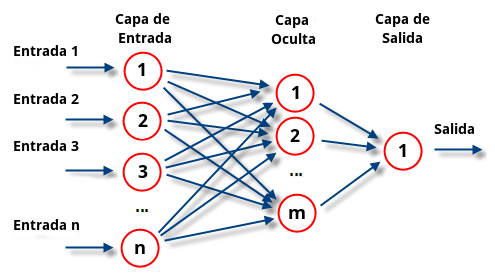


Ilustración 12: funcionamiento esquemático en una red neuronal

## 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). Su incidencia va en aumento. 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 ha puesto 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 enfemedades cardiovasculares. [35]

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 insula, que es difícil de medir en la práctica diaria, pero aceptó evidencias indirectas, como la alteración de la glucosa en ayunas.[36]

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[37]. En un estudio español [38] se comparó la prevalencia del SM en la misma población utilizando tres de los criterios *( Tabla 1: Comparativa de criterios del SM ).* Aquí se encontró que se epresentaba con más frecuencia en hombres que en mujeres y aumentaba con la edad, sin importar el criterio utilizado.

Cabe destacar el artículo [39] 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 periodo 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 esta patología se incrementa con la edad y se aprecia un aumento sustancial en enfermedades asociadas.

Existen varias investigaciones que utilizan diferentes algoritmos de machine learning para aplicarlos en el síndrome metabólico [40] . En Taiwan, en el 2018, nos encontramos con un paper [41] 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 Iran en 2016, emplea árboles de decisión y SVM (support vector machine) para predecir la incidencia en 7 años del SM[42]. Otra investigación de 2018 [43] propone modelos de ML en la predicción de la prevalencia del síndrome metabólico en población no obesa. Finalmente, en una revisión sistemática [44] 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.

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. 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 en relación al síndrome metabólico. 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 [45]

# materiales y métodos

## Base de datos NHANES

La base de datos National Health and Nutrition Examination Survey (NHANES) [46]lleva operativa desde el 1959 recopilando, analizando y difundiendo datos sobre el estado de salud de los residentes de los Estados Unidos. Cada año cerca de 5.000 individuos participan en sus encuestas (cada año 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 medico 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.

Por el tipo de información que ofrece NHANES, ha contribuido especialmente en el estudio de las siguientes cuestiones: Anemia, enfermedades cardiovasculares, diabetes, exposiciones ambientales, enfermedades oculares, pérdidas de audición, enfermedades infecciosas, enfermedades renales, nutrición, obesidad, salud bucodental, osteoporosis, condición física, historial reproductivo, enfermedades respiratorias y enfermedades de transmisión sexual.

### 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 un merge de todas las columnas por número de secuencia. [47]–[50]

Para llevar a cabo el desarrollo del TFM 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.

Tabla 2: 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 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 (en gris), de examinación (en naranja), de laboratorio (en rosa) y resultados del cuestionario (verde).

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.

Tabla 3: descripción de las características extraídas de la BBDD NHANES

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| nº | Column | Type | Long | Description | Values | Group |
| 1 | SEQN | Num | 8 | Sequence number | **62161 - 102956** | Demographics |
| 2 | RIAGENDR | Num | 8 | Gender | **1**: Male **2**: Female | Demographics |
| 3 | RIDAGEYR | Num | 8 | Age in years at screening | **0 - 80, .**:Missing mean: 41.76 std: 22.55 | Demographics |
| 4 | DMDEDUC2 | Num | 8 | Education level  (adults +20) | **1**: Less than 9th grade, **2** : 9-11th grade  (Includes 12th grade with no diploma),  **3**: High school graduate/GED or equivalent,  **4**: Some college or AA degree,  **5**: College graduate or above,  **7** : Refused, **9**: Don't Know **.** : Missing | Demographics |
| 5 | RIDRETH3 | Num | 8 | Race/Hispanic  origin w/ NH Asian | **1**: Mexican American, **2**: Other Hispanic  **3**: Non-Hispanic White,  **4** : Non-Hispanic Black  **6**: Non-Hispanic Asian,  **7:** Other Race - Including Multi-Racial | Demographics |
| 6 | BPXSY1 | Num | 8 | Systolic: Blood pres  (1st rdg) mm Hg | **66 - 238, .**:Missing mean: 122.56 std: 18.78 | Examination |
| 7 | BPXDI1 | Num | 8 | Diastolic: Blood pres  (1st rdg) mm Hg | **0 - 136**, .:Missing mean: 69, std: 13.6 | Examination |
| 8 | BMXWT | Num | 8 | Weight (kg) | **3.6 - 242**, .:Missing mean: 74.38, std: 27.75 | Examination |
| 9 | BMXBMI | Num | 8 | Body mass index  (kg/m\*\*2) | **11.5 - 86.2**, .:Missing mean: 27.95, std: 7.59 | Examination |
| 10 | BMXWAIST | Num | 8 | Waist circumference (cm) | **40-177.9**, .:Missing mean: 94.8, std: 19.76 | Examination |
| 11 | BMXHT | Num | 8 | Sanding height (cm) | **80.7 - 204.5**, .:Missing mean: 162.98, std: 16.62 | Examination |
| 12 | LBXTR | Num | 8 | Triglyceride (mg/dL) | **10 - 4233**, .:Missing mean: 114.94, std: 102.77 | Laboratory |
| 13 | LBXIN | Num | 8 | Insulin (uU/mL) | **0.14 - 682.48**, .:Missing mean: 14.11, std: 19.43 | Laboratory |
| 14 | LBXGLU | Num | 8 | Fasting glucose (mg/dL) | **21 - 479**, .:Missing mean: 109.35, std: 36 | Laboratory |
| 15 | LBDHDD | Num | 8 | Direct HDL-Cholesterol  (mg/dL) | **6 - 226**, .:Missing mean: 53.25, std: 15.73 | Laboratory |
| 16 | BPQ040A | Num | 8 | Taking prescription  for hypertension | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 17 | BPQ030 | Num | 8 | Told had high  blood pressure - 2+ times | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 18 | BPQ050A | Num | 8 | Taking prescription  for hypertension | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 19 | BPQ090D | Num | 8 | Told to take prescription  for cholesterol | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 20 | DIQ010 | Num | 8 | Doctor told you  have diabetes | **1**: Yes, **2**: No, **3:** Borderline  **7**: Refused , **9**: Don't know,  **.** :Missing | Questionnaire |
| 21 | DIQ070 | Num | 8 | Take diabetic pills  to lower blood sugar | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 22 | ALQ130 | Num | 8 | Avg # alcoholic  drinks/day - past 12 mos | **1 to 13**: Range of values,  **15**: 15 drinks or more  **777**: Refused , **999**: Don't know,  **.** :Missing | Questionnaire |
| 23 | HSD010 | Num | 8 | General health condition | **1**: Excellent, **2**: Very good, **3:** Good **4:** Fair, **5**: Poor **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 24 | DBQ700 | Num | 8 | How healthy is the diet | **1**: Excellent, **2**: Very good, **3:** Good **4:** Fair, **5**: Poor **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 25 | HIQ011 | Num | 8 | Covered by health  insurance | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 26 | MCQ080 | Num | 8 | Doctor ever said you were overweight | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 27 | MCQ010 | Num | 8 | Ever been told you have asthma | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 28 | MCQ220 | Num | 8 | Ever told you had  cancer? | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 29 | MCQ300C | Num | 8 | Close relative had diabetes | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 30 | DPQ020 | Num | 8 | Feeling down, depressed or hopeless | **0:** Not at all, **1**: Several days,  **2**: More than half the days,  **3:** Nearly every day,  **7**: Refused , **9**: Don't know,  **.** :Missing | Questionnaire |
| 31 | DPQ030 | Num | 8 | Trouble sleeping or sleeping too much | **0:** Not at all, **1**: Several days,  **2**: More than half the days,  **3:** Nearly every day,  **7**: Refused , **9**: Don't know,  **.** :Missing | Questionnaire |
| 32 | DPQ040 | Num | 8 | Feeling tired or having little energy | **0:** Not at all, **1**: Several days,  **2**: More than half the days,  **3:** Nearly every day,  **7**: Refused , **9**: Don't know,  **.** :Missing | Questionnaire |
| 33 | DPQ050 | Num | 8 | Poor appetite  or oveating | **0:** Not at all, **1**: Several days,  **2**: More than half the days,  **3:** Nearly every day,  **7**: Refused , **9**: Don't know,  **.** :Missing | Questionnaire |
| 34 | PAQ605 | Num | 8 | Vigorous work act | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 35 | PAQ620 | Num | 8 | Moderate work act | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 36 | PAQ635 | Num | 8 | Walk or bicycle | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 37 | PAQ650 | Num | 8 | Vigorous recreational activities | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 38 | PAQ665 | Num | 8 | Moderate recreational activities | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 39 | SMQ020 | Num | 8 | Smoked at least 100 cigarettes in life | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 40 | INQ020 | Num | 8 | Income from wages/salaries | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |
| 41 | INDFMMPI | Num | 8 | Family monthly poverty level index | **0 to 4.98**: range of values, **5**: greater or equal to 5,  **.** :Missing | Questionnaire |
| 42 | INDFMMPC | Num | 8 | Family monthly poverty level category | **1**: Monthly poverty level index <=1.30,  **2**: 1.30<=Monthly poverty level index <=1.85,  **3:** Monthly poverty index >1.85  **7**: Refused , **9**: Don't know, **.** :Missing | Questionnaire |
| 43 | SLQ050 | Num | 8 | Ever told doctor had trouble sleeping? | **1**: Yes, **2**: No, **7**: Refused ,  **9**: Don't know, **.** :Missing | Questionnaire |

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 punto *1.1.1* se ha añadido una columna inicial 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 *Tabla 4: requisitos del dataframe para SM:*

Tabla 4: requisitos del dataframe 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 (female) BMXWAIST => 102cm si RIDAGENDR = 1 (male) |
| 2 | Triglicéridos mayor o igual a 150mg/dL o Tomar medicamentos para el colesterol elevado | LBXTR => 150|| BPQ090D = 1 (yes) |
| 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 if RIDAGENDR = 2 (female) LBDHDD <= 40mg/dL if RIDAGENDR = 1 (male) || 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 (yes) |
| 5 | Glucosa en ayunas mayor o igual a 100mg/dL o Tomar medicamentos para la glucosa alta | LBXGLU >= 100mg/dL || DIQ070 = 1 (yes) |

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 evidencia [51] de que podría suponer un marcador incluso más fiable que el IMC a la hora de valorar el grado de salud de un paciente. 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 se muestra la categoría a la que pertenecería un paciente en función del rango en el que se encontrara este indicador.

Tabla 5: 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 |

## Herramientas empleadas

Las herramientas que se han empleado para llevar a cabo el desarrollo de este TFM se comentan a continuación:

* **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 para aprendizaje automático [52] .
  + **Numpy:** biblioteca de Python que ofrece funciones precompiladas para rutinas numéricas [53]
  + **Pandas:** biblioteca de Python que permite manipular, analizar y visualizar datos en grandes colecciones [54].
  + **Matplotlip:** biblioteca de Python para exportar gráficos y otras imágenes a formatos vectoriales. Muy útil para la visualización de gráficas [55] .
  + **Scipy:** biblioteca de Python con funciones algebraicas y estadísticas [56].
  + **Scikit-learn:** biblioteca de Python de software libre para aprendizaje automático. Integra una gran cantidad de algoritmos de ML además de mucha documentación [57].
  + **Imbalanced-learn:** librería 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 [58].
* **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 [59].
* **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 [60].

## Metodología

Las etapas del desarrollo del presente trabajo son las mencionadas a continuación:

1. Descripción y entendimiento de los parámetros de la base de datos NHANES
2. Extracción de columnas de la base de datos NHANES
3. Limpieza y preprocesado de los datos
4. Estudio de la prevalencia del síndrome metabólico entre los participantes
5. Análisis estadístico y de correlación entre las variables
6. Normalización del conjunto de datos y análisis de componentes principales
7. Obtención del conjunto de entrenamiento y del conjunto de test
8. Inyección en los modelos de ML
9. Ajuste de hiperparámetros y optimización de los modelos
10. Evaluación y comparativa de las métricas
11. Estudio de los resultados

### 3.3.1 Limpieza y preprocesado de los datos

El dataset, 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. A continuación se muestran las columnas con el número total de valores nulos.

Tabla 6: valores nulos de cada columna del dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **nº** | **Column** | **Null values** | **nº** | **Column** | **Null values** |
| 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 |  |  |  |

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.

Agrupación por edad

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

1. **1**: De 1 a 5 años
2. **2**: De 5 a 12 años
3. **3**: De 12 a 20 años
4. **4**: De 20 a 35 años
5. **5**: De 35 a 50 años
6. **6**: De 50 a 65 años
7. **7**: De 65 a 80 años

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 3.

**RIDRETH3**: Se realiza una codificación tipo ‘One hot encoding’, se crean seis nuevos atributos binarios correspondientes a las diferentes etnias representadas.

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 esta caso se puede realizar el cálculo directamente al haber truncado los valores de la altura y el peso. (peso entre altura al cuadrado )

Limpieza de variables NULAS de laboratorio

* **LBXTR:** Los niveles nulos de trigliceridos 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 reemplanzan 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.

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 el individuo toma medicación para la presión sanguínea elevada. 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 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

1. **1:** Monthly poverty level index <= 1.30
2. **2:** 1.3**<** Monthly poverty level index <= 1.85
3. **3:** Monthly poverty level index > 1.85

* **ALQ130:** Indica la media de bebidas alochólicas que ha consumido un paciente en el último año. En este caso se establecen grupos.

1. **0**: Si el paciente no ha consumido bebidas alcoholicas
2. **1**: Si el paciente ha consumido de 1 a 3
3. **2**: Si el paciente ha consumido de 4 a 6
4. **3**: Si el paciente ha consumido de 7 a 10
5. **4**: Si el paciente ha consumido de 11 a 14
6. **5**: Si el paciente ha consumido 15 bebidas alcohólicas o más

Imputación DE OUTLIERS

Los modelos son sensibles a la presencia de outliers. Cuando los datos no cumplen con estos supuestos disminuye la capacidad de detectar efectos reales [61].

Para las variables continuas de laboratorio y mediciones físicas: LBXIN, LBXGLU, LBTR, LBDHDD, BPXSY1, BPXDI1 y BXBMI (que además guardan relación directa con los requisitos del SM) se realiza un estudio para ver en qué casos merece la pena imputar los outliers. Debido a que hay que valorar en qué situación un valor atípico está relacionado con un estado patológico y cuando es una medida errónea. Esto útlimo puede suceder porque se ha procesado mal el dato o 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.

Chart, box and whisker chart

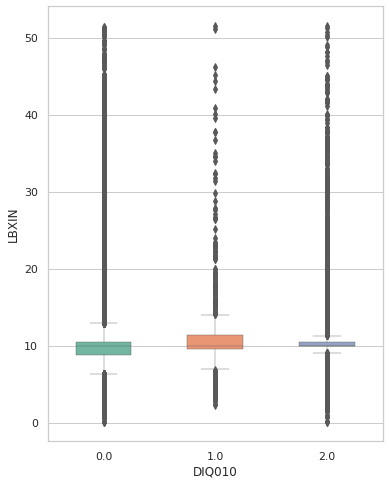
Description automatically generatedA continuación, se muestra la reducción de dispersión de los valores de insulina y glucosa tras haber imputado los valores atípicos.

Figura 1 : Boxplot de LBXIN

Figura 2: Boxplot de LBXIN tras borrado de outliers

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 es borderline 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.

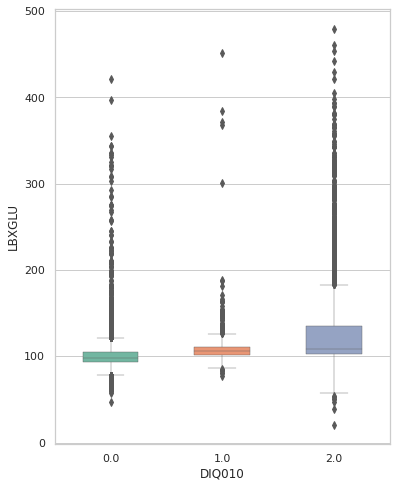
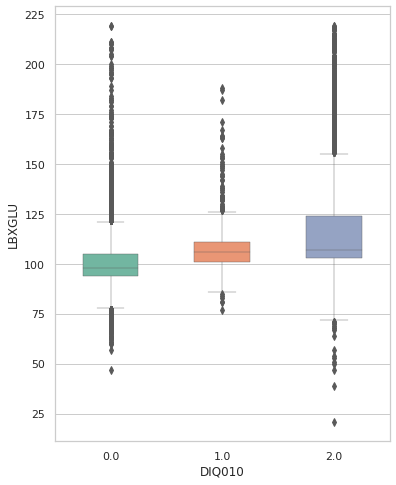


Figura 3: Boxplot de LBXGLU

Figura 4: Boxplot de LBXGLU tras borrado de outliers

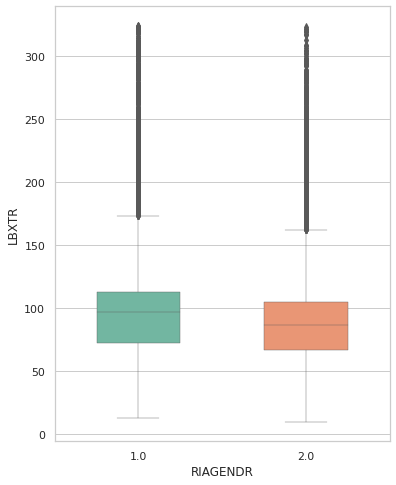
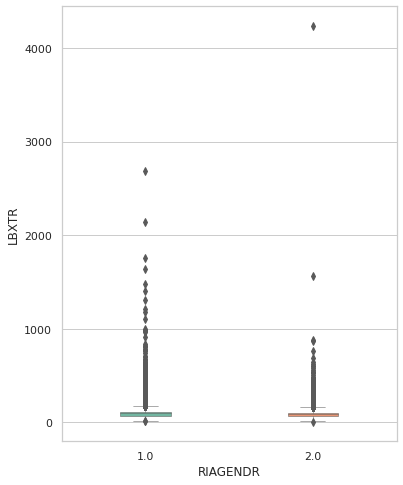
Asimismo, se realiza el mismo método para los niveles de triglicéridos y de HDL.

Figura 5: Boxplot de LBXTR

Figura 6: Boxplot de LBXTR tras borrado de outliers

Son parámetros muy influidos por las hormonas, por lo que pueden diferir entre hombres (RIAGENDR = 1) y mujeres (RIAGENDR = 2). Podemos ver la normalización de los datos tras la imputación de los outliers.

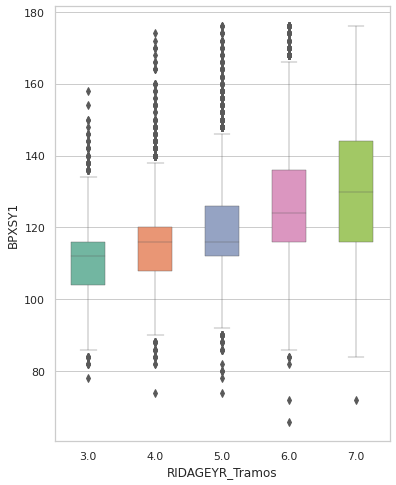
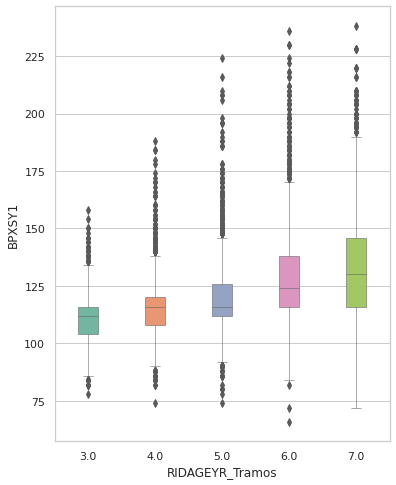
 Para la presión sistólica y diastólica, se observan valores que se podrían considerar desmesurados, por lo tanto se imputan también valores atípicos en este caso. Se puede apreciar a continuación la diferenciación por grupo de edad antes y después de la limpieza de valores atípicos.

Figura 7: Boxplot de la presión sistólica

Figura 8: Boxplot de la presión sistólica tras borrado de outliers

Chart, box and whisker chart

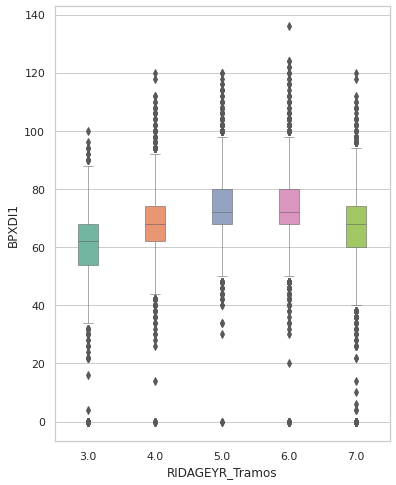
Description automatically generated

Figura 9: Boxplot de la presión sistólica tras borrado de outliers

Figura 10: Boxplot de la presión sistólica

Chart, box and whisker chart

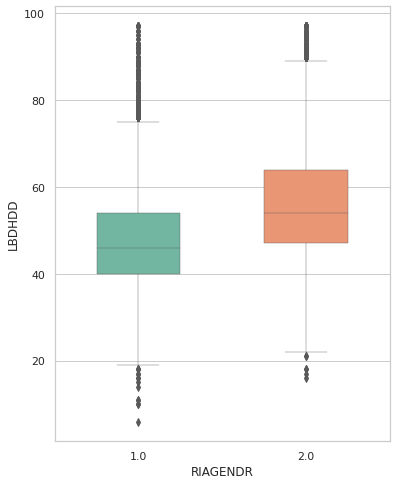
Description automatically generated

Figura 11: Boxplot del colesterol HDL

Figura 12: Boxplot de colesterol HDL tras borrado de outliers

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 outliers en este caso.

Tras la imputación de valores atípicos contamos con **283** registrosmenos. A continuación se presenta el diagrama de dispersión antes y despues del borrado de outliers para poder llevar a cabo una comparación visual.

Table

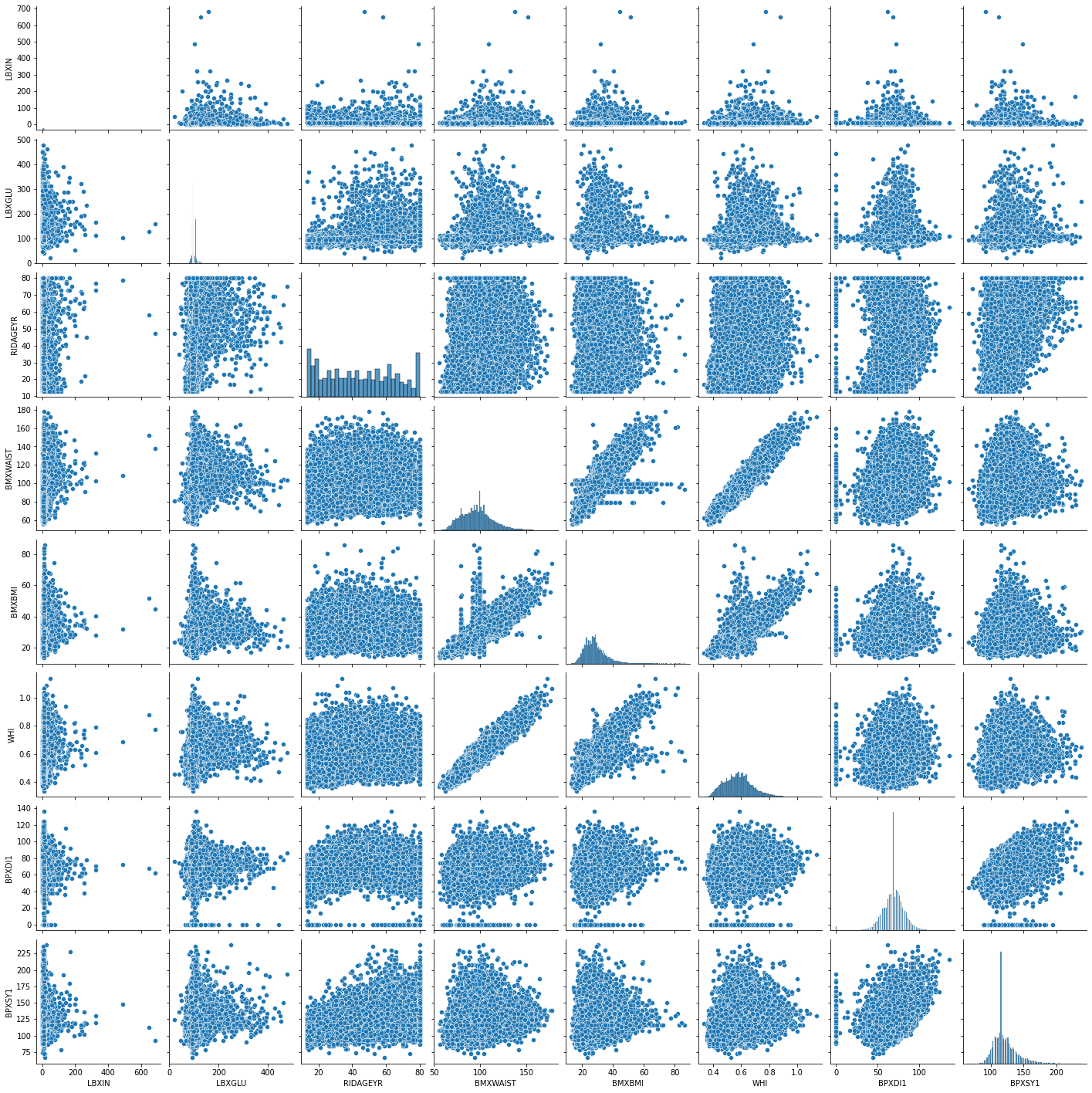
Description automatically generated

Figura 13: Diagrama de dispersión antes de borrado de outliers

Figura 14: Diagrama de dispersión después de borrado de outliers

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 (WHI) dividiendo directamente los valores de la cintura de los participantes entre su altura, ambas variables en centímetros. Adicionalmente, según los criterios establecidos en *Tabla 4* 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 dataset acaba sumando 45 columnas.

### Análisis estadístico de los datos

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 en la **Error! Reference source not found.**. De todas los registros el 50.7% representa a mujeres (RIAGENDR = 2) y el 49.3% a hombres (RIAGENDR = 1).

Chart, bar chart

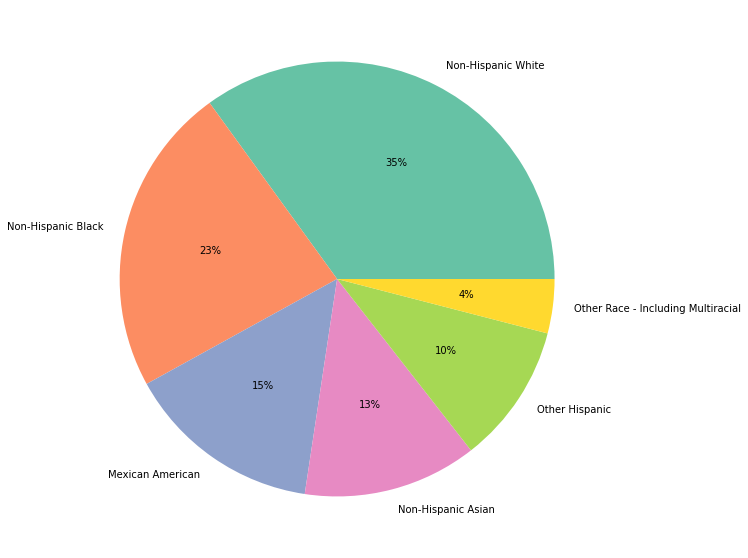
Description automatically generated

Figura 15: Distribución por género

Figura 16: Distribución por etnia

Chart, pie chart

Description automatically generated

Figura 17: Distribución por grupo de edad

Chart, histogram

Description automatically generated

Figura 18: Histograma de la edad de los participantes

En cuanto a la etnia, hay una mayor participación de personas blancas no hispanas (33%), seguido de no hispanos negros (24%), americanos mexicanos (16%), asiáticos (12%), hispanos (11%) y finalizando con un 5% de representación de otra raza.

Prevalencia de sindrome metabólico de los participantes

De todos los participantes, el 29.9% presenta síndrome metabólico. Tanto la edad avanzada como el genéro 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 a los dos primeros grupos de edad, tanto para el resto del análisis como para entrenar y probar los modelos de predicción. De esta forma, el dataset 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 [62] la prevalencia de SM en población adulta es del 35%, coincidiendo con otro que incorpora los años del 2011 hasta el 2016[63]. Bastante evidencia afirma, además, que cada año este porcentaje se incrementa [64]. Por lo tanto se puede considerar que los resultados son cifras razonables.

Por otro lado, se observa que los requisitos 1, 3 y 5 del SM son los que más se suelen dar entre los individuos de la muestra. Siendo el segundo el que menos.

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

Figura 19: Prevalencia de SM por género

Figura 20: Prevalencia de SM por grupo de edad

Distribución de los participantes por medidas corporales

Chart, box and whisker chart

Description automatically generatedEl 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 5: categorías según el índice cintura alturaTabla 5. A continuación se muestran los boxplots separados por grupos de edad sobre el IMC, el índice cintura altura y el perímetro de cintura.

Figura 21: Cintura de los participantes por grupo de edad

Chart, box and whisker chart

Description automatically generated

Figura 22: Indice cintura-altura por grupo de edad

Chart, box and whisker chart

Description automatically generated

Figura 23: IMC por grupo de edad

Se observa que el IMC y el WIH se distribuye de manerea 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, sobrepesa elevado y obesidad mientras que más de la mitad de las mujeres padecen obesidad. 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. [65], [66]

Tabla 7: ratios de sobrepeso entre los participantes

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

### 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 puede usar la siguiente guía.

* 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

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 se observa más relación entre el síndrome metabólico y ser blanco no hispano. Por otro lado, vemos como hay una ligera relación

* **Etnia\_1.0:** Mexican American
* **Etnia\_2.0:**  Other Hispanic
* **Etnia\_3.0:**  Non-Hispanic White
* **Etnia\_4.0:** Non-Hispanic Black
* **Etnia\_6.0:** Non-Hispanic Asian
* **Etnia\_7.0:** Other race

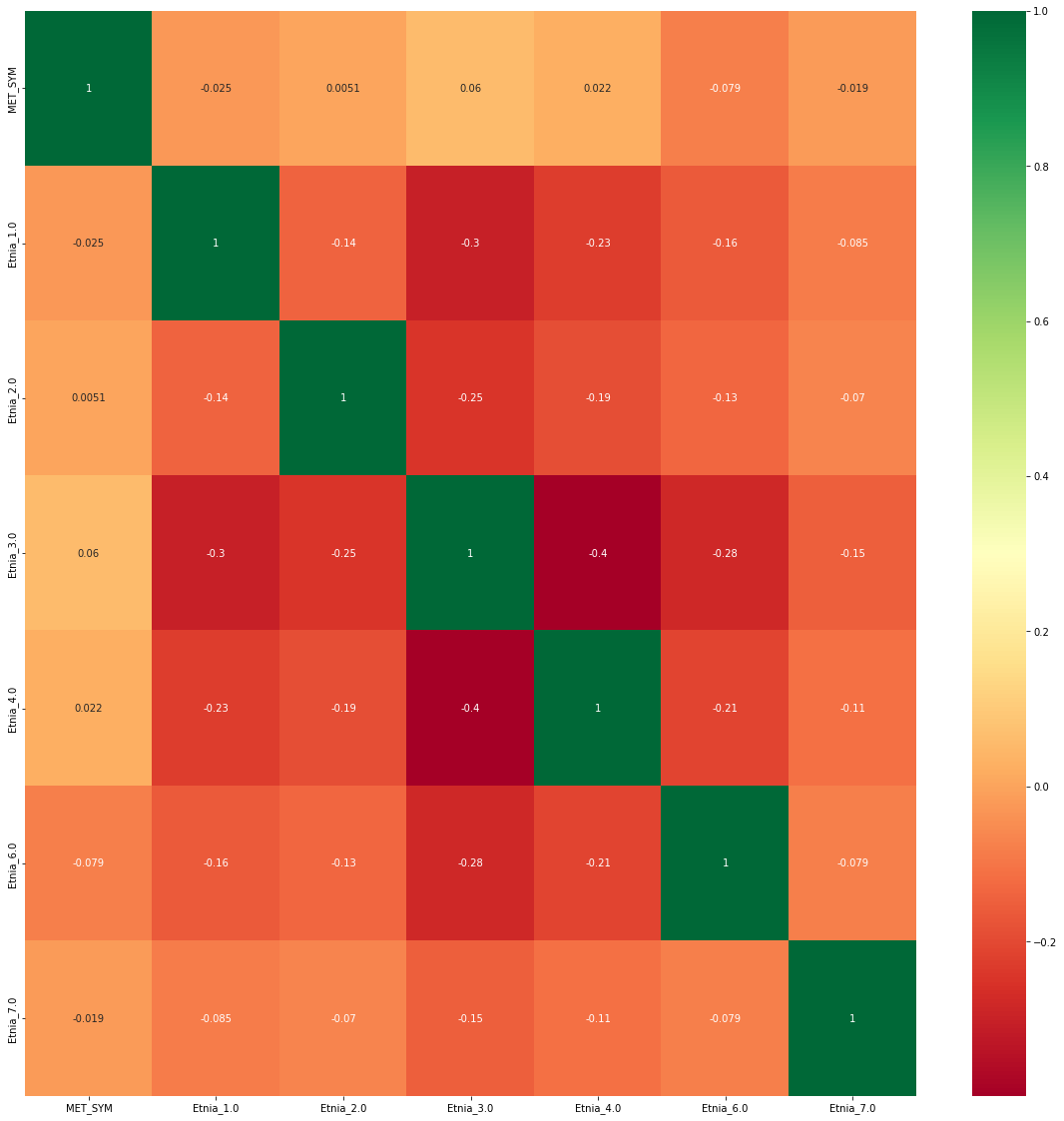


Figura 24: Correlación por etnia

Correlación por variables demográficas

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

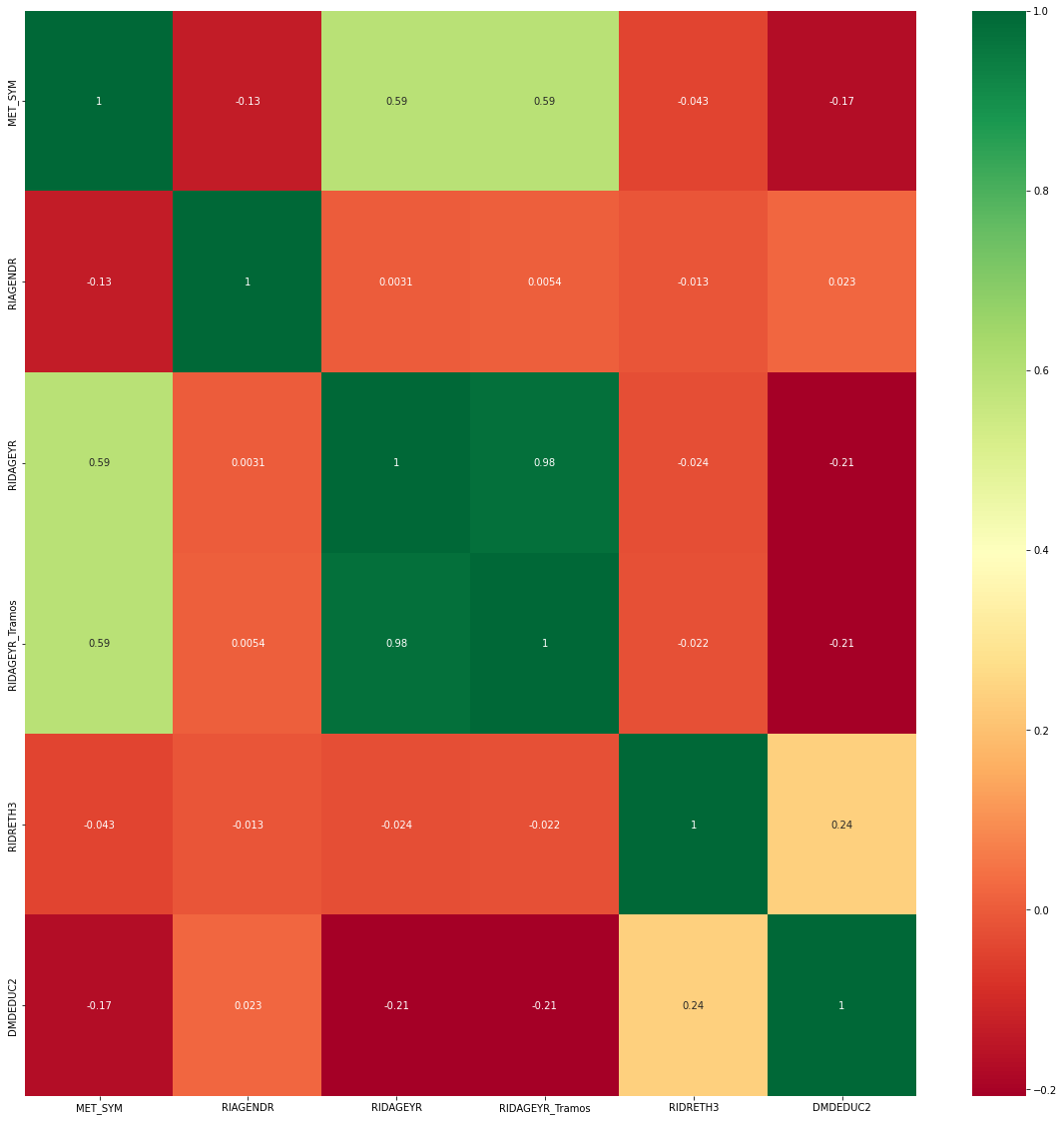


Figura 25: Correlación por variables demográficas

Chart

Description automatically generated

Figura 26: SM por nivel educativo

Correlación por ejercicio físico

A continuación se observa 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 MET\_SYM.

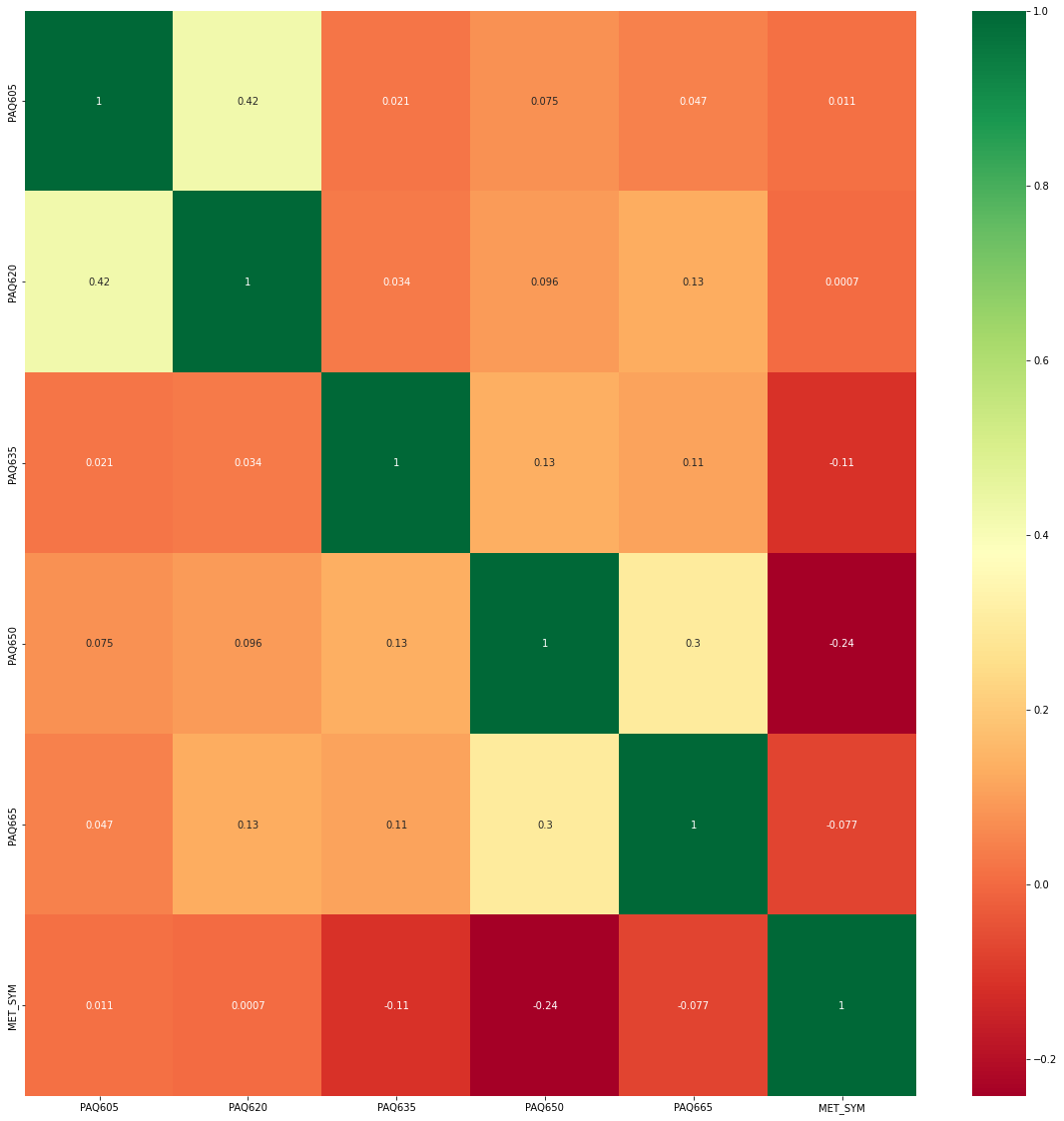


Figura 27: Correlación por ejercicio físico

Correlación por alcohol y tabaquismo

En este caso se observa una clara relación entre fumar y presentar SM. Mientras que el consumo de alcohol no parece afectar demasiado. De todos los participantes con síndrome metabólico el 48% son fumadores.

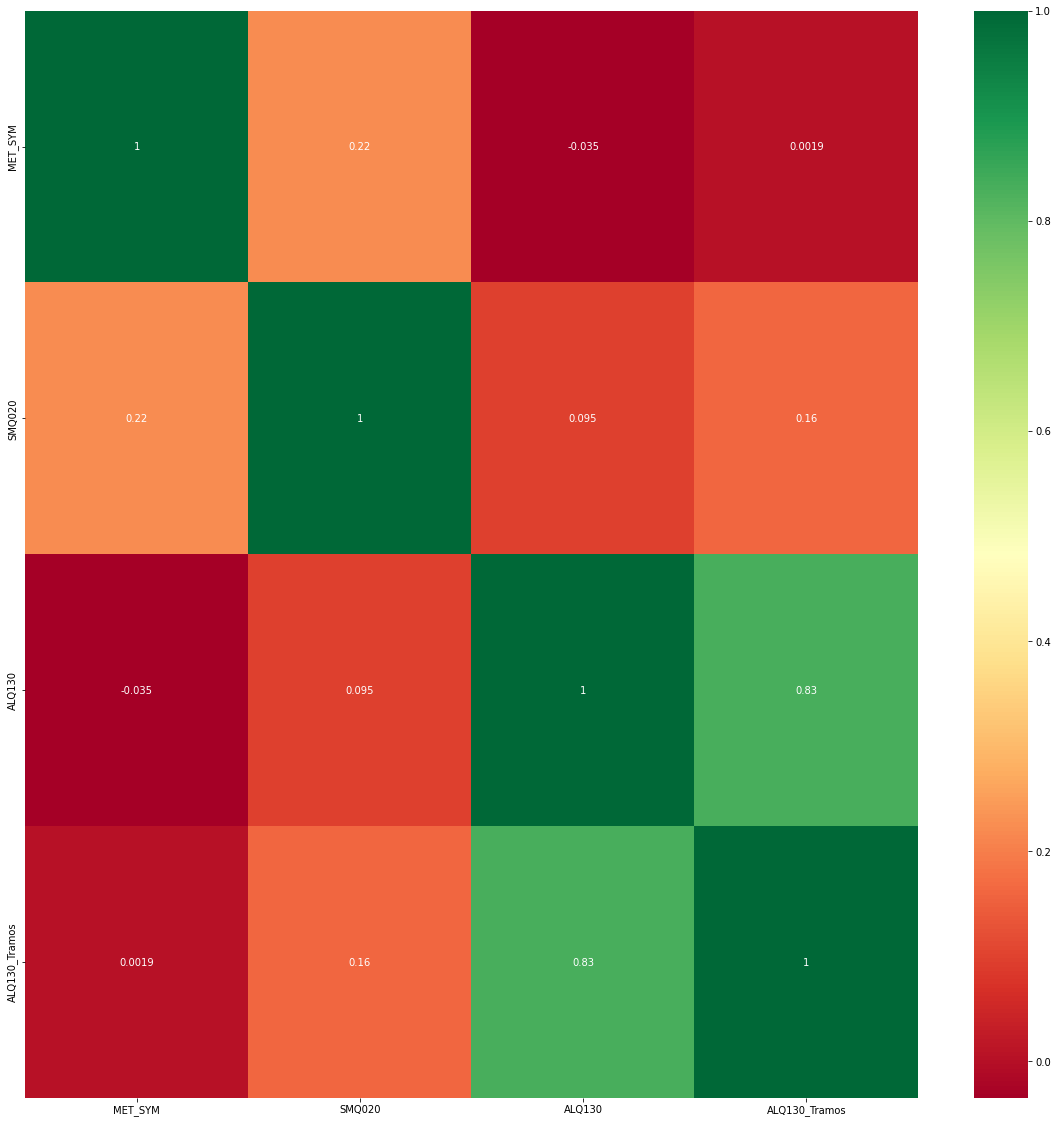


Figura 28: Correlación por alcohol y tabaquismo

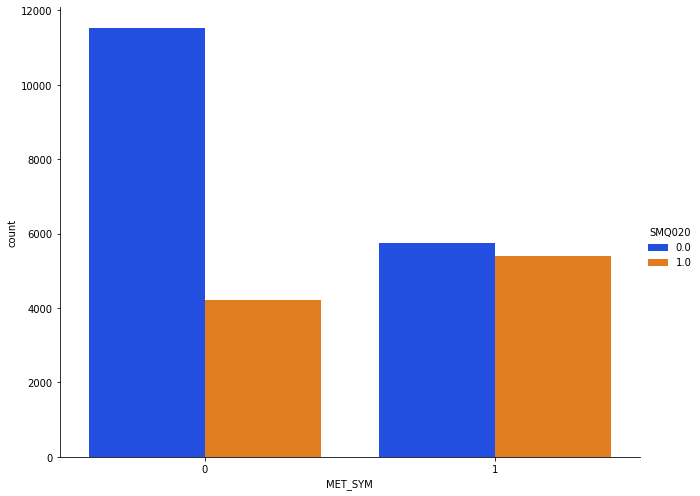


Figura 29: SM en función de ser fumador

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, mientras que el que evalúa la calidad de su dieta no. Cabe destacar que valorar 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.

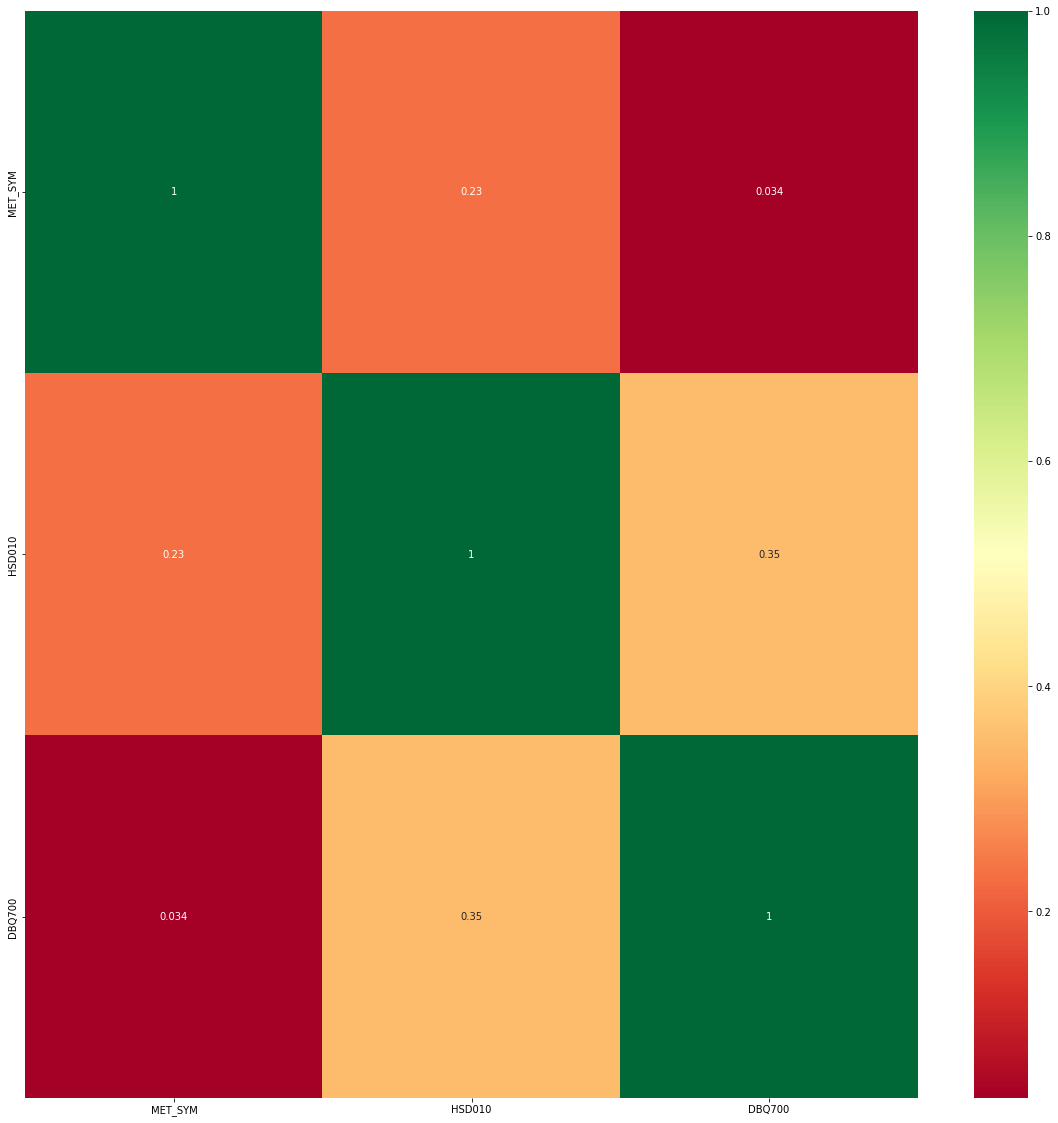


Figura 30: Correlación por calidad de dieta y estado general de salud

Correlación por indicaciones del doctor sobre presentar sobrepeso, asma, cáncer y diabetes

Se observa bastante correlación entre que el doctor le haga saber si estado de sobrepeso (MCQ080), cáncer (MCQ220) y/o diabetes (MCQ300C) y tener síndrome metabólica, mientras que tener asma (MCQ010) no dice mucho sobre el SM.

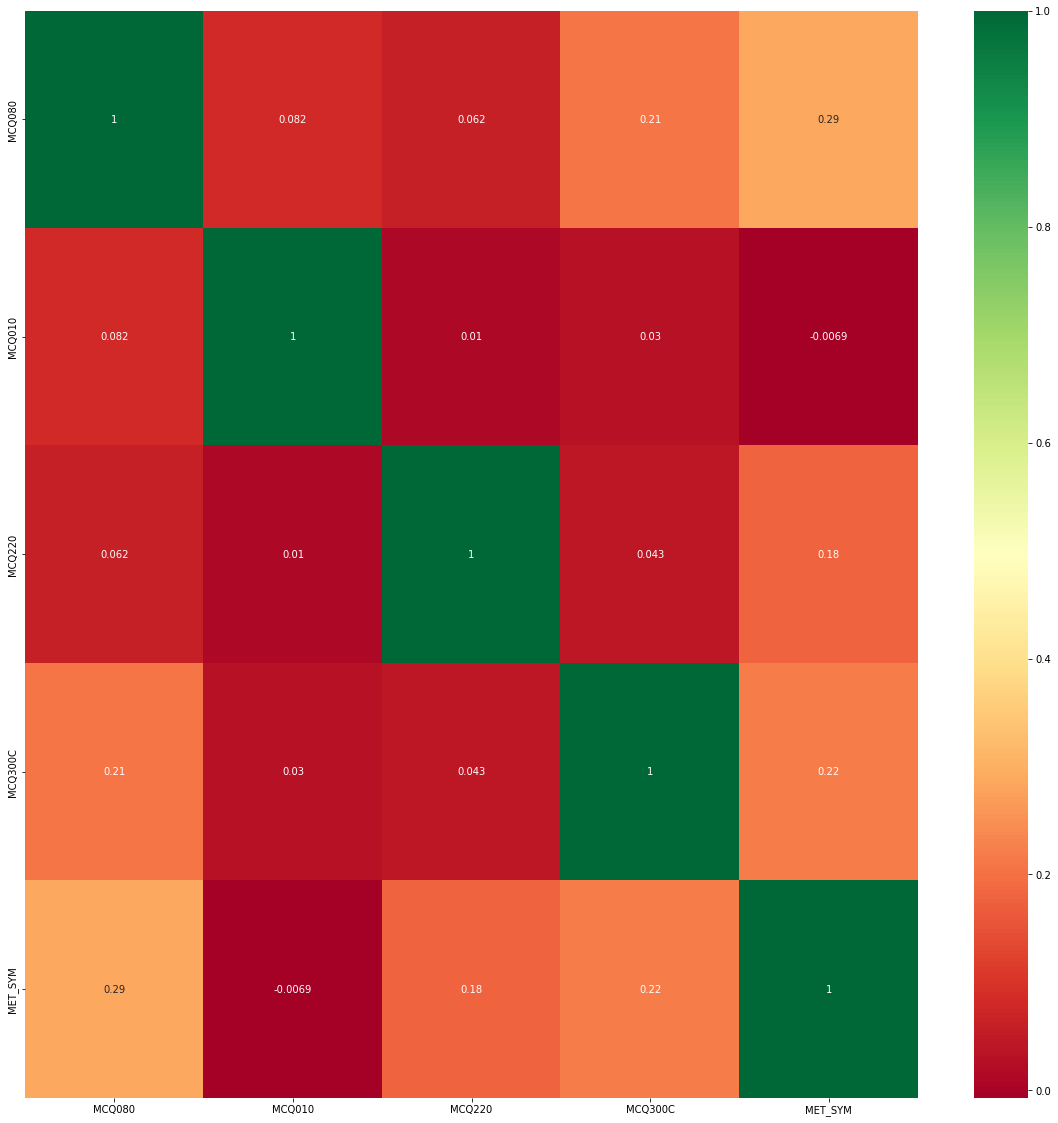


Figura 31: Correlación por sobrepeso, asma, cáncer y diabetes

Correlación por problemas para dormir y depresión

Se puede establecer una correlación débil entre DPQ020 (sentirse triste, deprimido o desesperanzado), DPQ030 (problemas para dormir o dormir demasisado), 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,

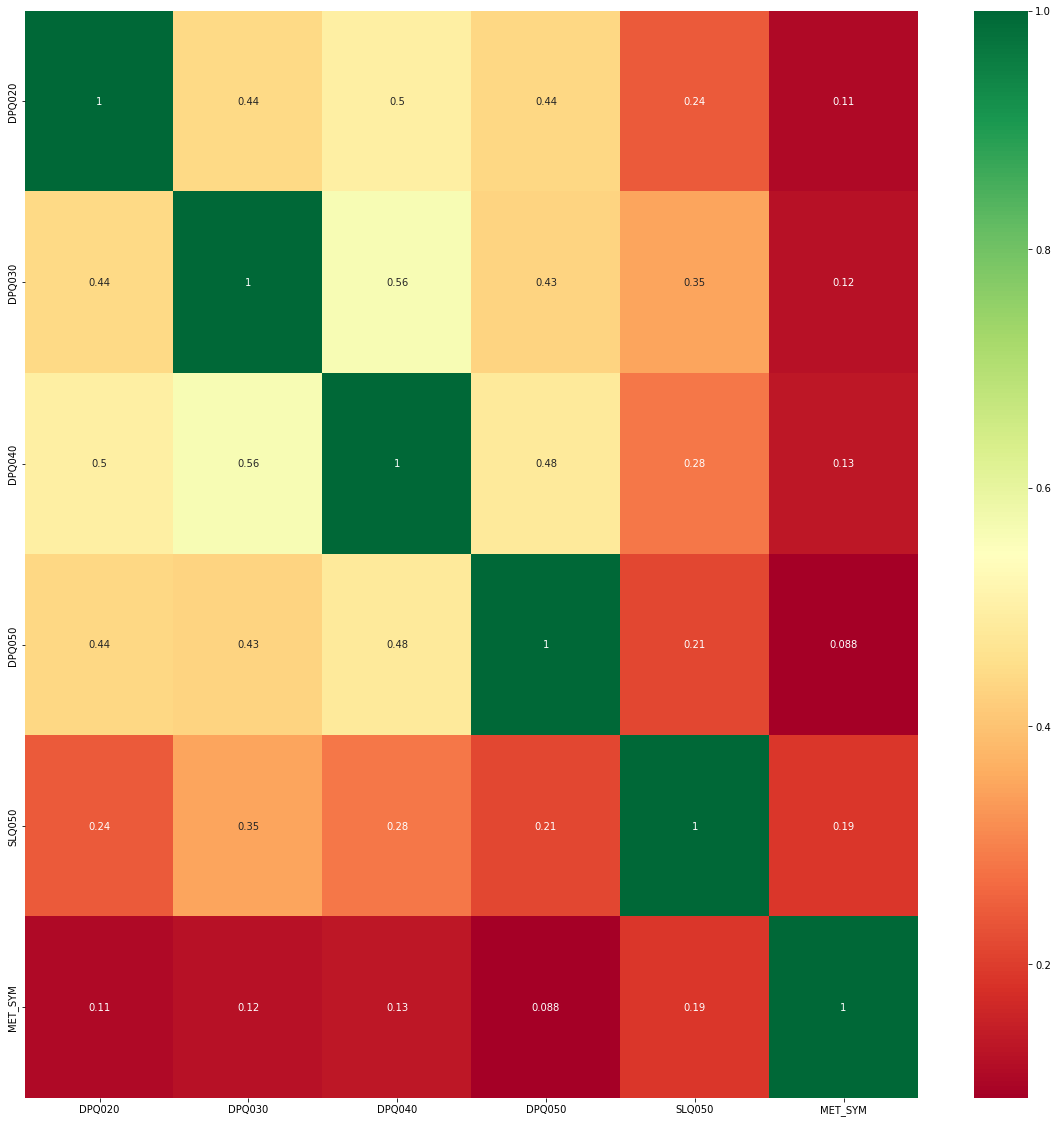


Figura 32: Correlacion con problemas de depresión y sueño

Correlación con nivel adquisitivo

Se muestra a continuación la matriz de correlación por nivel adquisito. 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. El 34.37% de individuos con SM no percibe ingresos de su salario.

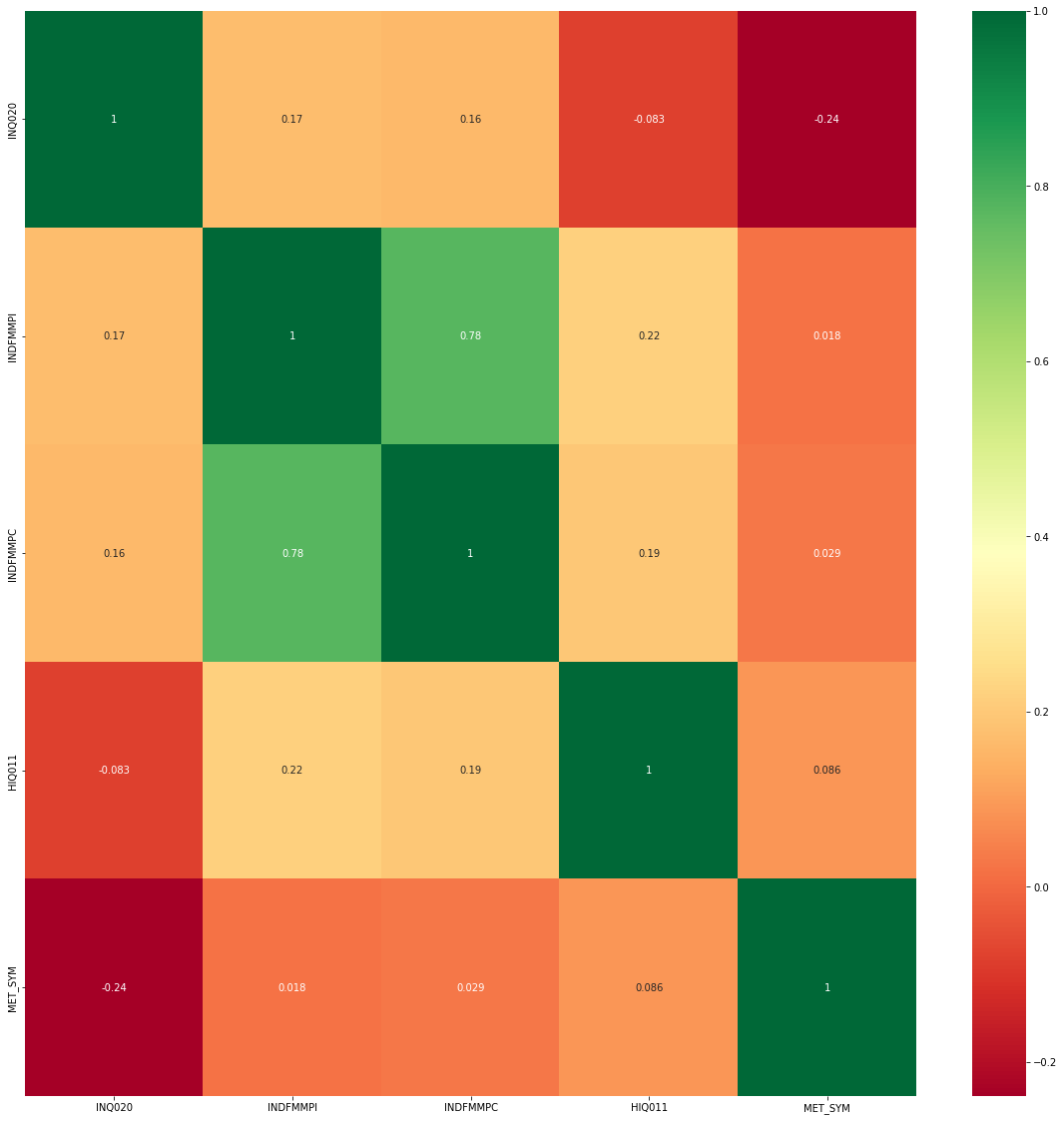


Figura 33: Correlación por nivel adquisitivo

## Conclusiones

…

## Líneas futuras

…

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

[2] C. K. Roberts, A. L. Hevener, and R. J. Barnard, “Metabolic Syndrome and Insulin Resistance: Underlying Causes and Modification by Exercise Training,” 2013, doi: 10.1002/cphy.c110062.

[3] 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.

[4] 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.

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

[6] 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.

[7] “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).

[8] 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.

[9] M. Cuesta *et al.*, “Incidence and regression of metabolic syndrome in a representative sample of the Spanish population: Results of the cohort di@bet.es study,” *BMJ Open Diabetes Research and Care*, vol. 8, no. 1, Oct. 2020, doi: 10.1136/BMJDRC-2020-001715.

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

[11] 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.

[12] “Obesidad: un desafío para las políticas públicas.” http://www.scielo.org.mx/scielo.php?script=sci\_arttext&pid=S0188-45572013000100007 (accessed May 05, 2022).

[13] “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).

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

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

[16] “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).

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] “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).

[23] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective.”

[24] 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.

[25] 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.

[26] 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.

[27] “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).

[28] “1.4. Support Vector Machines — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/svm.html (accessed May 05, 2022).

[29] “1.10. Decision Trees — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/tree.html (accessed May 05, 2022).

[30] 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.

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

[32] 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.

[33] 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.

[34] “8 Applications of Neural Networks | Analytics Steps.” https://www.analyticssteps.com/blogs/8-applications-neural-networks (accessed May 05, 2022).

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

[36] “(PDF) Síndrome metabólico: definición, historia, criterios.” https://www.researchgate.net/publication/26508500\_Sindrome\_metabolico\_definicion\_historia\_criterios (accessed May 05, 2022).

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] “(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).

[43] 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.

[44] 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.

[45] 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.

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

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

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

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

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

[51] 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.

[52] “Welcome to Python.org.” https://www.python.org/ (accessed May 05, 2022).

[53] “NumPy.” https://numpy.org/ (accessed May 05, 2022).

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

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

[56] “SciPy.” https://scipy.org/ (accessed May 05, 2022).

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

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

[59] “Project Jupyter | Home.” https://jupyter.org/ (accessed May 05, 2022).

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

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

[62] 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.

[63] 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.

[64] 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.

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

[66] “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).

# anexo a: aspectos ÉTICOS, económicos, sociales y ambientales

El apartado “Requisitos de las acreditaciones internacionales EUR-ACE y ABET” de la Normativa de TFT de la ETSIT-UPM establece que “*La memoria del TFT del GITST, GIB y MUIT, y en general la de aquellas titulaciones que hayan obtenido o para las que se desee solicitar una acreditación internacional EUR-ACE o ABET, debe mostrar conciencia de la responsabilidad de la aplicación práctica de la ingeniería, el impacto social y ambiental, el compromiso con la ética profesional, la responsabilidad y las normas de la aplicación práctica de la ingeniería, así como sobre las prácticas de gestión de proyectos, gestión, y control de riesgos, entendiendo sus limitaciones*”.

Este anexo obligatorio del TFT tendrá un carácter sintético con los siguientes apartados:

## A.1 Introducción

Breve descripción del contexto del proyecto, objetivos, necesidades que pretende cubrir o problemas que pretende resolver, centrándose en su relación con los temas sociales, económicos, éticos, legales y/o ambientales que se hayan identificado.

## A.2 DESCRIPCIÓN DE IMPACTOS RELEVANTES RELACIONADOS CON EL PROYECTO

Síntesis del trabajo realizado en la fase 2, de selección y descripción de impactos. Presentar y justificar las conclusiones a las que se haya llegado sobre cuáles son los asuntos más relevantes relacionados con la sostenibilidad social, económica o ambiental, así como los principales grupos de interés identificados y que se han considerado en los análisis posteriores.

## A.3 ANÁLISIS DETALLADO DE ALGUNO DE LOS PRINCIPALES IMPACTOS

Síntesis del trabajo de análisis realizado.

## A.4 CONCLUSIONES

Valorar el proyecto desde un punto de vista ético, social, económico y medioambiental y justificar si el uso de criterios de sostenibilidad ha aportado o puede aportar valor añadido al proyecto.

# anexo b: presupuesto económico

El apartado “Requisitos de las acreditaciones internacionales EUR-ACE y ABET” de la Normativa de TFT de la ETSIT-UPM establece que “*La memoria del TFT del GITST, GIB y MUIT, y en general la de aquellas titulaciones que hayan obtenido o para las que se desee solicitar una acreditación internacional EUR-ACE o ABET, … debe incluir un presupuesto económico*”. A modo de ejemplo, la tabla del presupuesto de un proyecto podría ser la siguiente:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **COSTE DE MANO DE OBRA (coste directo)** | | | | | **Horas** | | **Precio/hora** | **Total** |
|  | | | | | 300 | | 15 € | **4.500 €** |
|  | | | |  |  | |  |  |
| **COSTE DE RECURSOS MATERIALES (coste directo)** | | | | **Precio de compra** | **Uso en meses** | | **Amortización (en años)** | **Total** |
| Ordenador personal (Software incluido)....... | | | | 1.500,00 € | 6 | | 5 | 150,00 € |
| Impresora láser | | | | 500,00 € | 6 | | 5 | 50,00 € |
| Otro equipamiento | | | |  |  | |  |  |
|  | | | |  |  | |  |  |
| **COSTE TOTAL DE RECURSOS MATERIALES** | | | | | | | | **200,00 €** |
|  | |  | |  | |  | |  |
| **GASTOS GENERALES (costes indirectos)** | | 15% | | sobre CD | | | | **705,00 €** |
| **BENEFICIO INDUSTRIAL** | | 6% | | sobre CD+CI | | | | **324,30 €** |
|  | |  | |  | |  | |  |
| **MATERIAL FUNGIBLE** | |  | |  | |  | |  |
| Impresión | | | | | | | | **100,00 €** |
| Encuadernación | | | | | | | | **300,00 €** |
|  |  | |  | | |  | |  |
| **SUBTOTAL PRESUPUESTO** | | | | | | | | **6.129,30 €** |
| **IVA APLICABLE** | | | | | | | 21% | **1.287,15 €** |
|  |  | |  | | | |  |  |
| **TOTAL PRESUPUESTO** | | | | | | | | **7.416,45 €** |

Esta tabla podría ser rellenada mediante una sencilla hoja de cálculo como la siguiente:

