Universitat Oberta de Catalunya

uoc.edu

Estudio del conjunto de datos NHANES mediante el empleo de técnicas de aprendizaje no supervisado.

Raúl Sánchez Temporal
Máster Universitario en Ciencia de datos
Área: Machine learning en medicina

Consultora: Dra. Laia Subirats Maté
Profesor Responsable de la asignatura: Dr. Ferran Prados Carrasco

Enero 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio del conjunto de datos NHANES mediante el empleo de técnicas de aprendizaje no supervisado.*
Nombre del autor:	*Raúl Sánchez Temporal*
Nombre del consultor/a:	*Laia Subirats Maté*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster universitario Ciencia de datos*
Área del Trabajo Final:	*Machine learning en medicina.*
Idioma del trabajo:	*Castellano*
Palabras clave	*NHANES, Machine learning, Clustering*

Resumen del Trabajo (máximo 250 palabras):

El conjunto de datos National Health and Nutrition Examination Survey (NHANES) facilitado por el Centro de Control de enfermedades y Prevención (CDC) supone una oportunidad única para llevar a cabo investigaciones y análisis que puedan ayudar en la mejora de la salud de las personas.

En este trabajo se propone el uso de técnicas de aprendizaje no supervisado aplicado sobre los datos NHANES con el objetivo de detectar patrones que permitan perfilar a los pacientes en base a sus similitudes encontrando grupos (clústeres) naturales para estos. En concreto el trabajo se centra en el uso de métodos de agrupamiento basados en densidad y métodos jerárquicos. A su vez se crea una interfaz web que permite la clasificación de los pacientes en los diferentes clústeres que se generen.

Para el desarrollo del trabajo se sigue la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) ampliamente adoptada para proyectos de minería de datos que proporciona la descripción del ciclo de vida donde se definen las tareas necesarias para cada fase.

Abstract (in English, 250 words or less):

The National Survey of Health and Nutrition Survey (NHANES) data set provided by the Center for Disease Control and Prevention (CDC) is a unique opportunity to conduct research and analysis that can help improve the health of people.

This paper proposes the use of unsupervised learning techniques applied to NHANES data in order to detect patterns that adapt to patients based on their similarities by finding natural groups (clusters) for them. Specifically, the work focuses on the use of methods of grouping methods in density and hierarchical methods. In addition, a web interface is created that allows the classification of patients in the different clusters that are generated.

For the development of the work, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is followed, which is widely adopted for data mining projects that describe the life cycle where the necessary tasks are defined for each phase.

Índice

Lista de figuras

Lista de Tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

Desde los años sesenta el Centro Nacional de Estadísticas de la Salud (NCHS)[1] parte del Centro para el control y Prevención de enfermedades (CDC)[2] de los Estados Unidos de América mediante su programa NHANES[3] [1], realiza encuestas sobre diferentes temas relacionados con la salud y la nutrición de los ciudadanos estadounidenses. Esta encuesta se realiza sobre una muestra representativa de unas 5000 personas residentes en 15 condados que se visitan cada año.

Estas entrevistas incluyen datos demográficos, socioeconómicos, dietéticos y relativos a la salud. Paralelamente se realizan exámenes a los participantes que incluyen mediciones físicas, dentales y tests de laboratorios.

CDC pone a disposición de investigadores de todo el mundo los datos NHANES de forma pública para que sean usados en investigaciones con el objetivo mejorar la salud de la población, de hecho, son múltiples los logros conseguidos a partir de los datos recabados y sus posteriores análisis. Por ejemplo, tal como indica el CDC, la detección de diabetes no diagnosticadas facilita el lanzamiento de programas que permitan erradicar esta situación, así como la detección de plomo en sangra en las diferentes encuestas permitieron lanzar políticas para la eliminación de este compuesto de alimentos y como componente en la gasolina.

Lo indicado muestra la importancia de este programa, tanto por su continuidad en el tiempo, el volumen de datos recogidos así como su calidad al ser realizado por profesionales de la salud, ofrece por tanto la oportunidad de llevar a cabo análisis cuyos resultados pueden colaborar en la mejora de la salud de las personas.

Desde el punto de vista personal la elección de esta línea de trabajo es principalmente por mi interés, como hobby, en estudiar temas relacionados con hábitos saludables de vida a la vez que también capta mi interés la investigación social, aspectos que, dado el conjunto de datos NHANES y sus características, seguro permite que se puedan abarcar ambos en este trabajo. Desde el punto de vista de la ciencia de datos atrajo también mi

[1] National Center for Health Statistics (NCHS) , www.cdc.gov/nchs/index.htm

[2] Centers for Disease Control and Prevention (CDC), www.cdc.gov

[3] National Health and Nutrition Examination Survey (NHANES), www.cdc.gov/nchs/nhanes/

interés el conjunto de datos con el que se va a trabajar, tanto en su formato, su continuidad en el tiempo, o el método de recolección, el cual se demuestra tiene un gran potencial, dado la cantidad de investigaciones existentes que lo usan, permitiendo la aplicación de una gran variedad de técnicas de aprendizaje automático como las que se tratan en este trabajo para la obtención de conocimiento para la mejora de la salud.

En este trabajo se pretende aplicar técnicas de aprendizaje no supervisado para la creación de diferentes modelos sobre los datos facilitados con el objetivo de detectar patrones desconocidos que permitan la segmentación de pacientes en base a sus similitudes. Facilitando, si es posible, ya sea la detección de enfermedades de forma automática, o características descriptivas de grupos.

1.2 Objetivos del Trabajo

El objetivo principal de este trabajo es el de poder perfilar los pacientes mediante un estudio basado en técnicas de aprendizaje no supervisado sobre los datos del conjunto NHANES y su presentación en una interfaz web, el estudio se centrará en el uso de las técnicas y algoritmos de clustering jerárquico y basado en densidad como Density-based spatial clustering of applications with noise (DBSCAN).

Durante el estudio se establecen los siguientes objetivos:

- **Importar datos** desde la página web NHANES mediante el paquete estadístico SAS[4] y exportación posterior para llevar a cabo los análisis necesarios.

- **Preparar los datos** a aplicar en los modelos no supervisados que se crearán donde se evaluará su calidad, relevancia, se realizarán tareas de limpieza o transformación con el objetivo de construir un juego de datos apto.

- **Crear los modelos de clustering** para la detección de patrones ocultos que puedan aparecer en los datos que permitan el perfilado de los pacientes.

- **Describir y validar los clusters** generados con el objetivo de profundizar en su interpretación y entendimiento intentando medir sus índices internos como son la cohesión y separación.

[4] Los datos se encuentran almacenados en ficheros SAS transport (.XPT), lo cual hace necesario el uso del paquete estadístico SAS para su recolección y exportación. https://www.sas.com/es_es/software/stat.html

- **Construir una interfaz web** que a través de un formulario permita al usuario clasificar instancias y visualizar los resultados del clúster asignado. La aplicación web se desarrolla en lenguaje python mediante un notebook jupyter con extensiones necesarias, el framework Flask y publicación final en la plataforma Heroku.

1.3 Enfoque y método seguido

Se quiere obtener conocimiento de los participantes en la encuesta NHANES y presentar este conocimiento en una interfaz web, se trata de un producto nuevo a desarrollar para el cual se sigue la metodología Cross Industry Standard Process for Data Mining (CRISP-DM), ampliamente adoptada en proyectos de minería de datos convirtiéndose en la metodología del sector [2], se utiliza como guía de trabajos en los que se sigue una secuencia de fases de forma iterativa a través de una serie de ciclos que incluyen planificación, ejecución y revisión que aseguran la calidad del mismo.

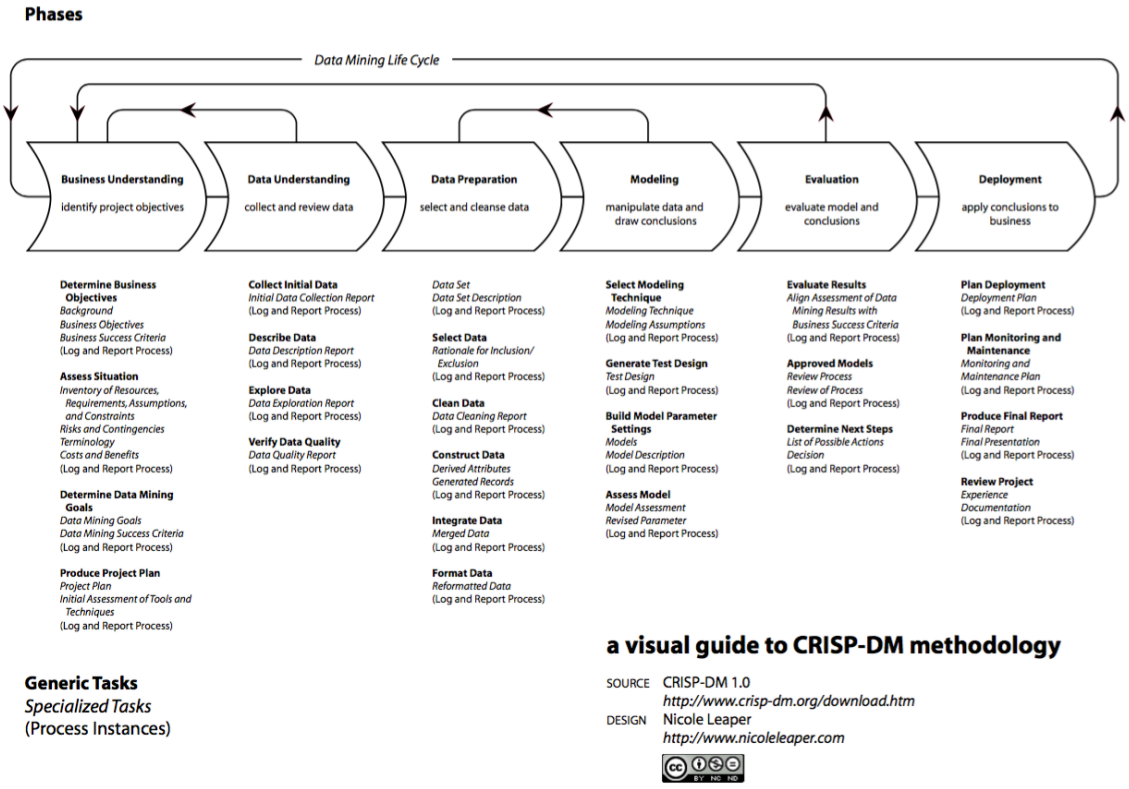


Fig 1. CRISP-DM process model (Nicole Leaper)

La metodología CRISP-DM consta de las fases siguientes:

Comprensión del negocio:

En esta fase se trata de conocer el dominio en el que se mueve el proyecto, establecer objetivos, evaluar su situación actual y una planificación del mismo. Definidas en secciones 1.1 a 1.4 del documento.

Comprensión de los datos:

Con el objetivo de comprender los datos, conocer su estructura y calidad. Se ejecuta el proceso de importación con el paquete estadístico SAS donde se realiza una selección previa de los datos que se usan en el estudio, se adjuntan y combinan los ciclos y componentes de interés. Finalmente se almacenan y exportan los datos. A continuación mediante el uso de notebook de jupyter en lenguaje python se exploran, se describen los datos y se evalúa su calidad para formar el conjunto final.

Preparación de los datos:

En esta fase se prepara el conjunto de datos para la aplicación de cada uno de los modelos de segmentación que se llevan a cabo a continuación. Se realizan en esta fase tareas de limpieza, detección de outliers y se aplican técnicas estadísticas para evaluar el uso de técnicas de reducción de dimensionalidad

Modelado:

Para las tareas no supervisadas de clustering (agrupamiento) se construyen modelos con algoritmos jerárquicos y basados en densidad, se verifica su calidad y se ajustan los modelos según los objetivos fijados.

Evaluación:

En esta fase se evalúan los resultados de los modelos, los agrupamientos descubiertos que se han podido generar y su capacidad para el perfilado de los pacientes en el conjunto. En función de estos resultados se itera sobre fases anteriores y se revisan procesos.

Despliegue:

En esta fase se presenta el conocimiento descubierto en forma de aplicación web que permitirá interactuar en la aplicación de los modelos sobre los datos y representar los resultados. Se desarrolla en este caso la interfaz web en lenguaje python sobre un notebook jupyter para la creación del modelo de clasificación y para el despliegue se usa el framework FLASK para la creación de los formularios y presentación de datos y finalmente la publicación usando los servicios de la plataforma Heroku.

La elección se basa en sus ventajas principales [3] como son que se trata de software open source, es ampliamente usado y soportado en múltiples disciplinas, multiplataforma y por último que los notebooks de jupyter permiten el uso de diferentes lenguajes. A parte de facilitar la colaboración, la publicación, la distribución de los resultados.

1.4 Planificación del Trabajo

La planificación del trabajo se realiza teniendo en cuenta las fechas de las entregas de las diferentes PEC a lo largo del proyecto.

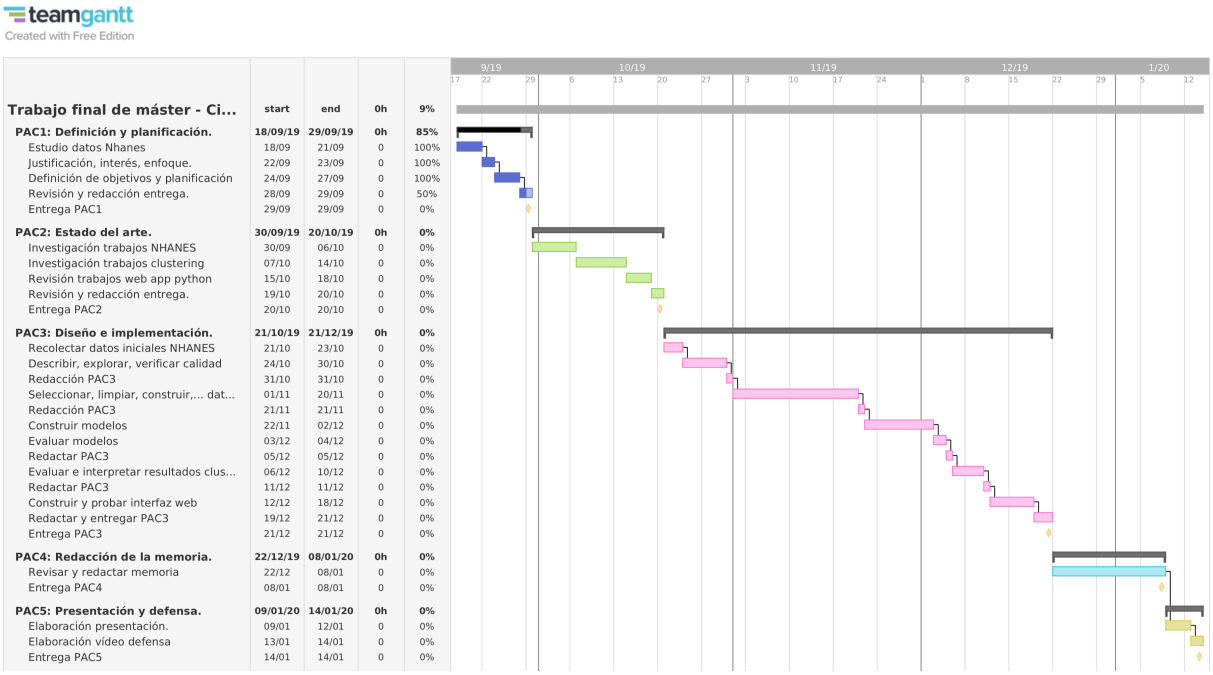
Tabla 1: Fechas de entrega de PACS

Entrega	Fecha
PAC1: Definición y planificación del trabajo final	29/09/19
PAC2: Estado del arte	20/10/19
PAC3: Diseño e implementación	21/12/19
PAC4: Redacción de la Memoria	08/01/20
PAC5: Presentación y defensa	14/01/20

Tabla 2: Planificación de las tareas.

Tarea	Días	Inicio	Fin
Definición y planificación	**12**	**18/09/19**	**29/09/20**
Comprensión del negocio			
Estudio datos NHANES	4	18/09	21/09

Tarea	Días	Inicio	Fin
Justificación trabajo	2	22/09	23/09
Definir objetivos y planificar	4	24/09	27/09
Redactar, revisar y entregar PAC1	2	28/09	29/09
Estado del arte	**21**	**30/09/19**	**20/10/19**
Comprensión del negocio			
Investigar trabajos NHANES	7	30/09	06/09
Investigar trabajos clustering	8	07/10	14/10
Investigar trabajos web app python	4	15/10	18/10
Revisar, redactar y entregar PAC2	2	19/10	20/10
Diseño en implementación	**62**	**21/10/19**	**21/12/19**
Comprensión de los datos			
Recolectar datos iniciales NHANES	3	21/10	23/10
Describir, explorar, verificar calidad datos	7	24/10	30/10
Redactar PAC3	1	31/10	31/10
Preparación de los datos			
Seleccionar, limpiar, construir, integrar datos	20	01/11	20/11
Redactar PAC3	1	21/11	21/11
Modelado			
Construir modelos	11	22/11	02/12
Evaluar modelos	2	03/12	04/12
Redactar PAC3	1	05/12	05/12
Evaluación			
Evaluar e interpretar resultados de clúster	5	06/12	10/12
Redactar PAC3	1	11/12	11/12
Despliegue			
Construcción y pruebas interfaz web	7	12/12	18/12
Redactar y entregar PAC3	3	19/12	21/12
Redacción de la Memoria	**18**	**22/12/19**	**08/01/20**
Revisar y redactar memoria.	18	22/12/19	08/01/20
Presentación y defensa del proyecto	**6**	**09/01/20**	**14/01/20**
Elaboración de presentación.	4	09/01	12/01
Elaboración vídeo defensa.	2	13/01	14/01



Trabajo final de máster - Ci...	start	end	0h	9%
PAC1: Definición y planificación.	**18/09/19**	**29/09/19**	**0h**	**85%**
Estudio datos Nhanes	18/09	21/09	0	100%
Justificación, interés, enfoque.	22/09	23/09	0	100%
Definición de objetivos y planificación	24/09	27/09	0	100%
Revisión y redacción entrega.	28/09	29/09	0	50%
Entrega PAC1	29/09	29/09	0	0%
PAC2: Estado del arte.	**30/09/19**	**20/10/19**	**0h**	**0%**
Investigación trabajos NHANES	30/09	06/10	0	0%
Investigación trabajos clustering	07/10	14/10	0	0%
Revisión trabajos web app python	15/10	18/10	0	0%
Revisión y redacción entrega.	19/10	20/10	0	0%
Entrega PAC2	20/10	20/10	0	0%
PAC3: Diseño e implementación.	**21/10/19**	**21/12/19**	**0h**	**0%**
Recolectar datos iniciales NHANES	21/10	23/10	0	0%
Describir, explorar, verificar calidad	24/10	30/10	0	0%
Redacción PAC3	31/10	31/10	0	0%
Seleccionar, limpiar, construir,... dat...	01/11	20/11	0	0%
Redacción PAC3	21/11	21/11	0	0%
Construir modelos	22/11	02/12	0	0%
Evaluar modelos	03/12	04/12	0	0%
Redactar PAC3	05/12	05/12	0	0%
Evaluar e interpretar resultados clus...	06/12	10/12	0	0%
Redactar PAC3	11/12	11/12	0	0%
Construir y probar interfaz web	12/12	18/12	0	0%
Redactar y entregar PAC3	19/12	21/12	0	0%
Entrega PAC3	21/12	21/12	0	0%
PAC4: Redacción de la memoria.	**22/12/19**	**08/01/20**	**0h**	**0%**
Revisar y redactar memoria	22/12	08/01	0	0%
Entrega PAC4	08/01	08/01	0	0%
PAC5: Presentación y defensa.	**09/01/20**	**14/01/20**	**0h**	**0%**
Elaboración presentación.	09/01	12/01	0	0%
Elaboración vídeo defensa	13/01	14/01	0	0%
Entrega PAC5	14/01	14/01	0	0%

Fig 2. Planificación - Diagrama de Gantt

1.5 Breve sumario de productos obtenidos

A partir del proceso de recolección que se describe en el capítulo 2 se ha creado, a partir de los datos públicos NHANES, un conjunto de datos de 29902 observaciones y 24 variables, estos abarcan un amplio espacio temporal de tres ciclos (6 años) de datos de carácter demográfico, médico o social. Este conjunto puede ser de interés para llevar a cabo distintos análisis estadísticos o tareas de aprendizaje automático. A su vez se encuentra disponible el script SAS para la extracción de los datos desde los ficheros NHANES.

Se han generado durante el proceso de análisis del trabajo cuatro notebooks de jupyter en lenguaje Python donde se describen y se ejecutan las tareas llevadas a cabo.

Los cuatro notebooks son:

- Data_prep_und.ipynb: Que contiene el código de entendimiento, limpieza y transformaciones necesarias ejecutadas sobre los datos para los diferentes modelos creados.
- hac_model.ipynb: Código necesario para la ejecución y evaluación del agrupamiento jerárquico aglomerativo.

- DBSCAN_model.ipynb: Código necesario para la ejecución y evaluación del agrupamiento basado en densidad.
- log_reg_model.ipynb: Código necesario para el entrenamiento, evaluación y serialización del modelo de clasificación.

Por último, se ha obtenido una aplicación web que permite al usuario clasificar individuos en base a las características de estos en los clústeres previamente generados mediante el modelo aglomerativo.

1.6 Breve descripción de los otros capítulos de la memoria

El capítulo 2 contiene el estado del arte donde se analiza la situación actual del aprendizaje automático no supervisado mediante el análisis de trabajos e investigaciones relevantes relacionados con la temática de este trabajo.

A continuación se recopilan los datos con los que se trabajará durante el estudio, estos provienen de los años 2011 a 2016, se explica el proceso de selección de variables y sus características principales, se realiza un proceso de limpieza y un análisis para el entendimiento de estos, finalmente se transforma para adaptarlos a la posterior aplicación de los algoritmos de aprendizaje automático.

En este mismo capítulo se crean los modelos de agrupamiento, uno jerárquico aglomerativo y un segundo basado en densidad, el DBSCAN, se parametrizan, se seleccionan métodos y métricas según los algoritmos a aplicar, a continuación se seleccionan los que presentan mejor comportamiento en el descubrimiento de patrones según indices analizados durante el proceso de evaluación.

El capítulo 3 explica el proceso de creación de una aplicación web destinada a la clasificación de individuos en clústeres según el conocimiento obtenido de la agrupación anterior que ha permitido el etiquetado del conjunto de datos. Con estos datos etiquetados se entrena un modelo de clasificación basado en regresión logística, se empotra el modelo entrenado en la aplicación y finalmente se publica para facilitar su acceso.

En el capítulo 4 se exponen las conclusiones finales sobre los resultados obtenidos y el proceso de ejecución de este estudio.

En los siguientes capítulos se incluye un glosario de términos usados en el documento, una bibliografía que relaciona las fuentes consultadas y anexos que incluye porciones de código generado durante la ejecución del estudio.

2. Diseño e implementación

2.1 Estado del arte

Al igual que en muchas otras áreas el aprendizaje automático está teniendo un gran impacto en la medicina. Su campo de aplicación es también muy amplio y abarca desde la predicción y diagnóstico de enfermedades mediante el análisis de imágenes, al desarrollo de tratamientos terapéuticos personalizados y automáticos para pacientes, la creación de nuevos medicamentos, recomendadores que ayudan en la toma de decisiones o predicción de epidemias. Esto implica que son múltiples los trabajos de investigación que se publican actualmente sobre la temática de aprendizaje automático en medicina, y son muchas también las instituciones, gobiernos y empresas privadas que patrocinan este tipo de estudios con el objetivo de avanzar y mejorar en un campo que está en pleno crecimiento.

Estos trabajos, principalmente, se basan en conjuntos de datos recopilados de pacientes de muy diferente tipo y forma de recolección, estos pueden ser los resultados de tests químicos de laboratorio, entrevistas, historial médico, datos genéticos o exámenes físicos. Este trabajo utiliza para su estudio el conjunto de datos NHANES, que es un ambicioso programa que trata de evaluar la salud de los estadounidenses a través de entrevistas, exámenes y test que realiza desde los años 60 y de forma continuada desde 1999. Dado su formato, extensión y calidad el conjunto NHANES es objeto de múltiples investigaciones, de muy variados objetivos, metodologías y técnicas de análisis que aplican sobre él. Según NHANES, enfermedades o indicadores de salud sobre los que se puede llevar a cabo diferentes estudios pueden ser, por ejemplo, enfermedades cardiovasculares, infecciosas, diabetes o indicadores como la actividad física. NHANES ha facilitado a los científicos hacer descubrimientos importantes a lo largo de su historia.

Por ejemplo, el artículo publicado en la revista Preventing Chronic Disease [4] analizando datos NHANES en tres periodos: 1988-1994, 1999-2006 y 2007-2012, examina la tendencia y prevalencia del síndrome metabólico por raza y sexo y tal como indican es uno de los mayores estudios realizados en cuanto a que abarca casi tres décadas. Consiguen los autores unos resultados que indican que la prevalencia es de 25.3% en el primer periodo, se reduce en el segundo y crece de forma sustancial en el tercero hasta alcanzar un 34.2%, dándose esta característica en ambos sexos y razas e indicando que no declina en ninguno de ellos, el mayor crecimiento se produjo en varones negros no hispanos, con un 55%, seguido de mujeres no hispanas blancas con

44% y no hispanas negras con 41%. Observan también que no solo se da el síndrome en personas obesas sino también en no obesas y que la prevalencia crece rápidamente con la edad y que se espera un aumento en el desarrollo de enfermedades crónicas relacionadas.

En línea con la mejora de enfoques de investigación y técnicas, basado también en el conjunto de datos NHANES, el siguiente artículo [5] utiliza en este caso datos del cuestionario dietético durante el periodo 2005-2006 para la predicción de factores de riesgo de enfermedades cardiovasculares, el objetivo del trabajo, más que la predicción de los niveles de triglicéridos o tipos de colesterol en sangre deseables que puedan ser usados como recomendaciones de hábitos saludables de consumo, es la propuesta de un nuevo enfoque a aplicar sobre los datos de dietas y lo hacen mediante el uso de un modelo LASSO [6] nunca antes aplicado sobre este tipo de datos. Los resultados indican que con el modelo de regresión se obtuvieron mayores tasas de precisión en la predicción que con los tradicionales de regresión lineal.

Más en la línea de interés de este trabajo sobre el uso de métodos no supervisados se encuentran estudios con enfoques y tareas diferentes como la combinación de métodos supervisados y no supervisados. Como ejemplo, la siguiente investigación [7] que pretende extraer conocimiento sobre la diabetes y alta presión en sangre localizando y cuantificando relaciones entre sus síntomas. A partir de los datos NHANES 1999-2008 adapta técnicas de clustering y reglas de asociación: en el caso de reglas de asociación modifican el algoritmo apriori y en cuanto a la agrupación jerárquica proponen un original definición de distancia entre estados de salud basado en frecuencias de co-ocurrencia de valores “Sí” entre indicadores. La idea sería, como indican que si hay dos síntomas y muchos encuestados sufren de ambas entonces estos síntomas están relacionados.

Los autores confirman la confianza en la aplicación de técnicas de minería de datos sobre datos médicos, la necesidad del conocimiento del dominio de los datos a la hora de aplicar estas técnicas, y un proceso iterativo de desarrollo similar al que se define en este trabajo.

Se evalúa a continuación el artículo “Using Total Correlation to Discover Related Clusters of Clinical Chemistry Parameters” [8] en el cual los autores proponen un nuevo enfoque con el uso de la correlación total en los casos de elevada dimensión o que la relación entre variables sea no lineal. En este caso también, sobre un cliclo de datos NHANES (2011-2012) formado por 39 variables de parámetros químicos. Los resultados muestran en la aplicación del agrupamiento jerárquico usando y sin usar la

correlación total como medida de similitud entre dos clústeres, que a pesar de tener estructuras globales diferentes mantienen gran parte de las estructuras internas mostradas en los dendrogramas.

En cuanto al agrupamiento basado en densidad, existen muchas variaciones sobre el algoritmo DBSCAN [9], entre ellas OPTICS (Ordering Points to Identify the Clustering Strusture) [10] y DENCLUE (DENsity-based CLUstEring) [11]. DBSCAN [9] destaca por su eficiencia en la localización de clústeres de forma arbitraria formando clústeres de alta densidad de objetos separadas por regiones de baja densidad. En la propuesta de desarrollo del algoritmo DVBSCAN (Density Variation Based Spatial Clustering of Applications with Noise) [12] mejoran a DBSCAN [9] en su capacidad de encontrar clústeres que representan regiones uniformes sin estar separadas entre ellas, es decir, sin las regiones de baja densidad que necesita DBSCAN [9] para identificar los diferentes clústeres.

Respecto a la agrupación jerárquica, técnica también usada en el presente trabajo, se comenta en el trabajo Modern hierarchical, agglomerative clustering algorithms [13], el uso ineficiente de algunos algoritmos implementados en diferentes paquetes estadísticos como hclust o agnes de R. Evalúa a continuación los más eficientes y aconseja cuáles usar, resultando como más óptimos el algoritmo MST-algorithm con “single linkage”, el NN-chain con los métodos “complete”, “average”, “weighted” y “ward” y el algoritmo genérico de clustering con métodos “centroid” y “median”

En otra comparación [14], en este caso sobre implementaciones de paquetes con algoritmos para clustering en software científico como R o Python, scipy.cluster.hierarchy, hclust, o flashclust, con el paquete fastcluster. Comenta que estas implementaciones ofrecen rendimientos más bajos que el comentado flastcluster que por otra parte tiene interfaces para su uso en R y Python.

En el siguiente trabajo [15] se presenta un procedimiento interesante para la exploración de datos multidimensionales con el objetivo de revelar la estructura subyacente de los datos para caracterizar e identificar subgrupos de pacientes así como obtener conclusiones que puedan ser descubiertas. En un dominio semejante al de este trabajo, sobre un conjunto de datos de 515 pacientes con 40 atributos nominales y numéricos, el procedimiento consiste en 1) Normalización de datos, 2) PCA (Principal component analysis, 3) Detección de outliers y 4) clustering jerárquico. Este último en dos etapas: primero un clúster sobre datos numéricos y a continuación un clúster que incluye la incorporación de los datos nominales.

No menos importante es la validación de las particiones que se han generado tras la aplicación de los algoritmos sobre los datos, para ello se suelen usar una serie de índices, CVI (cluster validation index), que miden la compacidad y separación de estas. En el siguiente trabajo, "An extensive comparative study of cluster validity indices" [16], se realiza una extensiva comparación de 30 índices. Los resultados indican que hay diferencia estadísticamente significante entre los diferentes CVI, resultando los indices de mejor comportamiento: Silhouette, Davies–Bouldin, Calinski–Harabasz, generalized Dunn, COP and SDbw. El solapamiento, el ruido en los datos así como el número de dimensiones afecta a los resultados de los CVI.

Estos trabajos, entre otros muchos no incluidos, dan una idea de las posibilidades de investigación que ofrece el conjunto de datos NHANES objeto de este trabajo, así como la idoneidad de la elección de métodos no supervisados para la extracción de conocimiento de ellos, se observa que ofrecen buenos resultados y son ampliamente usados.

2.2 Entendimiento de los datos

2.2.1 Recolección de datos del conjunto NHANES

Los datos necesarios para el estudio han sido obtenidos de la web del programa NHANES [1], esta pone a disposición de forma pública un basto conjunto de datos recopilados desde los años 60 hasta la actualidad, que se publican en ciclos de dos años desde 1999. Para este trabajo se seleccionan una serie de variables de los tres ciclos más recientes publicados: 2011-2012, 2013-2014 y 2015-2016.

EL conjunto final de datos obtenidos consiste en un total 29902 observaciones (tabla 3) y 25 variables. La selección de las variables (tabla 4) se ha basado en estudios y trabajos previos analizados que tratan y estudian también la patología diabetes [7] [17] [18] [19] [20] y se ha llevado a cabo, para este trabajo, una combinación de las variables que se estudian en estos trabajos y a la vez están presentes en NHANES [1]. El objetivo es obtener los mejores resultados posibles en el perfilado de este tipo de pacientes.

El proceso de recolección consiste, para cada uno de los ciclos, desde la página web de NHANES [1], en localizar el componente y los archivos de este componente que contienes las variables necesarias, a continuación mediante el software SAS Studio se seleccionan variables de interés, se agrupan por ciclos y estos se agrupan para formar el conjunto final objetos de este estudio, finalmente se exportan a un archivo con extensión csv para su posterior tratamiento.

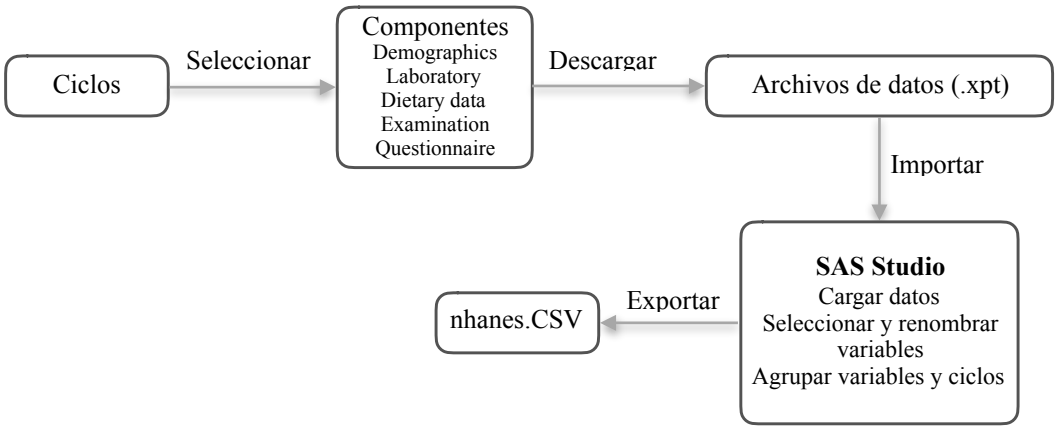


Fig 3. Diagrama de recopilación de datos NHANES

Tabla 3: Número de observaciones.

Ciclo	Número de observaciones
2011-2012	9756
2013-2014	10175
2015-2016	9971
Total	**29902**

Tabla 4: Información sobre las variables seleccionadas.

Listado alfabético de variables y atributos					
Nº	**Variable**	**Tipo**	**Long**	**Etiqueta**	**Valores**
1	**SEQN**	Num	8	Respondent sequence number	**62161 - 93702**
2	**BMXBMI**	Num	8	Body Mass Index (kg/m**2)	**11.5** to **67.3 ; .** :Missing
3	**BMXWAIST**	Num	8	Waist Circumference (cm)	**40** to **171.6 ; .** :Missing
4	**BPXDI1**	Num	8	Diastolic: Blood pres (1st rdg) mm Hg	**0** to **120 ; .** :Missing
5	**DIQ010**	Num	8	Doctor told you have diabetes	**1**:Yes **2**:No **3**: Borderline **7**:Refused **9**:Don't know **.** :Missing
6	**LBXGLU**	Num	8	Fasting Glucose (mg/dL)	**21** to **479; 0:** No lab result **.**: Missing
7	**LBXIN**	Num	8	Insulin (uU/mL)	**0.71** to **324.06 ; .** :Missing
8	**HIQ011**	Num	8	Covered by health insurance	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
9	**HOD050**	Num	8	Number of rooms in home	**1-12**, **13** or more., **777**: Refused **999**: Don't know **.** :Missing
10	**HOQ065**	Num	8	Home owned, bought, rented, other	**1**:Owned or being bought **2**:Rented **3**:Other arrangement **7**:Refused **9**:Don't know **.** :Missing
11	**HSD010**	Num	8	General health condition	**1**:Excellent **2**:Very good **3**: Good **4**:Fair **5**:Poor? **7**:Refused **9**:Don't know **.** :Missing
12	**INQ020**	Num	8	Income from wages/salaries	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
13	**MCQ010**	Num	8	Ever been told you have asthma	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
14	**MCQ300C**	Num	8	Close relative had diabetes?	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
15	**PAQ650**	Num	8	Vigorous recreational activities	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
16	**PAQ665**	Num	8	Moderate recreational activities	**1**:Yes **2**:No **7**:Refused **9**:Don't know **.** :Missing
17	**PAQ710**	Num	8	Hours watch TV or videos past 30 days	**0**: Less than 1 hour **1-4**: n_hours **5**: 5 or more hours **8**: don't/doesn't watch videos **77**:Refused **99**:Don't know **.** :Missing

18	**PAQ715**	Num	8	Hours use computer past 30 days	**0**: Less than 1 hour **1-4**: range **5**: 5 or more hours **8**: don't/doesn't use computer **7**:Refused **9**:Don't know **.** :Missing
19	**RHQ160**	Num	8	How many times have been pregnant?	**1-10**: Range **11**: 11 or more **77**: Refused **99**: Don't know
20	**SLD012**	Num	8	How much sleep do you get (hours)?	**2** to **14.5 77**: Refused **99**: Don't know
21	**SMQ040**	Num	8	Do you now smoke cigarettes	**1**:Every day **2**:Some days **3**: Not at all **7**:Refused **9**:Don't know **.** :Missing
22	**RIAGENDR**	Num	8	Gender	**1**: Male **2**: Female **.**:Missing
23	**RIDAGEYR**	Num	8	Age in years at screening	**0 to 79**: Range **80**: 80 or more **.** :Missing
24	**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 **.** :Missing
25	**DMDEDUC 2**	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

2.2.2 Preprocesado de los datos

No se ha detectado la existencia de observaciones duplicadas por SEQN en ninguno de los archivos descargados ni en el conjunto concatenado de los diferentes ciclos.

Nuestro conjunto de datos de partida para las siguientes fases es ahora el archivo descrito y creado en el apartado anterior nhanes.csv, como se ha dicho consta de 29902 observaciones y las 25 variables seleccionadas.

Durante el proceso de combinación de las variables seleccionadas en cada ciclo se observa la aparición de valores ausentes (missing) debido a la diferencia de observaciones existente entre los distintos archivos descargados y posteriormente combinados que acaba generando este efecto, por ejemplo, en el ciclo 2013-2014, el archivo del componente laboratorio Plasma Fasting Glucose (GLU_H) cuenta con 3329 observaciones mientras que el archivo Blood Pressure (BPX_H) del componente examen cuenta con 9813 observaciones, esto es debido a la operativa en el desarrollo de la encuesta NHANES y los diferentes exámenes, análisis y fases que en ella se desarrollan donde los participantes pueden o no formar parte de las diferentes pruebas que se realizan.

En la tabla 5 se puede observar el elevado número de valores ausentes en las variables, algunas de ellas con valores sobre el 80% del total en valores ausentes, lo cual hace muy difícil su imputación con garantías.

En la tabla 6 se muestra el número de observaciones según su número de variables con valores ausentes, el número de estas con los datos completos es de 897, conjunto a usar si no se imputaran variables que permitan utilizar un mayor número de las observaciones disponibles en el conjunto de datos.

Tabla 5: Valores ausentes por variable.

Variable	Etiqueta	N Miss	N
SEQN	Respondent sequence number	0	29902
BMXBMI	Body Mass Index (kg/m**2)	3489	26413
BMXWAIST	Waist Circumference (cm)	4724	25178
BPXDI1	Diastolic: Blood pres (1st rdg) mm Hg	8829	21073
DIQ010	Doctor told you have diabetes	1195	28707
LBXGLU	Fasting Glucose (mg/dL)	20725	9177
LBXIN	Insulin (uU/mL)	21007	8895
HIQ011	Covered by health insurance	0	29902
HOD050	Number of rooms in home	499	29403
HOQ065	Home owned, bought, rented, other	499	29403
HSD010	General health condition	11377	18525
INQ020	Income from wages/salaries	501	29401
MCQ010	Ever been told you have asthma	1195	28707
MCQ300C	Close relative had diabetes?	12855	17047
PAQ650	Vigorous recreational activities	9013	20889
PAQ665	Moderate recreational activities	9015	20887
PAQ710	Hours watch TV or videos past 30 days	2167	27735
PAQ715	Hours use computer past 30 days	2167	27735
RHQ160	How many times have been pregnant?	23654	6248
SLD012	How much sleep do you get (hours)?	10981	18921
SMQ040	Do you now smoke cigarettes	22532	7370
RIAGENDR	Gender	0	29902
RIDAGEYR	Age in years at screening	0	29902
RIDRETH3	Race/Hispanic origin w/ NH Asian	0	29902
DMDEDUC2	Education level - Adults 20+	12854	17048

Tabla 6: Número de observaciones por número de variables ausentes.

Nº variables ausentes	Count	Nº variables ausentes	Count
3	4541	**17**	1177
11	3860	**0**	**897**
4	3602	**15**	861

1	3410	**8**	489
2	3384	**13**	435
10	2358	**12**	344
5	1600	**14**	60
7	1524	**9**	48
6	1261	**16**	24

Se detalla a continuación las posibles tareas de imputación a realizar en las variables del conjunto de datos, con ellas se espera aumentar el número de casos a utilizar en la aplicación de los modelos de clustering objeto de este estudio e introducir el menor sesgo posible en los datos finales, la elección de las variables imputables se basa en su menor proporción de datos ausentes, o que por su naturaleza parece lógica una posible imputación como es el caso de la variable RHQ160, número de embarazos, en observaciones con clase hombre:

- En la variable **DIQ010** que indica si el caso tiene o no diabetes se opta por la eliminación de todas las observaciones que tengan este campo ausente ya que no se considera imputable decidir si un individuo es diabético o no.

- Las variables **RHQ160**, nº de embarazos, cuyo Target en la entrevista son mujeres de 20 a 150 años muestra que existen 23564 casos de valores ausente, un 78.8% del total. En este caso se propone la imputación del valor 0 en los casos cuyo género sea hombre.

- La variable **SLD012** que representa las horas que se duermen muestra 10981 casos de valores ausentes. Se propone la imputación de estos casos con la media de las observaciones presentes, siendo esta de 7.197 horas de sueño al día.

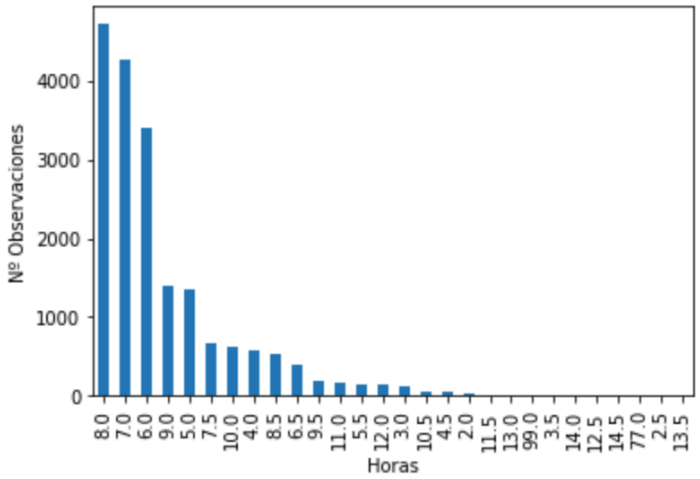


Fig 4. Diagrama de barras de horas de sueño. (Antes de la imputación)

- La variable **INQ020**, si la unidad familiar tiene ingresos o no, tiene 501 observaciones con valores ausentes, se imputará en ellas el valor más frecuente que aparece en los datos y que es que sí tiene ingresos la familia. Dentro del conjunto la existencia de ingresos representa aproximadamente el 80% de los encuestados.

- Sobre las variables **HOD050** y **HOQ065**, número de habitaciones en casa y si esta es en propiedad, alquilada u otra situación, respectivamente, tienen una situación similar a la anterior variable INQ020, con apenas un 1.5% de valores ausentes se propone imputar con el valor de la media en la variable HOD050 y el valor más frecuente en HOQ065.

El resto de variables con valores ausentes no se consideran imputables debido a su elevado número respecto al total, en algunos casos llegando al 70-75%, y dado el caso de aplicar una imputación se produciría una disminución alta de la varianza en estas variables imputadas creando un conjunto de datos muy artificial que ofrecería resultados distorsionados.

Se eliminan a su vez las observaciones que presentan valores como “no sabe”, “rehusa contestar” o “sin resultados de laboratorio” representados en los datos con diferentes valores, tal como se explica en la página web de NHANES [1], como 7, 9, 77, 99 ó 0 según la variable.

La aplicación de las imputaciones descritas crea un conjunto de datos sin variables con datos ausentes de 2481 observaciones. De estas, el mayor aumento en número de observaciones se ha producido al imputar la variable RHQ160, el resto de las imputaciones aplicadas a las variables SLD012, INQ020, HOD050 y HOQ065 apenas han producido un aumento de 20 observaciones para el conjunto final.

El clustering aglomerativo es muy sensible al ruido en los datos y a los valores extremos (outliers), dependiendo según métodos de enlace usado y estructura de los datos, por su parte DBSCAN es robusto a estos valores. Se analiza a continuación la posible existencia de valores extremos en las variables continuas BPXDI1, LBXIN, LBXGLU y BMI.

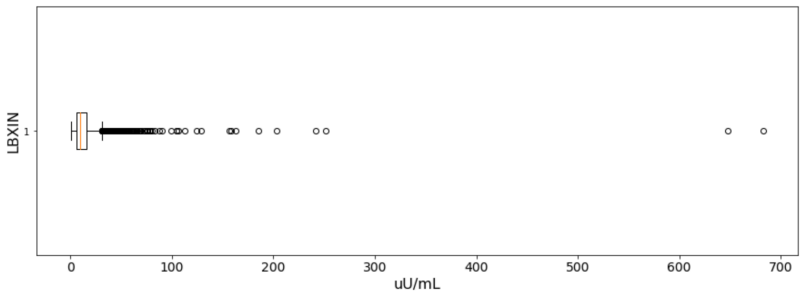


Fig 5. Insulina (LBXIN)

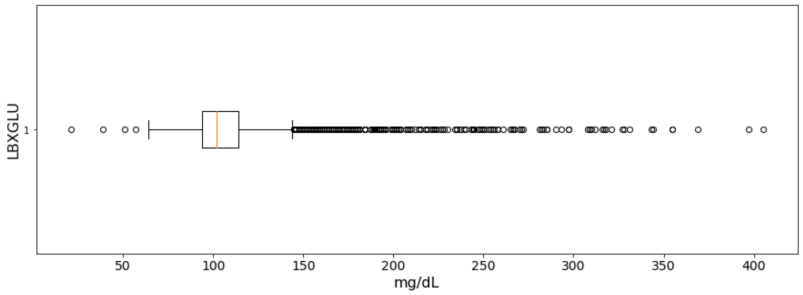


Fig 6. Glucosa **(**LBXGLU)

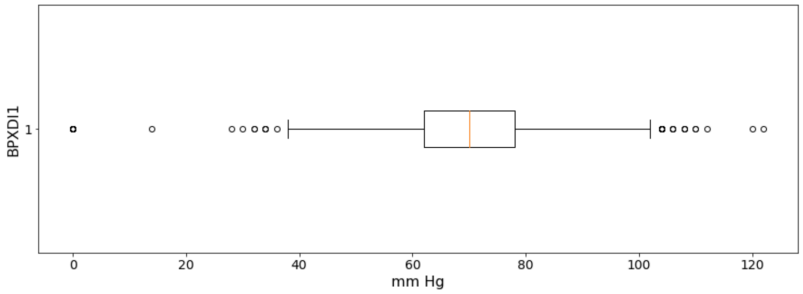


Fig 7. Presión en sangre diastólica (BPXDI1)

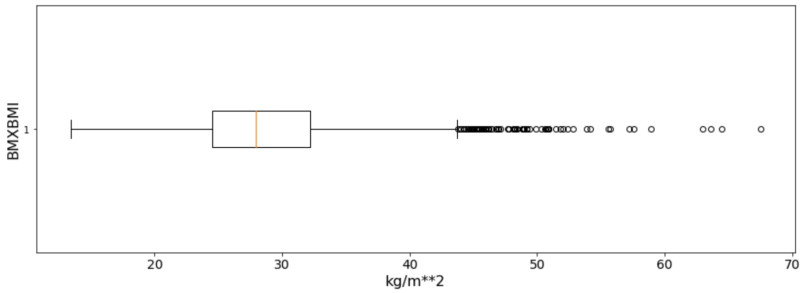


Fig 8. Índice de masa corporal (BMXBMI)

Los gráficos muestran la presencia elevada de observaciones que han sido identificadas como valores extremos en todas las variables, figuras 5 a 8, son diversos factores los que pueden ocasionar la aparición de estos valores, errores en la introducción de los datos, mala codificación, mala calibración de herramientas de análisis o que los encuestados no hayan ayunado antes de las pruebas de glucosa o insulina como es necesario.

Los valores extremos parecen representados de forma consistente y constante en los gráficos y no parece que apuntan a los errores mencionados como la causa, más bien a síntomas de diferentes patologías, se intenta perfilar en este trabajo precisamente a los participantes de la encuesta NHANES y entre ellos a los pacientes diagnosticados de diabetes, y estos últimos debido a su patología pueden presentar valores y mediciones que caen fuera de lo que serían considerados los parámetros normales en una persona sana. Por otro lado, la eliminación de estas observaciones identificadas como valores extremos podría suponer la eliminación de individuos cuyo valor en este estudio puede ser importante y ayudar a identificar patrones de patologías como la diabetes.

Se compara en el gráfico siguiente (fig. 9) si la presencia de "outliers" es debido la patología y se comparan los mismos en grupos de diabéticos, no diabéticos e individuos en el umbral para la variables de laboratorio y examen: Insulina en sangre, LBXIN, presión en sangre BPXDI1y glucosa LBXGLU.

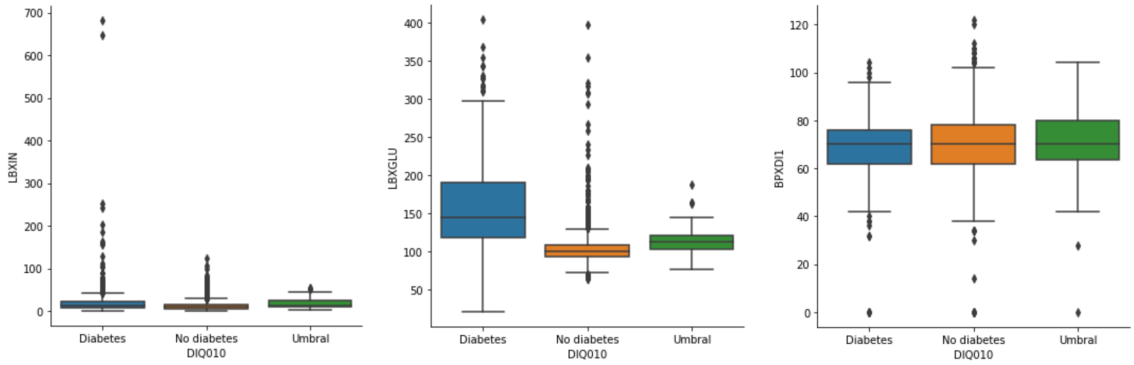


Fig 9. Comparación variables por grupos: LBXIN, LBXGLU, BPXDI1

De los gráficos se desprende que en la variables LBXIN y LBXGLU hay mayor dispersión en enfermos de diabetes o en el umbral y valores ligeramente más altos, sobre todo en los niveles de glucosa (LBXGLU), en la variable BPXDI1 sucede lo contrario, con una muy ligera mayor dispersión en no diabéticos y los que están en el umbral. Los valores extremos se dan en todos los grupos, por lo que no parece su causa el tener diagnosticada la patología.

Teniendo en cuenta lo comentado se usa el método Z-score [21] para la identificación y eliminación de los "outliers", se aplica sobre las variables BPXDI1, LBXIN, LBXGLU y BMI, se sitúa el límite en tres desviaciones típicas que conlleva la eliminación de 131 observaciones, número considerablemente menor que si se utiliza el método IQR [22].

Una vez aplicado el método Z-score y eliminados los outliers, se comparan de forma visual las variables.

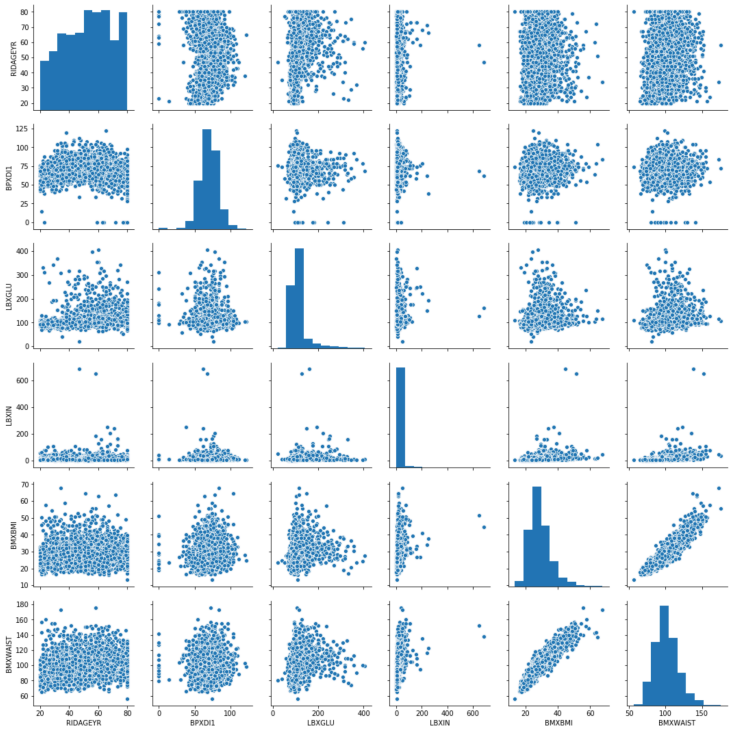


Fig 10. Diagrama de dispersón antes de eliminar “outliers”

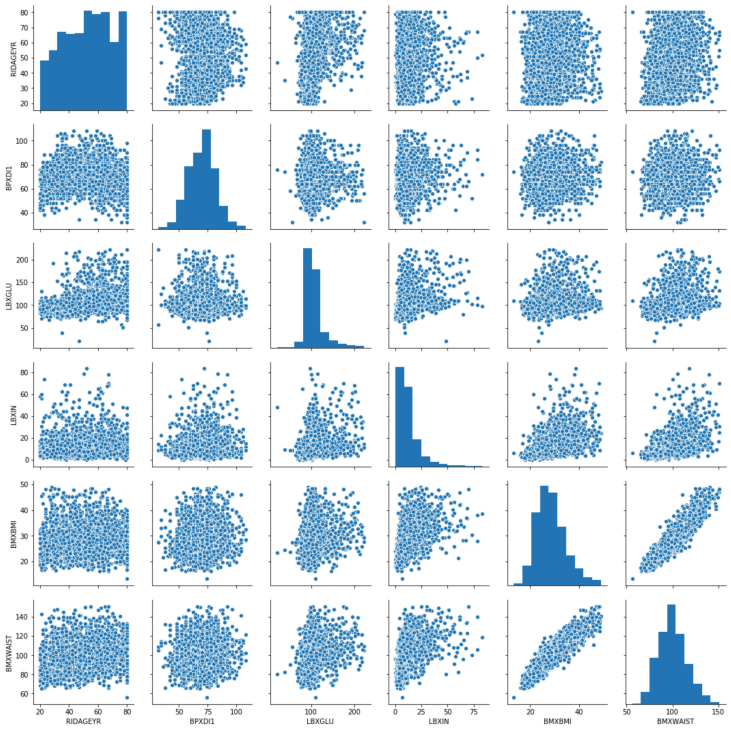


Fig 11. Diagrama de dispersión tras eliminar “outliers”

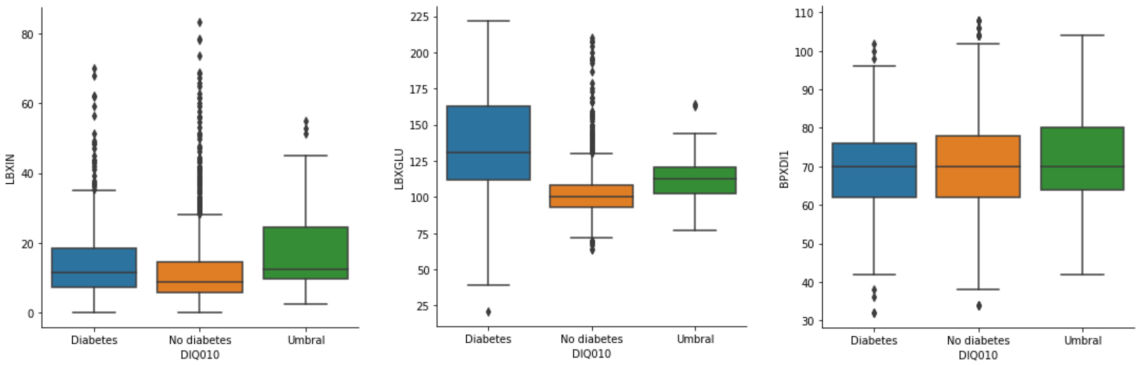


Fig 12. Comparación variables por grupos: LBXIN, LBXGLU, BPXDI1

Se observa comparando las figuras 10 y 11 que los datos han ganado normalidad de forma leve en todas las variables donde se ha aplicado la eliminación de valores extremos. En la figura 12 se puede observar, tras la eliminación de los valores, como han quedado las variables según si son diabéticos, están en el umbral o no lo son.

2.2.3 Comprensión de los datos

Se realiza continuación una descripción de los datos una vez imputadas las variables, eliminadas las observaciones con presencia de outliers y con todas las observaciones restantes completas con la intención de entenderlos sobre su dominio y estudiar las posibles tareas de preparación final antes de aplicar los modelos de clústering. En la tabla 7 se identifican a modo de resumen las características principales de la muestra del análisis descriptivo.

Tabla 7: Resumen de datos descriptivos del conjunto de datos limpios.

Concepto	Valor
Número de observaciones participantes	2350
Participantes a los que un doctor les ha dicho que tienen diabetes	299 (12.7%)
Participantes de género masculino	1521 (64.7%)
Media del número de embarazos de mujeres	3.49
Todas las mujeres participantes han tenido al menos 1 embarazo	
Media del índice de masa corporal de los participantes	28.62
Rango de edad participantes. (Años)	20 a 80
Edad Media participantes. (Años)	52.28
Participantes mayores de 60 años	892 (37.95%)
Participantes que padecen asma	398 (16.93%)
Participantes cubiertos por un seguro médico	1817 (77.32%)
Participante sí tiene pariente cercano con diabetes	1008 (42.89%)
Participantes que practican actividad deportiva de alta intensidad (+10 min)	439 (18.68%)
Participantes que practican actividad deportiva de moderada intensidad (+10 min)	951 (40.47%)
Media de horas que duermen los participantes (horas)	7.08
Participantes no fumadores	1295 (55.1%)
Participantes que consideran no tener un estado de salud bueno	641 (27.27 %)
Media de horas frente a la televisión (últimos 30 días)	2.71
Media de horas frente al ordenador (últimos 30 días)	1.14
La raza blanca es la más numerosa representada	1124 (47.83%)
Media niveles de insulina. (uU/mL)	12.36
Media niveles de glucosa en ayunas (mg/dL)	106.89
Media presión en sangre Diastólica (1st rdg) (mm Hg)	69.78

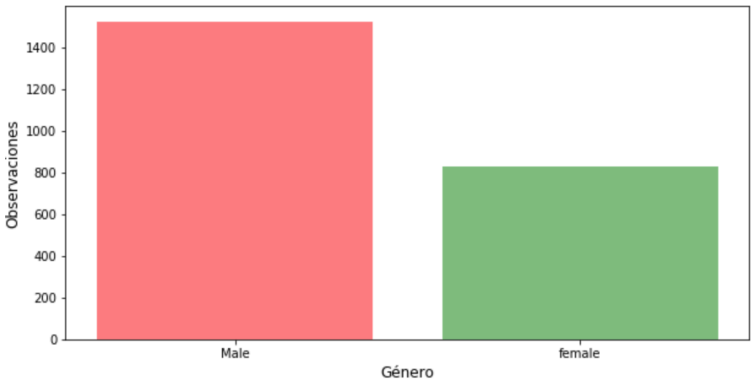


Fig 13. Distribución participantes por género

En la figura 13 se observa que la representación del género hombre es mayor que mujeres, la muestra está formada en un 64.7% por hombres con un total de 1521 observaciones, este aspecto difiere sobre la población real de Estados Unidos según el censo del año 2000 [23] donde ambos géneros se distribuyen al 50% aproximadamente.

Tabla 8: Distribución por raza.

Raza	**Nº Obs. (%)**
Non-hispanic white	1124 (47.82)
Non-hispanic black	474 (20.17)
Other Hispanic	254 (10.8)
Mexican American	251 (10.68)
Non-Hispanic Asian	169 (7.19)
Other Race - Including Multi-Racial	78 (3.32)

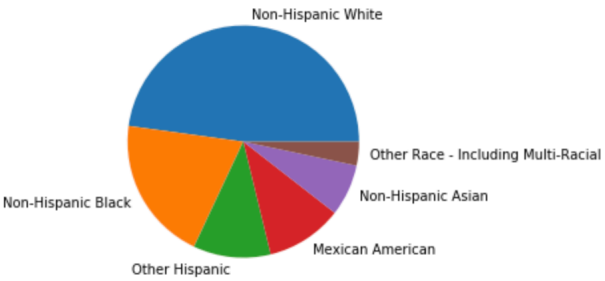


Fig 14. Distribución participantes por raza.

De la tabla 8 y figura 14 se desprende que la población blanca es la más representada seguida de la población negra, en este caso comparado con el censo [23] se ve que la muestra sigue el mismo orden por volumen aunque difieren los porcentajes.

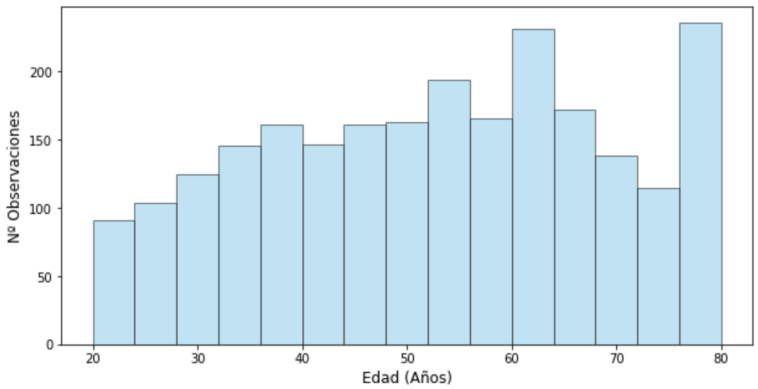


Fig 15. Distribución participantes por edad.

Sobre la distribución de edad de los participantes seleccionados en el conjunto, esta muestra una edad media de 52.28 años, el cual se antoja muy alto debido al hecho de que los participantes mayores de 60 años representan en la muestra el 37.96% de las observaciones, dato muy alto en referencia al censo [23] comentado donde representan alrededor del 16% de la población.

En la figura 16 se observa que los participantes no fumadores son 1295 y representa un 55% del total, el resto son fumadores diarios y ocasionales, juntos son 1055 participantes.

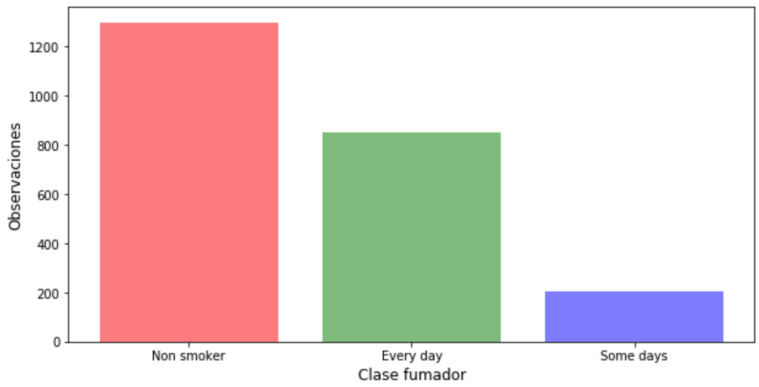


Fig 16. Distribución fumadores.

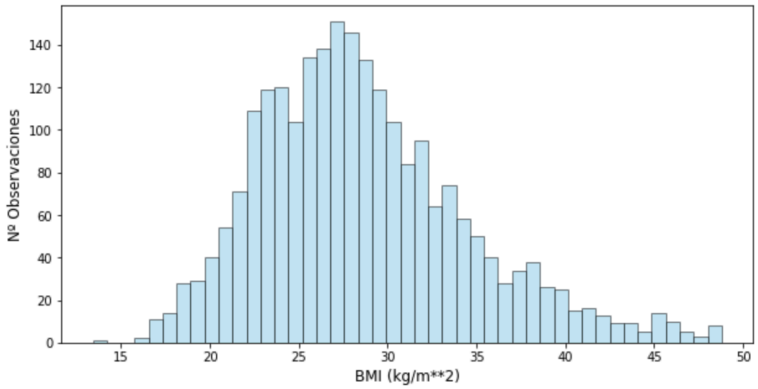


Fig 17. Distribución Índice masa corporal (BMI)

Respecto al índice de masa corporal (BMXBMI) de los participantes (Fig. 17) su media se sitúa en 28.62 kg/m², dato que les situaría en el umbral de la obesidad [24]. Si se compara por rangos de edad, mayores y menores de 40 años, se observa en la figura 18 que tienen distribuciones similares y la media de BMI entre ambos grupos es muy similar también, situándose en 28.06 para menores de 40 años y 28.85 para los mayores de 40 años.

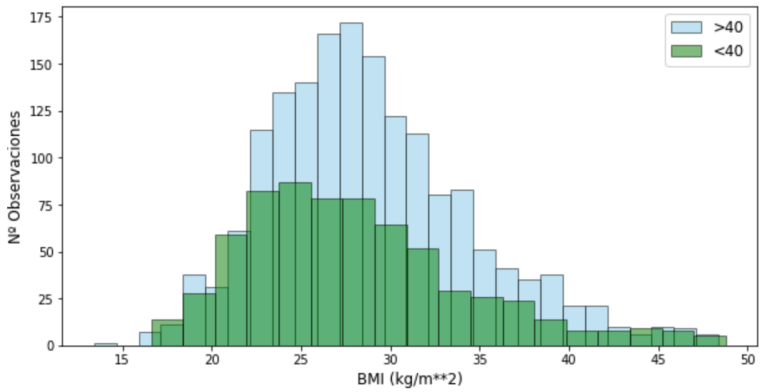


Fig 18. Comparación BMI por grupo de edad: Menores y mayores de 40 años

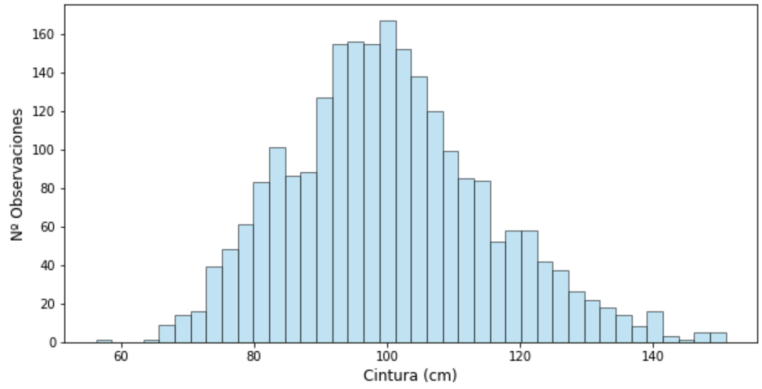


Fig 19. Distribución diámetro cintura (cm)

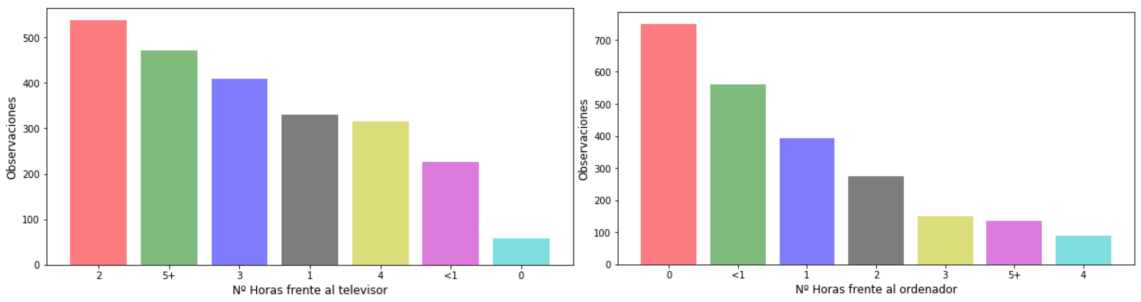


Fig 20. Distribución de nº de horas frente a la televisión (Izq) Ordenador (der)

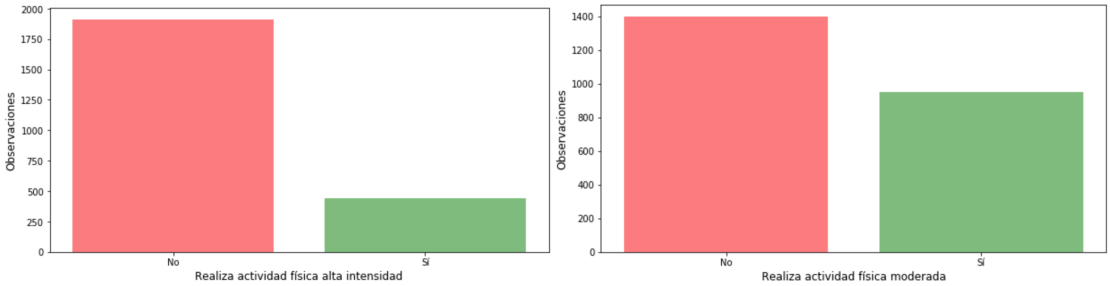


Fig 21. Realiza actividad física de alta intensidad (izq) o moderada (der)

Respecto a los indicadores relacionados con el sedentarismo disponibles, las horas que dedican frente al televisor (fig. 20) varían mayoritariamente entre 2 y 5 o más horas situándose la media en 2.71 horas diarias en los últimos 30 días. Menor es el uso del ordenador, un 32 % de los participantes dice no usarlo y su media se sitúa en 1.14 horas al día. En cuanto a la actividad física que realizan fuera del trabajo (fig 21), de tipo fuerte o moderado, mayoritariamente no realizan ninguna, siendo mayor el número de personas que las realizan de tipo moderado.

Se visualiza a continuación variables de laboratorio referentes a análisis realizados a los participantes, niveles de insulina (LBXIN) y glucosa en ayunas (LBXGLU). Según figuras 22 y 23, ambas muestran sesgo positivo en sus distribuciones. En el caso de la insulina la media se sitúa en 12.36 uU/mL mientras que la glucosa se sitúa en 106.89 mg/dL

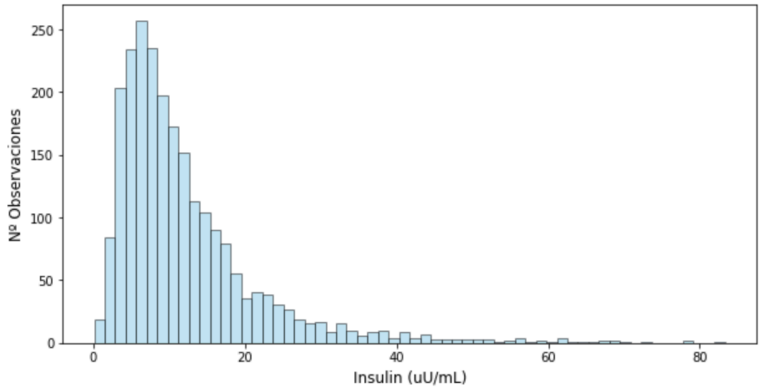


Fig 22. Distribución niveles de insulina (uU/mL)

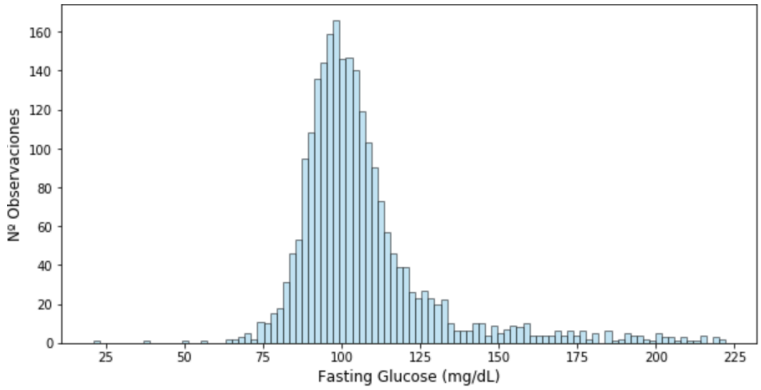


Fig 23. Distribución de niveles de glucosa en sangre. (mg/dL)

Los valores de la presión en sangre (BPXDI1) de los participantes cuentan con una media de 69.78 mm Hg.

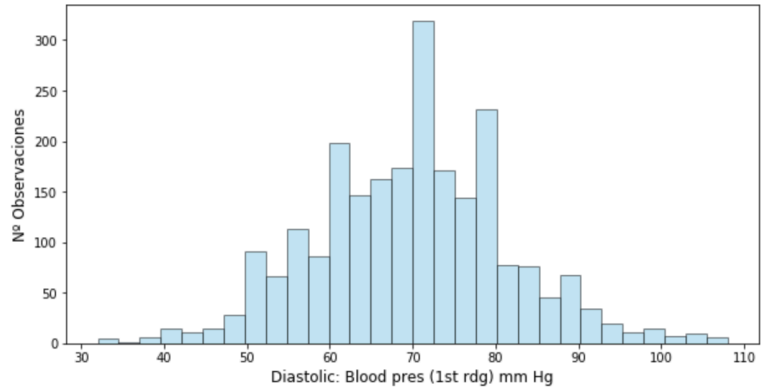


Fig 24. Distribución de presión en sangre (Diastólica 1st rdg) mm Hg

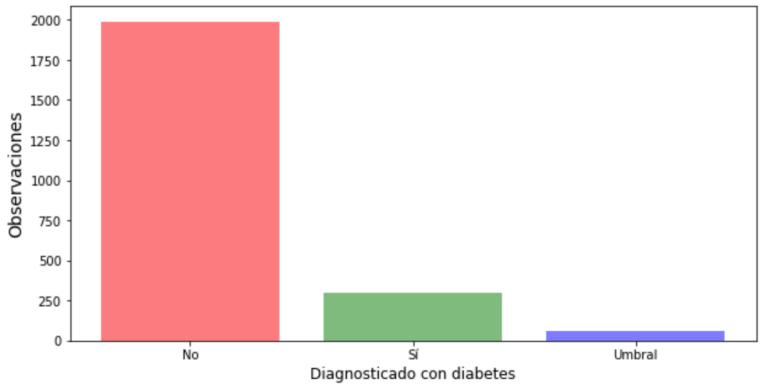


Fig 25. Diagnóstico de diabetes entre participantes

En la figura 25 se observa que 299 de los participantes (12.72%) han sido diagnosticados con diabetes por un doctor, similar porcentaje se produce en el diagnóstico de hombres, un 12.62% y mujeres un 12.9%. La figura 26 muestra si parientes cercanos al participante han sido diagnosticados con diabetes o no, de los cuales un 42.9 % han contestado que sí.

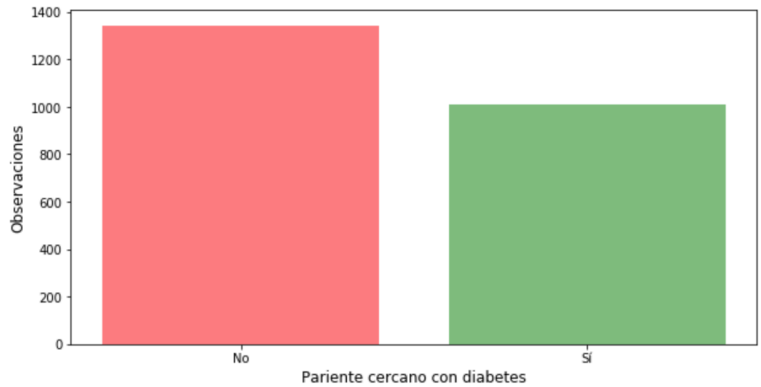


Fig 26. Distribución de participantes con parientes cercanos con diabetes

Si se analiza esta proporción entre observaciones con diagnóstico o no de diabetes, entre los participantes diabéticos, un 70.5 % de estos tienen un pariente cercano diabético mientras que entre los no diabéticos este porcentaje se reduce hasta el 38.3% según figuras 27 y 28 a continuación.

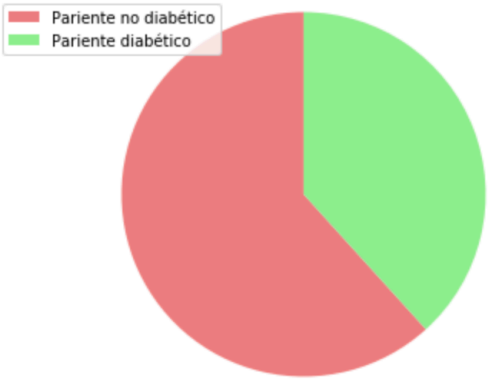


Fig 27. No diabéticos: Proporción de parientes diagnosticados con diabetes

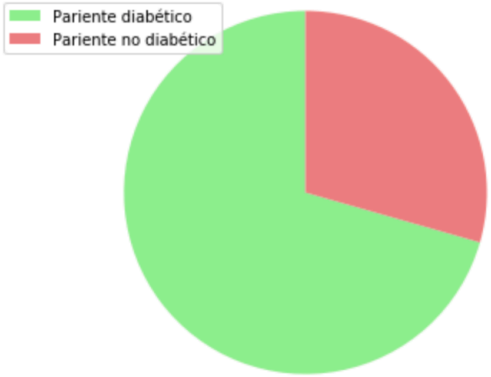


Fig 28. Sí diabéticos: Proporción de parientes diagnosticados con diabetes

Algunos de los indicadores anteriores analizados muestran valores que informan que de media el estado de salud de los participantes no es muy buena o que es mejorable, como es el caso del índice de masa corporal que los sitúa en niveles cercanos a la obesidad, tanto en personas mayores como jóvenes, o el caso de la insulina en ayunas que muestra niveles bastante altos así como de glucosa [25], también la presión diastólica en sangre que aunque su media se sitúa en niveles ligeramente bajos [26], hay en la muestra 436 participantes (18.5%) con medidas por debajo de 60 mm Hg.

Otro aspecto a tener en cuenta es la edad media de los participantes, que como se ha indicado se sitúa en los 52.4 años y que el rango de edad de estos es de los 20 a los 80 años.

2.2.4 Transformaciones

Como paso de preparación de los datos final antes de la aplicación de los modelos y dado que el conjunto de datos está formado por diferentes tipos como datos continuos y categóricos no se puede, en principio, sobre estos calcular la matriz de disimilitud necesaria para el modelo de clúster jerárquico ya que no son comparables usando una métrica como la euclidean, por ejemplo, en la variable raza la codificación usada (tabla 4) no tiene un orden natural que permita determinar si las observaciones están más o menos alejadas en base a esta variable si una persona es blanca y otra asiática o asiática e hispana.

En base a esta problemática se analiza las posibles opciones para el cálculo de la matriz de distancia, una de ellas es la recodificación de las variables categóricas y permitir su comparación cuando sea posible según las características de estas y se pueda establecer un orden/escala natural de medición. De las variables categóricas presentes filtradas del conjunto total en la siguiente tabla se observa qué variables se pueden recodificar estableciendo así una posible cuantificación de sus disimilitudes, se observa que todas las variables excepto en RIDRETH3, raza del individuo y HOQ065, tipo de propiedad de vivienda, permiten establecer una ordenación natural de sus valores, como es el caso de la variable HSD010 que indica el estado general de salud en un rango de 1 a 5 y muestra una mayor distancia entre un estado excelente (1) y uno pobre (5) que un individuo con un estado regular (4) sobre uno con un estado (5) excelente.

Tabla 9: Listado de variables categóricas.

Listado de variables categóricas				
Nº	**Variable**	**Tipo**	**Etiqueta**	**Valores**
1	**DIQ010**	Num	Doctor told you have diabetes	**1**:Yes **2**:No **3**: Borderline
2	**HIQ011**	Num	Covered by health insurance	**1**:Yes **2**:No
3	**HOQ065**	Num	Home owned, bought, rented, other	**1**:Owned or being bought **2**:Rented **3**:Other arrangement
4	**HSD010**	Num	General health condition	**1**:Excellent **2**:Very good **3**: Good **4**:Fair **5**:Poor
5	**INQ020**	Num	Income from wages/salaries	**1**:Yes **2**:No
6	**MCQ010**	Num	Ever been told you have asthma	**1**:Yes **2**:No
7	**MCQ300C**	Num	Close relative had diabetes?	**1**:Yes **2**:No
8	**PAQ650**	Num	Vigorous recreational activities	**1**:Yes **2**:No
9	**PAQ665**	Num	Moderate recreational activities	**1**:Yes **2**:No
10	**SMQ040**	Num	Do you now smoke cigarettes	**1**:Every day **2**:Some days **3**: Not at all
11	**RIAGENDR**	Num	Gender	**1**: Male **2**: Female

12	**RIDRETH3**	Num	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
13	**DMDEDUC2**	Num	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

Las acciones que se realizan en consecuencia para intentar hacer comparables estas variables, RIDRETH3 y HOQ065, es la de una recodificación tipo one_hot_encoding [27] crea 9 nuevos atributos todos ellos de tipo binario (0,1) que forman un vector binario que representa dicha observación.

Sobre la variable DIQ010, si el individuo ha sido diagnosticado con diabetes, no se encuentra en el umbral, se recodifica cambiando el orden del rango para una más natural ordenación estableciendo en Sí: 0, Umbral: 1 y No:2

Se recodifican a su vez las variables binarias con rango (1, 2) a (0,1) para facilitar su lectura.

Debido a las diferentes escalas presente en las diferentes variables se realiza un re-escalado de los datos para transformar estas variables a una escala similar con el objetivo de evitar que variables que se mueven en un rango más grande de valores tengan una mayor influencia que otras en el modelo de clúster.

Se selecciona el escalado MinMax, frente a otros como el escalado estándar, la estandarización z-score o el normalizado de observaciones, debido a que mantiene la forma de la distribución, aun sabiendo que la mayoría de las variables seleccionadas no siguen una distribución normal, y no reduce la influencia de los outliers que no se consideraron para su eliminación como sí hacen otros métodos. Se realiza, por tanto, un escalado de las variables usando el método MinMaxScaler disponible en el módulo sklearn.preprocessing [28], el cual permite para cada variable individualmente llevar a cabo la transformación que establecerá los valores en un rango [0,1] a través de la siguiente fórmula:

$$x_i' = \frac{x_i - min(x)}{max(x) - min(x)}$$

2.3 Modelado

Los métodos de clustering seleccionados para el estudio son el jerárquico aglomerativo y DBSCAN.

La selección del método jerárquico aglomerativo se debe a su flexibilidad respecto a los datos en los que se aplica como de facilidad manejo de las agrupaciones que realiza, es aplicable a cualquier tipo de atributo, aspecto importante debido a la heterogeneidad presente en el conjunto de datos seleccionado, y respecto a su facilidad de uso permite seleccionar y jugar con diferentes tipos de granularidad en las agrupaciones a lo que hay que sumar la facilidad de interpretación de los resultados obtenidos. Destacable es también la variedad de estrategias de enlace como métricas de distancia entre clústeres que permite la parametrización que mejor se ajuste a los datos disponibles.

Sus desventajas serían que no dispone de criterio de parada, esto es, que una vez creada una agrupación, esta ya no se puede mejorar. En caso de aplicación sobre grandes conjuntos de datos su capacidad de interpretación se reduce y pierde eficiencia en cuanto a tiempos de computación.

La elección del método DBSCAN, con un enfoque diferente al anterior método, basado en localizar grupos de alta densidad separados por zonas de baja densidad permite descubrir clústeres de formas y tamaños diferentes. La generación de los clústeres la hace de forma automática, al contrario que en el caso del método jerárquico aglomerativo. Es un algoritmo robusto frente a outliers.

Como desventajas indicar que su rendimiento decrece en conjuntos de datos de alta dimensionalidad, la parametrización a través de sus dos parámetros, MinPtses y Eps, es difícil y le hace muy sensible a estos. En casos de agrupamientos solapados su comportamiento no es bueno así como a la presencia de diferencias de densidad en grupos.

2.3.1 Modelo Jerárquico aglomerativo

Se aplica a continuación el clustering jerárquico aglomerativo [29] sobre los datos previamente preparados, el objetivo es el de agrupar las observaciones que presentan más similitud entre si en grupos lo más homogéneos posible, y a continuación tratar de perfilar los integrantes de los diferentes grupos en base a sus características.

Se usa la métrica o distancia euclídea [30], ampliamente usada sobre atributos numéricos, para computar la similitud entre las características de las observaciones, esta distancia permite evaluar lo similares que son los objetos, cuanto menor es la distancia entre estos más similares serán y al contrario cuanto más alejados estén según la distancia euclídea.

Se evalúa a continuación algunos de los criterios de enlace de los algoritmos jerárquicos aglomerativos [31] en base a la distancia euclídea seleccionada: Complete, average, single, centroid y ward.

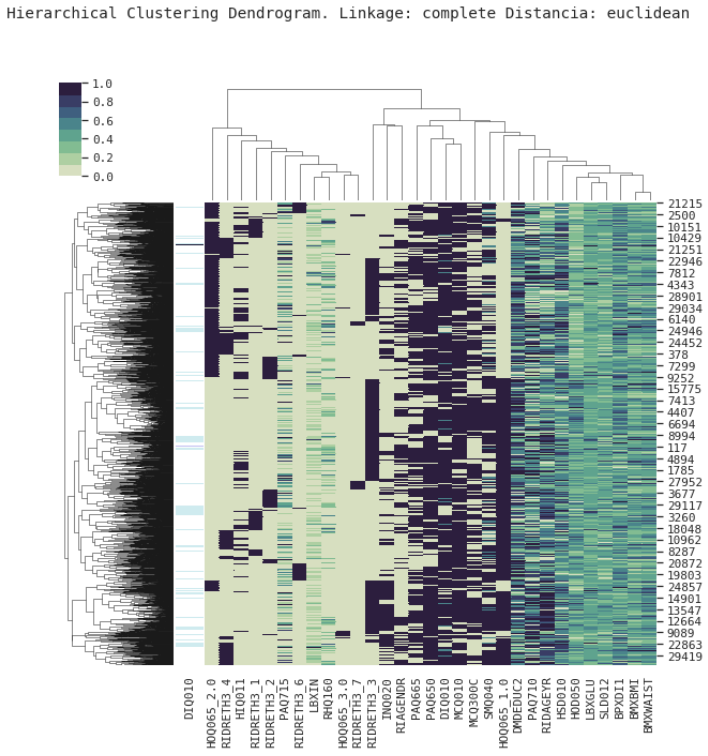


Fig 29. Clustermap: Linkage = Complete Dist = euclidean

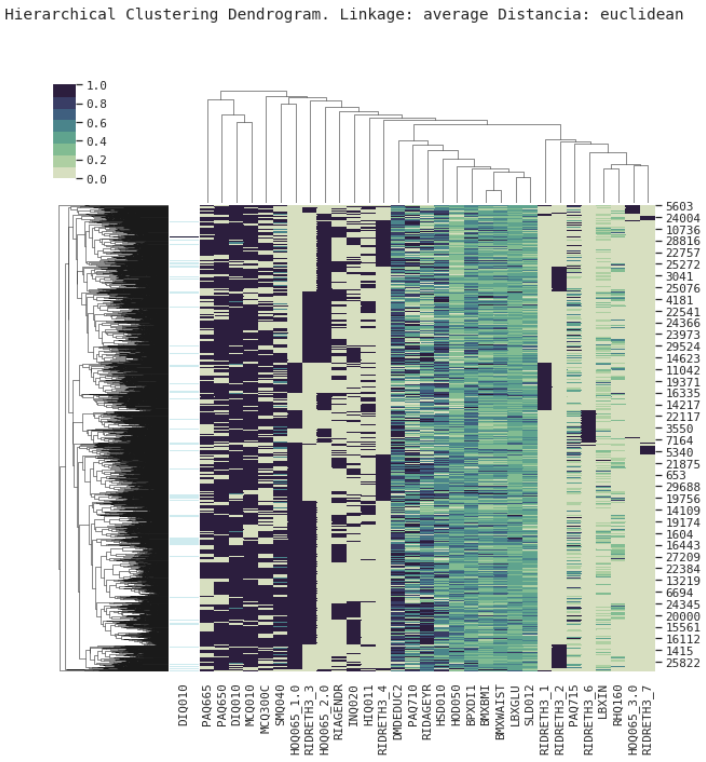


Fig 30. Clustermap: Linkage = Average Dist = euclidean

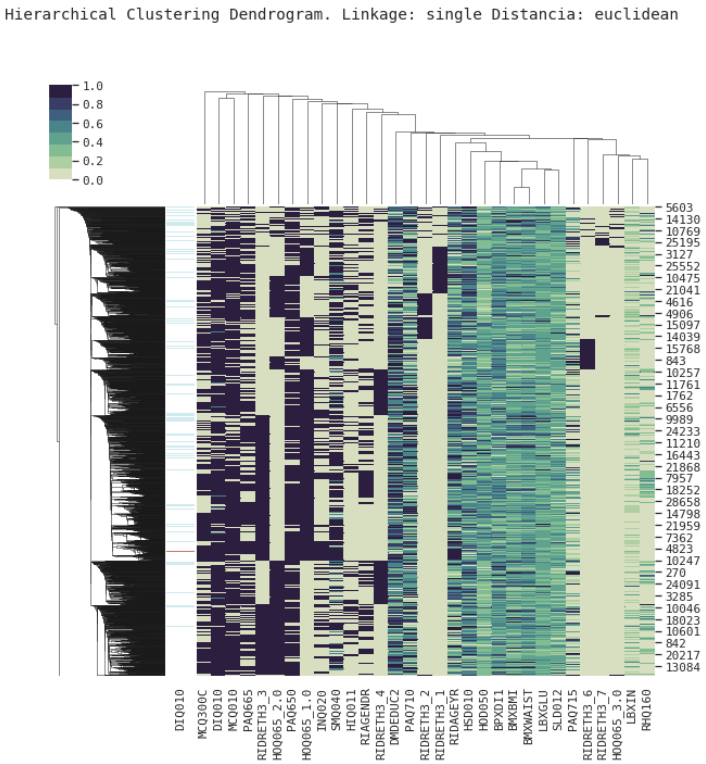


Fig 31. Clustermap: Linkage = Single, Dist = euclidean

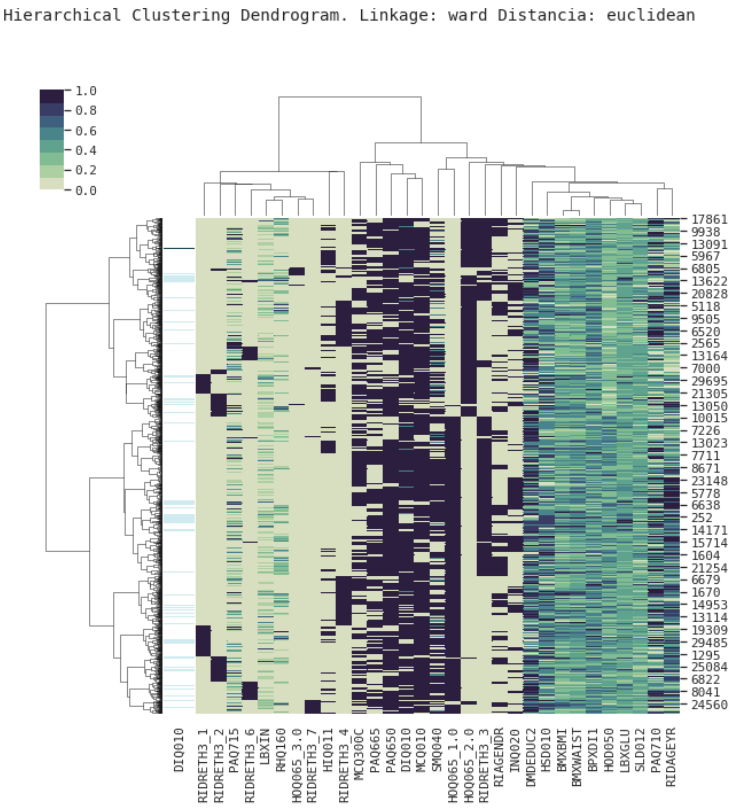


Fig 32. Clustermap: Linkage = ward, Dist = euclidean

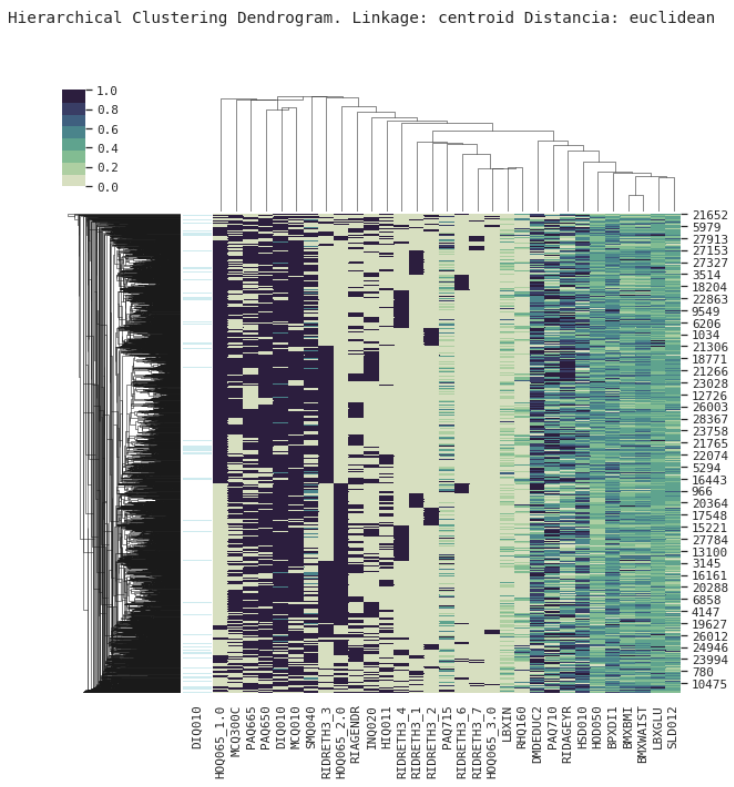


Fig 33. Clustermap: Linkage = centroid, Dist = euclidean

De forma visual se observa a partir de las figuras de los dendrogramas que el más equilibrado y que parece que muestra mayor distancia entre clústeres puede ser el clúster generado por el método Ward (figura 32), mientras que el método centroid no parece mostrar un clustering útil. En cuanto al método single parece mostrar una agrupación desequilibrada en cuanto al número de observaciones asignadas a cada clúster. Estas características comentadas son comunes y debidas a las estrategias de fusión de cada algoritmo usado.

Visualmente es complicado, más con un alta dimensionalidad en los datos, seleccionar un método a partir del dendrograma aglomerativo generado que permita una elección de un mejor clúster, mediante el coeficiente silhouette [32], como método cuantitativo de validación entre clústeres, se evalúa la distancia entre los clústeres generados, se compara a continuación la calidad de estas estructuras para cada método utilizado, permite también seleccionar el número de clústeres óptimo.

Tabla 10: Comparación de algoritmos jerárquicos: Coeficiente silhouette

Método	Distancia	Nº de clústeres óptimo	Coeficiente silhouette
Complete	Euclidean	2	0,134
Average	Euclidean	2	0,146
Single	Euclidean	2	0,180
Ward	Euclidean	2	0,142
Centroid	Euclidean	2	0,147

Se observa en la tabla de comparación que los valores más altos de coeficiente silhouette en todos los casos es para un número de dos clústeres, por otro lado se observa que los coeficientes son en todos los casos muy similares, indicar que ha sido en el método centroid donde han aparecido valores negativos indicando error en la asignación de clúster, aspecto que se intuye en el dendrograma de la figura 33. El coeficiente ligeramente mayor ha sido obtenido con el método single con medida euclídea, aun así el valor de coeficiente cercano a cero, 0.18, con valores en todos los casos muy lejanos de lo aceptable, cercano a 1, indica que las muestras de los clústeres se encuentran muy cercanas al clúster vecino o incluso solapadas con este.

Se comparan los clústeres a continuación a través del coeficiente de correlación cofenético [33][34], este permite observar la fidelidad de la representación en las distancias entre pares mediante la correlación entre estas representadas en el dendrograma respecto a sus distancias originales en los datos.

Tabla 11: Comparación algoritmos jerárquicos: Coeficiente correlación cofenético

Método	Distancia	Coeficiente de correlación cofenético
Complete	Euclidean	0,527
Average	Euclidean	0,625
Single	Euclidean	0,563
Ward	Euclidean	0,546
Centroid	Euclidean	0,468

Se desprende de los coeficientes de correlación de la tabla 11 que el método average es el que menor distorsión genera sobre las distancias originales en el dendrograma y parece representar mejor las distancias originales, aun así se siguen obteniendo valores.

Con un número dos clústeres seleccionado para la segmentación, en cuanto al tamaño de estos obtenido con cada uno de los métodos se observa que es muy diferente de un clúster a otro por método tal y como se podía prever en la visualización de los dendrogramas, los métodos de enlace average, centroid y single forman unos clústeres muy desequilibrados, mientras que los métodos ward y complete los forman de una forma más equilibrada y compacta, aunque esta situación tampoco puede llevar a la conclusión de cuál de los métodos es el idóneo, ward o complete sobre average y single ya que puede que los datos seleccionados no contengan ningún tipo de patrón a identificar y realmente no sea posible segmentar las observaciones como se discutirá más adelante.

Tabla 12: Número de observaciones por clúster según método.

Método	Distancia	Clúster 1 (nº obs)	Clúster 2 (nº obs)
Complete	Euclidean	899	1451
Average	Euclidean	2346	4
Single	Euclidean	2349	1
Ward	Euclidean	946	1404
Centroid	Euclidean	2349	1

Debido a la igualdad entre los métodos de enlace ward y complete, entre los que crean dos particiones analizables, al contrario que en el caso de los métodos average, centroid y single que con una gran partición creada se intuye inútil describir sus clústeres, se selecciona describir y analizar los clústeres creados por el método ward frente a

complete. La selección por tanto se basa, debido a su igualdad, en que visualmente sobre el dendrograma los clústeres parecen más compactos.

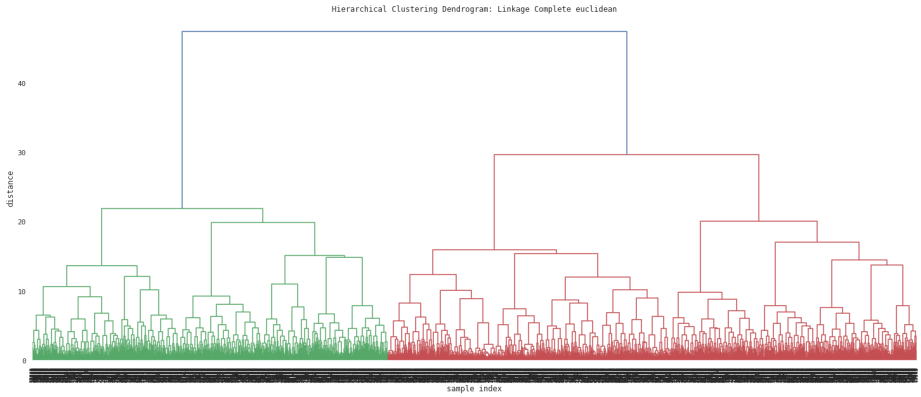


Fig 34. Dendrograma: Linkage = Ward, Dist = euclidean

Se describe a continuación los clústeres creados por el método ward, para ello se usará el coeficiente de variación (cv) [35], medida de dispersión que permitirá determinar qué variables describen mejor cada uno de los clústers a través de sus diferencias.

Tabla 13: Comparación coeficiente de variación variables clústeres 1 y 2.

Nº	Variable	CV Clúster 1	CV Clúster 2
2	**BMXBMI**	21,40 %	20,10 %
3	**BMXWAIST**	15,48 %	14,74 %
4	**BPXDI1**	16,65 %	16,90 %
5	**DIQ010**	17,54 %	21,46 %
6	**LBXGLU**	20,24 %	21,42 %
7	**LBXIN**	79,83 %	83,61 %
8	**HIQ011**	35,35 %	31,65 %
9	**HOD050**	35,49 %	28,46 %
10	**HOQ065**	10,84 %	13,03 %
11	**HSD010**	31,53 %	32,17 %
12	**INQ020**	35,17 %	35,14 %
13	**MCQ010**	21,69 %	19,63 %
14	**MCQ300C**	31,07 %	31,78 %
15	**PAQ650**	22,27 %	20,96 %
16	**PAQ665**	29,91 %	31,31 %
17	**PAQ710**	61,16 %	55,71 %
18	**PAQ715**	126,35 %	119,22 %
19	**RHQ160**	169,02 %	163,59 %

20	**SLD012**	23,58 %	21,02 %
21	**SMQ040**	47,41 %	39,02 %
22	**RIAGENDR**	35,33 %	35,31 %
23	**RIDAGEYR**	35,45 %	27,76 %
24	**RIDRETH3**	40,71 %	43,25 %
25	**DMDEDUC2**	37,40 %	34,89 %

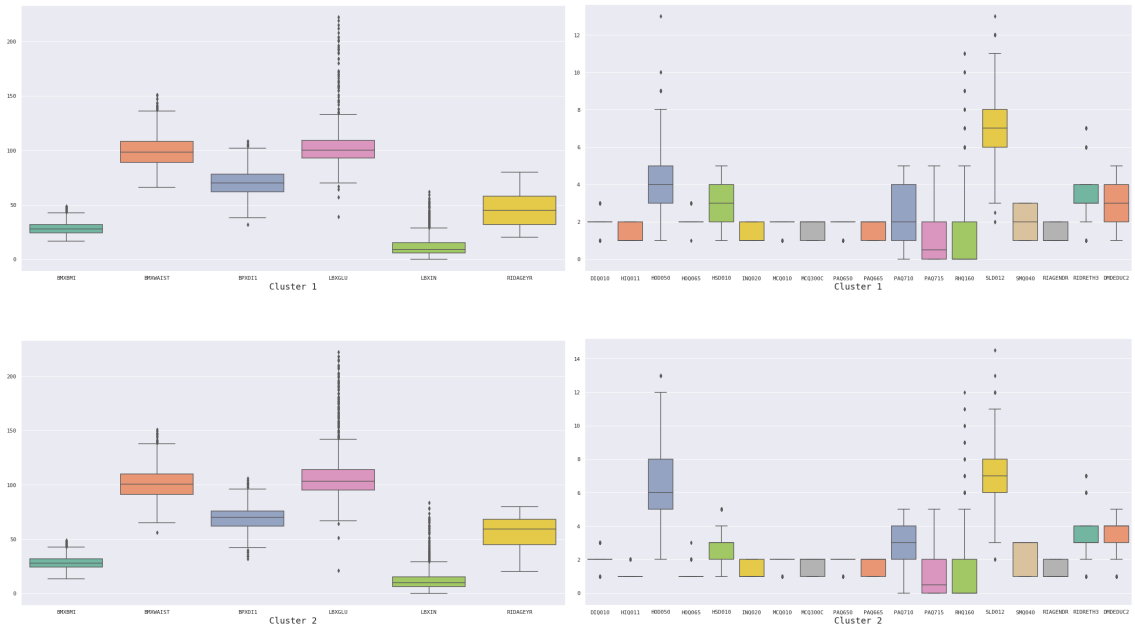


Fig 35. Comparación dispersión de variables entre clústeres. (2 clústeres)

Se observa, tanto en la tabla 13 como en la figura 35, que las variables en ambos clústeres tienen un comportamiento similar, no existe apenas diferencias en la dispersión entre las variables comparadas vis-a-vis entre los dos clústeres, se hace difícil identificar qué variables han podido suponer una mayor importancia en la segmentación creada. Destacan ligeramente sobre el resto de variables por su diferencia entre clústeres, PAQ710, HOD050, HIQ011, HOQ065 Y DMEDUC2, presentando todas ellas una menor dispersión en el clúster 2 a excepción de HOQ065 que muestra mayor dispersión en este clúster número 2.

Respecto al perfil de las observaciones en los clústeres cabe destacar, dentro de la similitud que se observa, que un 95% de los integrantes del clúster 1 viven en régimen de alquiler mientras que un 98% de los integrantes del clúster 2 tiene casa en propiedad (HOQ065). El resto de variables tienen comportamientos muy similares que no

permiten perfilar al individuo sobre ninguna otra característica del conjunto de estudio como se observar en tablas 14 y 15.

Tabla 14: Descripción estadísticos de clúster 1

	SEQN	BMXBMI	BMXWAIST	BPXDI1	DIQ010	LBXGLU	LBXIN	HIQ011	HOD050	HOQ065	HSD010	INQ020
count	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000
mean	78050.021142	28.597780	99.238795	70.454545	1.931290	104.310782	12.020359	1.324524	4.517970	2.049683	3.024313	1.289641
std	9234.771020	6.124573	15.374896	11.738263	0.338896	21.125191	9.601027	0.468445	1.604291	0.222218	0.954211	0.453835
min	62172.000000	16.400000	65.800000	32.000000	1.000000	39.000000	0.140000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	69338.750000	24.300000	88.725000	62.000000	2.000000	93.000000	5.815000	1.000000	3.000000	2.000000	2.000000	1.000000
50%	78468.000000	27.700000	98.300000	70.000000	2.000000	100.000000	8.860000	1.000000	4.000000	2.000000	3.000000	1.000000
75%	86019.000000	31.900000	108.225000	78.000000	2.000000	109.000000	15.085000	2.000000	5.000000	2.000000	4.000000	2.000000
max	93695.000000	48.400000	151.000000	108.000000	3.000000	222.000000	61.930000	2.000000	13.000000	3.000000	5.000000	2.000000

MCQ010	MCQ300C	PAQ650	PAQ665	PAQ710	PAQ715	RHQ160	SLD012	SMQ040	RIAGENDR	RIDAGEYR	RIDRETH3	DMDEDUC2
946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000	946.000000
1.809725	1.585624	1.799154	1.621564	2.655391	1.213531	1.289641	7.066056	1.987315	1.348837	46.278013	3.191332	3.160677
0.392726	0.492875	0.400845	0.485254	1.624924	1.534046	2.180879	1.666768	0.942724	0.476854	16.415372	1.299723	1.182580
1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	2.000000	1.000000	1.000000	20.000000	1.000000	1.000000
2.000000	1.000000	2.000000	1.000000	1.000000	0.000000	0.000000	6.000000	1.000000	1.000000	32.000000	3.000000	2.000000
2.000000	2.000000	2.000000	2.000000	2.000000	0.500000	0.000000	7.000000	2.000000	1.000000	45.000000	3.000000	3.000000
2.000000	2.000000	2.000000	2.000000	4.000000	2.000000	2.000000	8.000000	3.000000	2.000000	58.000000	4.000000	4.000000
2.000000	2.000000	2.000000	2.000000	5.000000	5.000000	11.000000	13.000000	3.000000	2.000000	80.000000	7.000000	5.000000

Tabla 15: Descripción estadísticos de clúster 2.

	SEQN	BMXBMI	BMXWAIST	BPXDI1	DIQ010	LBXGLU	LBXIN	HIQ011	HOD050	HOQ065	HSD010
count	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000
mean	78662.042023	28.637536	101.025570	69.331909	1.878205	108.629630	12.589195	1.160969	6.449832	1.016382	2.881766
std	9047.940565	5.759339	14.900393	11.723147	0.403277	23.277425	10.529387	0.367633	1.836483	0.132478	0.927282
min	62218.000000	13.400000	56.200000	32.000000	1.000000	21.000000	0.140000	1.000000	2.000000	1.000000	1.000000
25%	70329.000000	24.400000	91.000000	62.000000	2.000000	95.000000	6.077500	1.000000	5.000000	1.000000	2.000000
50%	79476.500000	27.900000	100.300000	70.000000	2.000000	103.000000	9.755000	1.000000	6.000000	1.000000	3.000000
75%	86171.750000	31.800000	109.925000	76.000000	2.000000	114.000000	15.390000	1.000000	8.000000	1.000000	3.000000
max	93655.000000	48.800000	150.600000	106.000000	3.000000	222.000000	83.340000	2.000000	13.000000	3.000000	5.000000

INQ020	MCQ010	MCQ300C	PAQ650	PAQ665	PAQ710	PAQ715	RHQ160	SLD012	SMQ040	RIAGENDR	RIDAGEYR
1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000	1404.000000
1.285613	1.844729	1.561254	1.822650	1.577635	2.750712	1.101852	1.190171	7.091591	2.325499	1.355413	56.325499
0.451867	0.362292	0.496411	0.382101	0.494112	1.532901	1.314141	1.947681	1.490966	0.907813	0.478809	15.642802
1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	2.000000	1.000000	1.000000	20.000000
1.000000	2.000000	1.000000	2.000000	1.000000	2.000000	0.000000	0.000000	6.000000	1.000000	1.000000	45.000000
1.000000	2.000000	2.000000	2.000000	2.000000	3.000000	0.500000	0.000000	7.000000	3.000000	1.000000	59.000000
2.000000	2.000000	2.000000	2.000000	2.000000	4.000000	2.000000	2.000000	8.000000	3.000000	2.000000	68.000000
2.000000	2.000000	2.000000	2.000000	2.000000	5.000000	5.000000	12.000000	14.500000	3.000000	2.000000	80.000000

RIDRETH3	DMDEDUC2
1404.000000	1404.000000
3.253561	3.405271
1.407579	1.188685
1.000000	1.000000
3.000000	3.000000
3.000000	4.000000
4.000000	4.000000
7.000000	5.000000

Si se analiza la agrupación con 3 clústeres en las mismas condiciones, con igual método ward y medida de distancia euclídea, el comportamiento es muy similar a la segmentación con dos clústeres.

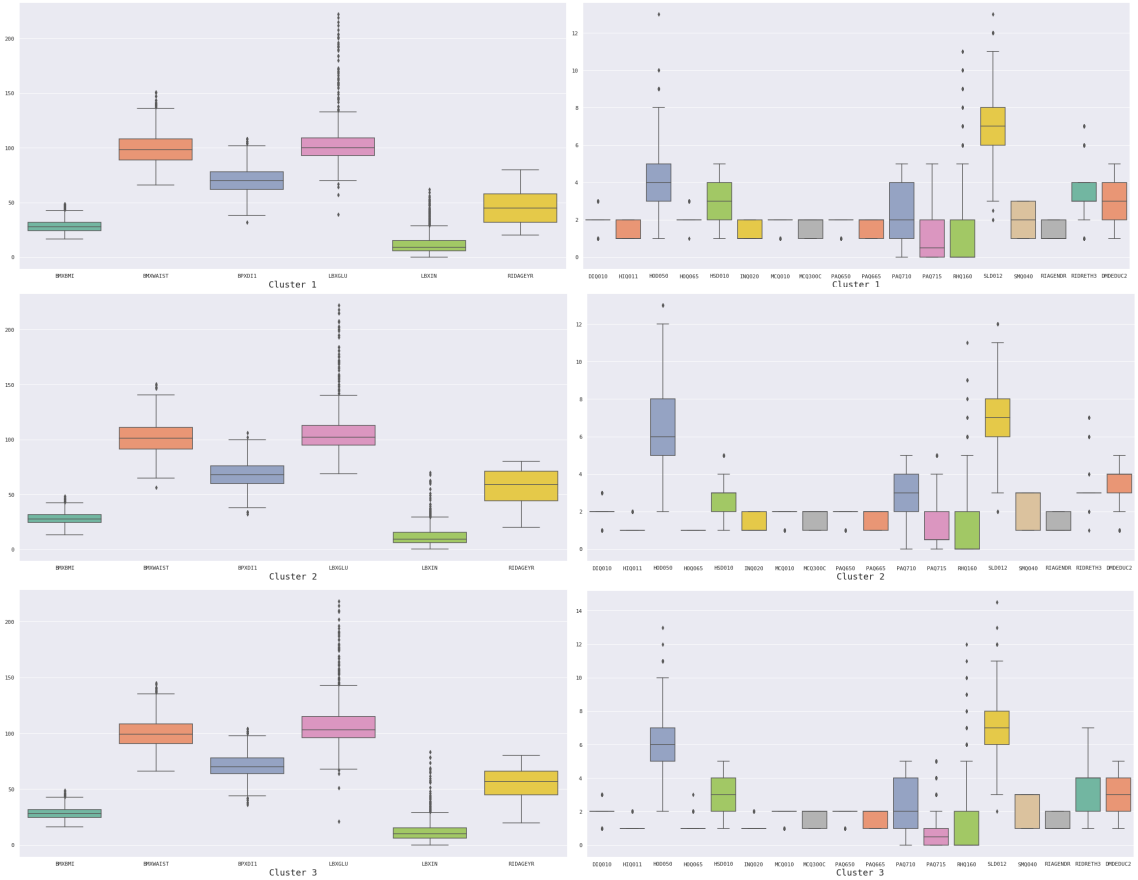


Fig 36. Comparación de dispersión de variables entre clústeres. (3 clústeres)

Con el objetivo de entender mejor las variables y la agrupación realizada se proyecta el conjunto sobre un espacio de dos dimensiones mediante la técnica PCA(Análisis de componentes principales) para representar de forma gráfica el clúster.

Fig 37. Clustering jerárquico aglomerativo con proyección PCA

Teniendo en cuenta que los dos componentes consiguen únicamente capturar el 25% de la variabilidad total de los datos (fig. 38), se observa en la figura 37 dos clústeres bien definidos y cercanos con la posible presencia de outliers del segundo clúster que se sitúan sobre el primero.

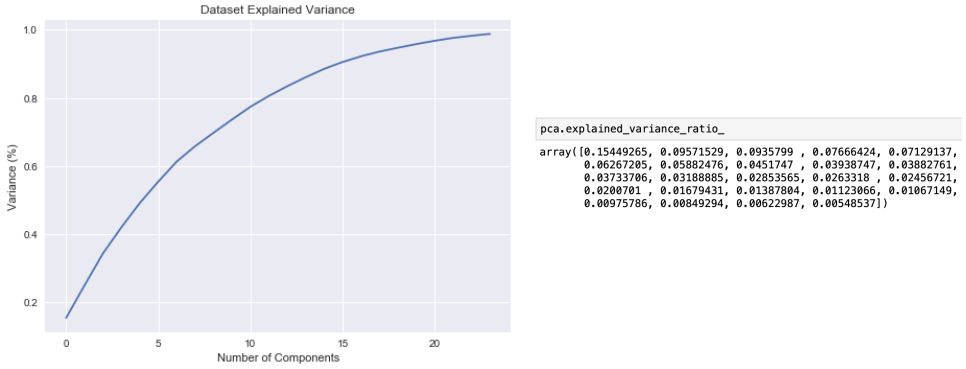


Fig 38. Ratio varianza explicada de componentes PCA

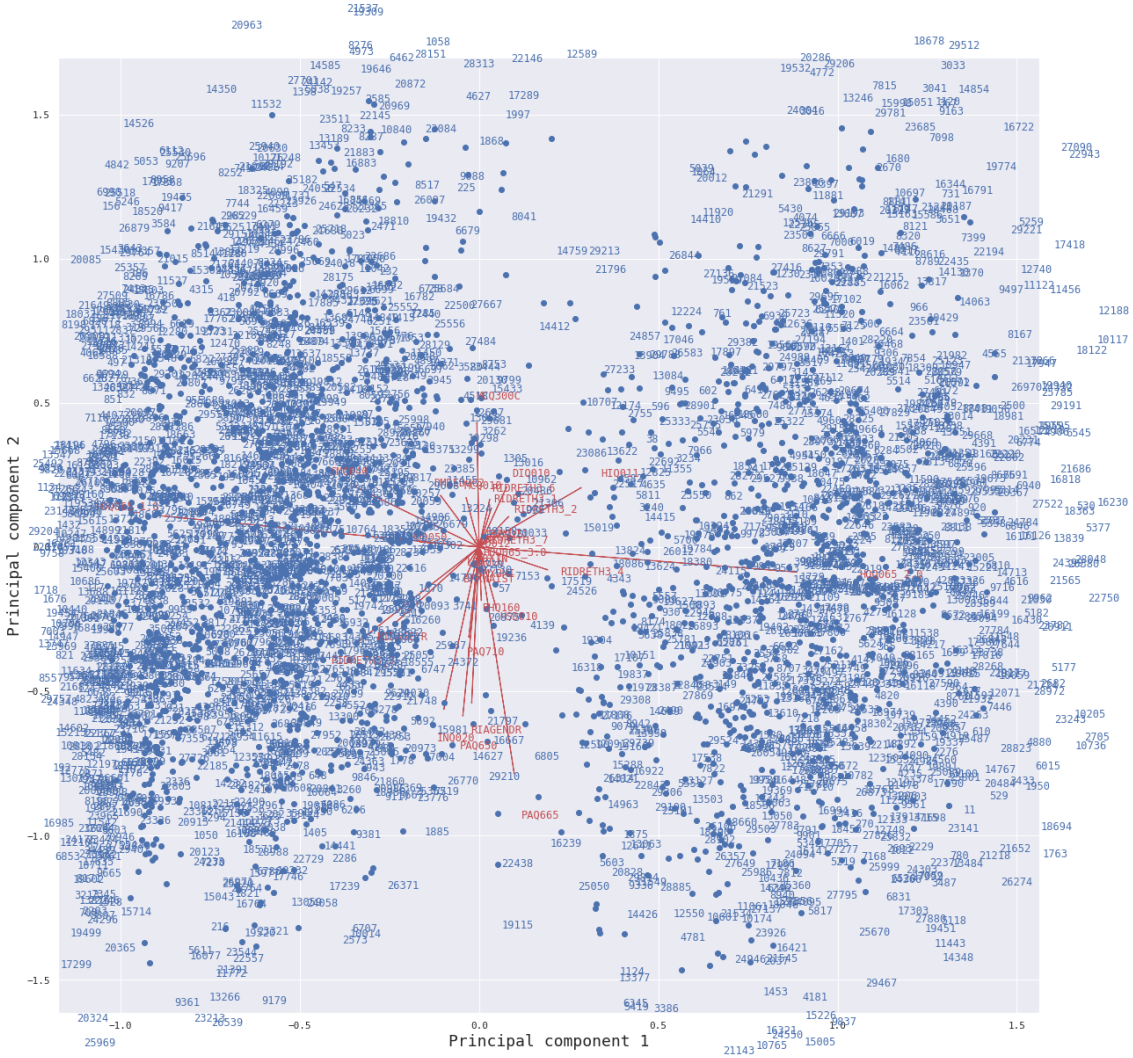


Fig 39. PCA biplot

En la figura 39 se puede observar la influencia de las diferentes variables sobre los componentes principales 1 y 2, sobre el componente 1 cabe destacar por encima del resto la influencia importante de la variable HOQ065 (tipo de propiedad en vivienda), con fuerte correlación negativa entre el alquiler y la propiedad de esta. Este aspecto ya se detectó en el análisis de las variables en el clúster jerárquico.

La baja varianza capturada por el PCA puede ser debido a que las observaciones disponibles no son capaces de explicar el modelo y posiblemente sea necesario aumentar la muestra para aumentar la varianza.

2.3.2 Modelo DBSCAN

Se realiza a continuación un agrupamiento basado en densidad mediante el algoritmo DBSCAN [9], se aplica sobre los mismos datos preparados anteriormente sobre los que se aplicó la agrupación jerárquica aglomerativa.

El primer paso para la aplicación del algoritmo DBSCAN consiste en fijar los parámetros de radio épsilon (ε) y minPts, mínimo número de vecinos, para ello se han seguido las recomendaciones de Erich Schubert et al. [36]. Se ha fijado ε en el valor menor en el que se ha podido obtener una agrupación con la menor estimación posible de outliers, en cuanto a minPts, siguiendo la recomendación que indica que para grupos multidimensionales se puede fijar un valor de dos veces el número de dimensiones presentes, este valor sería de 48.

El proceso de selección de los parámetros ha consistido en la prueba de los parámetros, estableciendo como punto de partida las recomendaciones anteriores, mediante la generación del modelo y evaluación los resultados de estimación de número de clústeres, número de outliers y el coeficiente Silhouette. La mejor combinación en cuanto a resultados obtenidos consiste en un radio de 7 y un minPts de 30. El resultado ha sido de dos clústeres, una estimación de 3 outliers y un coeficiente silhoutte de 0.254.

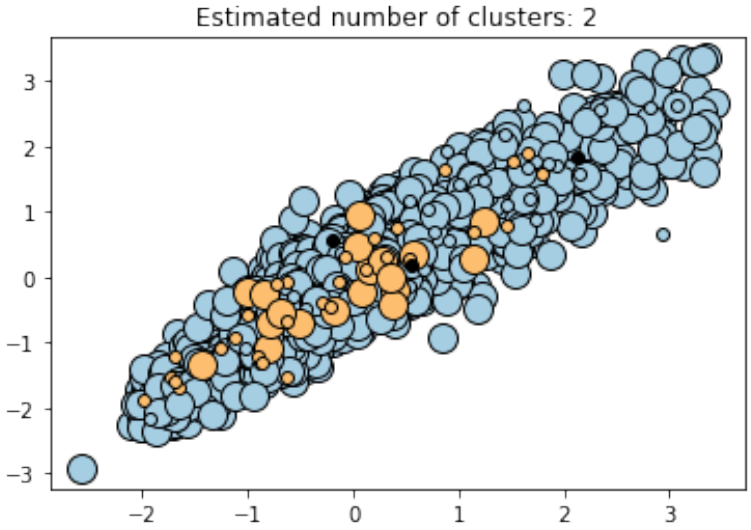


Fig 40. DBSCAN clustering

La partición resultante ha generado dos clústers, el primero cuenta con 2299 observaciones, el segundo con 48 observaciones y 3 outliers. Aspecto que se puede

observar también en la figura 40 donde no se identifican patrones y estructuras que permita la agrupación de los elementos. El valor del coeficiente silhoutte, cercano a cero indica también el solapamiento de los clústeres. Cabe destacar que en el clúster segundo con 48 observaciones, el 100% de estas presentan el mismo valor en la variable HOQ065, tipo de propiedad de vivienda, aspecto similar al presentado en el agrupamiento jerárquico y en el análisis PCA sobre la aparente importancia de esta variable, destacando por encima del resto, en los agrupamientos realizados. El resto de variables presentan características muy similares entre clústeres.

A la vista de los resultados obtenidos y tras haber jugado con los parámetros del algoritmo sin encontrar mejora evidente, todo hace pensar que los datos no poseen particiones basadas en densidad que se puedan detectar.

2.4 Evaluación

Los métodos utilizados de clustering, tanto el jerárquico aglomerativo com DBSCAN no han encontrado agrupaciones en la muestra del conjunto de datos, sus resultados han sido similares, cuando se ha evaluado la posibilidad de generar dos o más clústeres durante la parametrización de los dos modelos los clústeres han resultado en todos los casos con una gran similitud entre ellos y con diferencias de tamaños muy importantes.

Al tratarse de un aprendizaje no supervisado no se tenía ninguna idea preconcebida sobre los resultados a obtener ni unos datos etiquetados con los que contrastar los resultados obtenidos, en consecuencia, para evaluar la calidad del agrupamiento se pueden seguir básicamente tres tipos de criterios:

- Externo, que no se puede aplicar en este caso al no disponer de datos etiquetados.
- Criterio relativo, que se basa en la comparación con otros métodos de agrupamiento diferentes niveles de agrupamiento, en este caso sí aplicable ya que se han creado dos modelos, el jerárquico y el basado en densidad. Y según se ha comentado su comportamiento es muy similar una vez aplicados sobre los datos disponibles.
- Criterio interno, medidas de calidad interna de los grupos creados en cuanto compatibilidad y separación de estos, el ya analizado coeficiente silhoutte indicaba en el apartado anterior solapamiento en los casos de más de un grupo con valores cercanos a 0.

Se analiza a continuación el indice Davies-Bouldin [37] (maximum interclass-intraclass distance ratio), otro criterio de calidad interna para corroborar conclusiones sobre los agrupamientos realizados.

Tabla 16: Comparación índice Davies-Bouldin.

Modelo	Nº grupos	Índice Davies-Bouldin
HAC Ward	2	2.45
HAC Ward	3	2.70
DBSCAN	2	1.79

Los índices Davies-Bouldin obtenidos son bastante elevados según se puede observar en tabla 16, teniendo en cuenta que los valores cercanos a 0 son indicativos de buenas particiones, se puede concluir que según también este índice no se está realizado unas buenas particiones.

No parece por tanto haber agrupaciones con significado, tanto en términos de distancia como de densidad, para crear más de un clúster a partir de los datos disponibles, parece que se está ante unos datos muy homogéneos y de baja varianza. Este aspecto se detectó ya en la fase de comprensión y entendimiento de los datos durante la eliminación de casos con datos ausentes que provocó la creación de grupos predominantes en la muestra, por ejemplo que la muestra tuviera una media de edad elevada generando probablemente una muestra no balanceada con individuos de muy similares características.

Otra causa puede haber sido la selección de variables para el estudio, que estas resulten también similares y no presenten discriminación entre las observaciones, posiblemente presentando colinealidad entre ellas, aspecto que se estudió ligeramente en la fase preparación de los datos debido a la presencia de variables categóricas y continuas en los datos. En cuanto a la cantidad de observaciones utilizadas, este parece suficiente en base a otros estudios similares con parecido número de observaciones.

3. Interfaz web

3.1 Diseño

Se implementa en esta fase una simple aplicación web que permite la clasificación de individuos mediante la introducción de las características de este y perfilarlo en uno de los clústeres del modelo jerárquico aglomerativo de dos clústeres implementado en la fase anterior.

Para ello es necesario empotrar el modelo de clasificación, ya entrenado y validado, en la aplicación, presentar un formulario que facilite la entrada de los datos por el usuario, a continuación pasar los datos introducidos al modelo para que realice la predicción oportuna y finalmente presentar el resultado al usuario.

Para crear la aplicación se usa el framework Flask [38], que permite un desarrollo web en Python, lenguaje usado durante todo el trabajo, y ofrece sencillez en la implementación. Una vez creada la aplicación se hace pública a través de la Heroku, un servicio de plataforma en la nube (PaaS) [39].

3.2 Modelo de clasificación

El modelo de clasificación usado es una regresión logística donde la variable dependiente es el número de clúster a predecir, esta es binaria, y las variables independientes son los atributos analizados en fases anteriores. El conjunto de datos de entrenamiento para el modelo de clasificación es el resultado del etiquetado, se ha añadido la variable 'cluster' con el número al que pertenece, 1 ó 2, de cada una de las observaciones en la agrupación jerárquica, está formado por 2350 observaciones y 25 variables.

Tabla 17: Muestra conjunto de datos.

	SEQN	BMXBMI	BMXWAIST	BPXDI1	DIQ010	LBXGLU	LBXIN	HIQ011	HOD050	HOQ065	...	PAQ710	PAQ715	RHQ160	SLD012	SMQ040	RIAGENDR	RIDAGEYR	RIDRETH3	DMDEDUC2	CLUSTER
11	62172	33.3	120.4	70.0	2.0	104.0	18.62	1	4.0	2.0	...	5.0	5.0	3.0	8.0	1.0	2	43	4	3.0	1
38	62199	28.0	107.8	70.0	2.0	100.0	10.02	1	4.0	2.0	...	1.0	1.0	0.0	8.0	3.0	1	57	3	5.0	1
241	62402	23.3	82.0	70.0	2.0	107.0	7.81	2	6.0	3.0	...	4.0	0.5	0.0	9.0	1.0	1	48	6	1.0	1
250	62411	26.3	100.1	62.0	2.0	96.0	16.07	1	3.0	2.0	...	4.0	0.0	4.0	7.0	1.0	2	51	4	4.0	1
269	62430	37.9	114.2	74.0	2.0	97.0	15.90	1	5.0	2.0	...	3.0	3.0	0.0	6.0	3.0	1	27	2	5.0	1

Durante la validación del modelo de regresión logística se ha realizado una validación cruzada de 5 iteraciones en la que se ha obtenido una precisión media del 99%

```
print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

```
Accuracy: 0.99 (+/- 0.01)
```

Fig 41. Precisión 5-Fold Cross validation

El coeficiente de determinación, como métrica que evalúa lo bien que el modelo explica la variabilidad y su capacidad de predicción, ha arrojado un resultado de 0.947, lo cual indica que el modelo parece explicar en un alto grado la variabilidad de la variable cluster por los datos.

Finalmente de la matriz de confusión sobre las predicciones en el conjunto de test los valores de precisión, 0.975, accuracy, 0.987 y recall, 0.99 muestran la bondad del modelo en sus predicciones.

```
y_pred=lr.predict(X_test)
c_matrix = metrics.confusion_matrix(y_test, y_pred)
c_matrix
```

```
array([[195,   1],
       [  5, 269]])
```

Fig 42. Matriz de confusión de predicciones sobre conjunto de test.

3.3 Uso y funcionamiento

La aplicación creada se puede visitar en la plataforma Heroku donde está alojada a través del siguiente enlace: https://flask-tfm-app.herokuapp.com

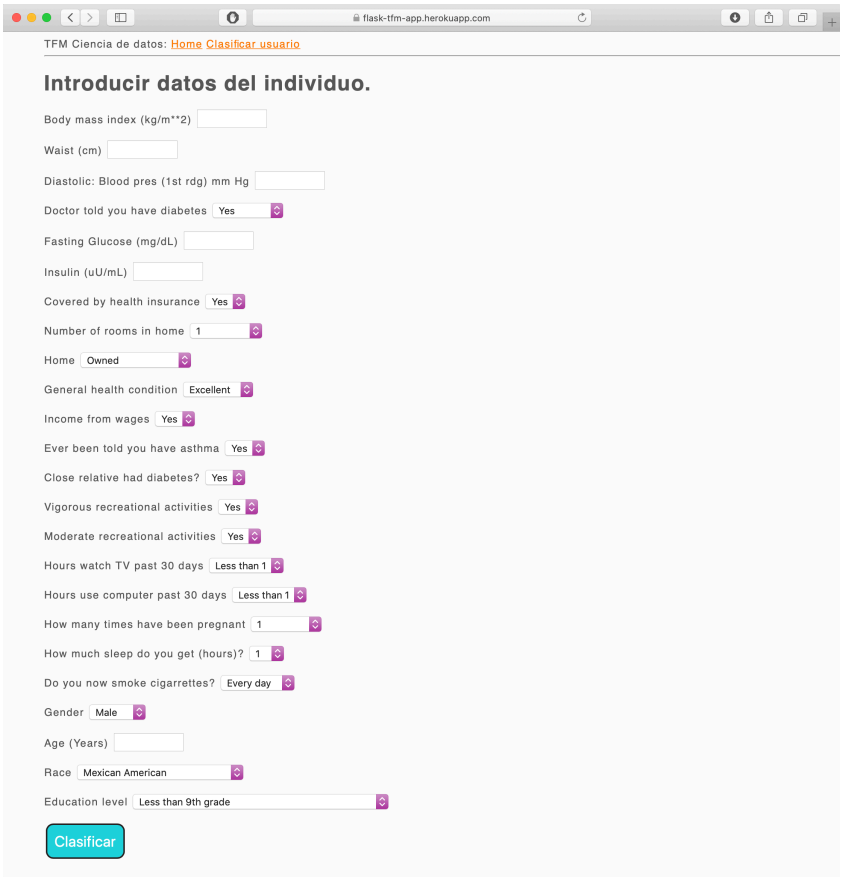


Fig 42. Formulario de entrada de datos aplicación.

En la página de entrada se encuentra una breve descripción de los clústeres a los cuales se pueden asignar los individuos en el momento de la clasificación.

A través del enlace “Clasificar usuario” en el menú superior se pasa a la página donde se deben introducir los datos del individuo a clasificar. Los campos poseen una serie de validadores que aseguran que los datos introducidos sean correctos. Una vez los datos han sido introducidos mediante el botón “Clasificar” situado en la parte inferior de la página se mostrará la predicción del clúster más afín con los datos del individuo tal y como se muestra en la siguiente figura.

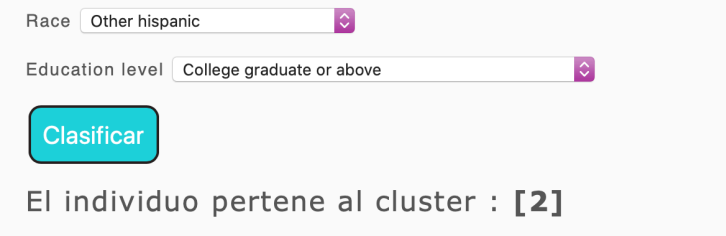


Fig 43. Resultado de clasificación.

4. Conclusiones

Una de las principales lecciones aprendidas durante la ejecución del trabajo ha sido, durante la recolección de los datos, el primer objetivo marcado, la dificultad en la creación de un conjunto de datos que sea idóneo para los objetivos que se marcan en una tarea de aprendizaje automático, más si cabe en uno de aprendizaje no supervisado como éste.

Ante esta dificultad se optó por una selección de variables basada en la combinación de atributos presentes en literatura consultada sobre trabajos similares, los resultados no han sido los esperados y no se ha obtenido una segmentación clara de pacientes de diabetes ya que estos se distribuían uniformemente en las agrupaciones obtenidas.

La causa ha sido principalmente la selección de las variables cuya combinación no parece ofrecer una visión relevante de la salud del individuo en el estudio de esta enfermedad a la vez que se ha producido una uniformidad elevada de las características de los individuos seleccionados tras la limpieza y tratamiento de los datos.

A raíz de lo comentado se desprende la importancia que tiene el conocimiento del dominio a la hora de construir un conjunto de datos y en el establecimiento de los objetivos de la tarea de aprendizaje que se va a llevar a cabo, tanto en el momento de la ejecución del estudio como después en la aplicación de los resultados obtenidos.

Por otro lado esta fase de recolección de datos ha resultado ser un sencillo proceso gracias al uso del paquete estadístico SAS que hace fácil los procesos de extracción, combinación y exportación de los datos recolectados.

Respecto a la creación de los modelos de agrupamiento, tanto el jerárquico aglomerativo como el DBSCAN, en ambos ha sido complicado encontrar agrupaciones que permitieran perfilar a los individuos con características diferenciadas entre grupos generados. En ambos casos con índices de calidad de clústeres ofreciendo unos valores pobres indicando baja calidad de los grupos y alto solapamiento entre ellos, siendo imposible en el caso del modelo basado en densidad extraer conocimiento alguno.

En cuanto al modelo jerárquico, con una agrupación de dos clústeres, permite agrupar las observaciones de forma que se pueden extraer las siguientes conclusiones como principales características definitorias de cada clúster:

Los clústeres formados han sido dos, el primero con 946 observaciones y el segundo con 1404.

El cluster 1 está formado en su mayoría (94.5%) por personas con vivienda habitual en alquiler con una media de 4,5 habitaciones, por contra el clúster 2 está formado personas que en su mayoría (98%) cuentan con vivienda en propiedad y número medio de habitaciones que se sitúa en 6.5. El clúster 1 corresponde a personas más jóvenes que en el clúster 2, siendo la media de edad de 46 años en el primero y 56 años en el segundo clúster.

Entre los hábitos de los individuos de los dos grupos, el uso de ordenador y televisión es ligeramente diferente entre ambos, el grupo 1 formado por personas más jóvenes el uso de ordenador es mayor frente al grupo 2, al contrario sucede con el consumo de televisión que es mayor en el grupo 2 frente al grupo 1. Respecto a los hábitos de actividad física que se realiza, esta es entre el grupo 1 ligeramente mayor la actividad intensa mientras en el grupo 2 es mayor la actividad física moderada.

El resto de características como son la educación, el diagnóstico de la diabetes, edad o el asma presentan ratios de presencia muy similares entre ambos grupos.

En cuanto al análisis de los logros sobre los objetivos que se marcaron inicialmente, comentar que se logró la creación del conjunto de datos objeto del estudio a partir de NHANES, la creación de los modelos de agrupamiento así como su validación y la implementación y despliegue de una aplicación web con el modelo de clasificación, según etiquetado de la tarea de agrupamiento previa, empotrado que permite la clasificación de individuos en clústeres según sus características. Por el contrario no se ha conseguido un perfilado útil de los individuos respecto a la patología tal como era la intención inicial del objetivo marcado al no conseguir descubrir patrones relevantes en los datos.

Respecto a la planificación de los trabajos inicial, esta ha sido correcta, todas las entregas se han realizado a tiempo, las fases intermedias se han cumplido todas prácticamente al día a excepción de la fase de selección de las variables para el conjunto de datos que se demoró en cuatro días aproximadamente respecto a la planificación inicial. La metodología empleada, CRISP-DM ha resultado también ser válida para completar los objetivos, la definición de las fases y tareas y su carácter cíclico ha permitido mejorar los modelos cuando no se obtuvieron los resultados esperados.

Como tareas de trabajo futuro que no se han podido llevar a cabo por falta de tiempo, incluiría en primer lugar la mejora del conjunto de datos, mejorar la selección de variables, más apropiadas a los objetivos, para los modelos de agrupamiento e intentar conseguir unos clústeres de mejor calidad, así como también intentar balancear más las observaciones presentes y evitar que el conjunto de datos contenga individuos muy similares entre ellos. Si se consigue obtener un conjunto de características que mejore las agrupaciones también la aplicación web construida podría ser más útil en la clasificación de pacientes según características de la patología como objetivo.

Respecto a la aplicación web, una posible mejora podría ser el almacenamiento de los datos de los individuos que se van clasificando y usarlos tanto para los modelos de agrupamiento como el entrenamiento del modelo de clasificación e ir actualizando estos modelos con los nuevos datos disponibles.

5. Glosario

- **Aprendizaje no supervisado**: Método de aprendizaje automático que trata de encontrar patrones antes desconocidos en un conjunto de datos sin información previa conocida.

- **CDC**: Acrónimo de Centro de Control de enfermedades y Prevención.

- **Clúster**: Partición o grupo de objetos de similares características.

- **Clustering**: Aplicación más común del aprendizaje no supervisado, tarea que busca crear grupos entre objetos de forma que los objetos en un mismo grupo son similares entre ellos y difieren con los objetos de otros grupos.

- **CRISP-DM**: Acrónimo de Cross Industry Standard Process for Data Mining.

- **CVI**: Acrónimo de Cluster Validity Indices, relación de estimaciones capaces de cuantificar la calidad de las particiones (clústeres) creadas por los algoritmos de de clustering.

- **DBSCAN**: Acrónimo de Density-based spatial clustering of applications, algoritmo de agrupamiento de datos.

- **Dendrograma**: Diagrama que muestra la relación jerárquica entre objetos.

- **Estandarización**: El proceso de cambiar las escalas de uno o más atributos para que tengan media igual a 0 y desviación estándar igual 1.

- **FLASK**: Framework escrito en python destinado a la creación de aplicaciones web.

- **Granularidad**: Define la escala o el nivel de detalle de una partición.

- **HEROKU**: Heroku es una plataforma como servicio de computación en la Nube que soporta distintos lenguajes de programación.

- **Imputación**: El proceso de reemplazo de datos ausentes en un conjunto de datos por valores atribuidos.

- **Interfaz web**: Conjunto de elementos que permiten la interacción con el usuario implementada en formato web que se puede navegar mediante un navegador web.

- **Matriz de confusión**: Herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.

- **Matriz de distancia**: Es una matriz cuyos elementos representan las distancias entre los puntos, tomados por pares, de un conjunto.

- **Modelo**: En aprendizaje automático puede ser una representación matemática de un proceso del mundo real, para generarlo es necesario un conjunto de datos con el que entrenar un algoritmo.

- **NHANES**: Acrónimo de National Health and Nutrition Examination Survey.

- **NCHS**: Acrónimo de National Center for Health Statistics.

- **Notebook jupyter**: Jupyter Notebook es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo.

- **One-hot-encoding**: Es un proceso mediante el cual las variables categóricas se convierten para poder tratarse por algoritmos de aprendizaje automático. El valor categórico representa el valor numérico de la entrada en el conjunto de datos.

- **Outlier**: Es una observación que es numéricamente distante del resto de los datos.

- **PaaS**: Acrónimo de Platform as a service, Servicios de infraestructura en la nube ofrecidos por un proveedor mediante la contratación del cual un usuario puede desplegar una aplicación.

- **Paquete estadístico**: Aplicación informática destinada a solucionar problemas relacionados con la estadística.

- **Parametrización**: Asignar valores a parámetros declarados para modificar o influir en su comportamiento.

- **PCA**:Acrónimo de Principal Component Analysis, técnica utilizada para describir un conjunto de datos en términos de nuevas variables ("componentes") no correlacionadas.

- **Scikit-learn**: Es una biblioteca para aprendizaje automático de software libre para el lenguaje de programación Python.

- **SAS®**: Acrónimo de Statistical Analysis Software. Paquete estadístico para el análisis de datos.

- **Z-score**: Es el número de desviaciones estándar que una observación se encuentra alejada de la media del conjunto.

6. Bibliografía

[1] About the National Health and Nutrition Examination Survey, https://www.cdc.gov/nchs/nhanes/about_nhanes.htm Accedido: 24/09/2019

[2] Gironés, J., Casas, J., Minguillón, J., Caihuelas, R. (2017). Minería de datos. Modelos y algoritmos, Capítulo 1.

[3] Jupyter dashboards, Software Carpentry Foundation, https://annefou.github.io/jupyter_dashboards/02-dashboards/index.html Accedido: 29/09/2019

[4] Moore, JX., Chaudhary, N., Akinyemiju, T. Metabolic Syndrome Prevalence by Race/Ethnicity and Sex in the United States, National Health and Nutrition Examination Survey, 1988–2012. Prev Chronic Dis 2017; 14:160287. DOI: https://doi.org/10.5888/pcd14.160287.

[5] Zhang, F., Tapera, TM. and Gou, J. "Application of a New Dietary Pattern Analysis Method in Nutritional Epidemiology." *BMC Medical Research Methodology* 18, no. 1 (October 29, 2018): 119. https://doi.org/10.1186/s12874-018-0585-8.

[6] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58, no. 1 (1996): 267-88. http://www.jstor.org/stable/2346178.

[7] Jun won, L. and Giraud-Carrier, C. "Results on Mining NHANES Data: A Case Study in Evidence-Based Medicine." *Computers in Biology and Medicine* 43, no. 5 (June 2013): 493–503. https://doi.org/10.1016/j.compbiomed.2013.02.018.

[8] Ferenci, T., and Kovács, L. "Using Total Correlation to Discover Related Clusters of Clinical Chemistry Parameters." In *2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY)*, 49–54, 2014. https://doi.org/10.1109/SISY.2014.6923614.

[9] Ester, M., Kriegel, HP., Sander, J. and Xu, X. 1996. "A Density-Based Algorithm for Discovering Clusters a Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231. KDD'96. AAAI Press. http://dl.acm.org/citation.cfm?id=3001460.3001507.

[10] Ankerst, M., Breunig, MM., Kriegel, HP. and Sander, J. 1999. "OPTICS: Ordering Points to Identify the Clustering Structure." In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 49–60. SIGMOD '99. New York, NY, USA: ACM. https://doi.org/10.1145/304182.304187.

[11] Hinneburg, A., and Keim, DA. 1998. “An Efficient Approach to Clustering in Large Multimedia Databases with Noise.” In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 58–65. KDD’98. AAAI Press. http://dl.acm.org/citation.cfm?id=3000292.3000302.

[12] Ram, A., Jalal S., and Kumar M. 2010. “A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases.” *International Journal of Computer Applications* 3 (June). https://doi.org/10.5120/739-1038.

[13] Müllner, D. 2011. “Modern Hierarchical, Agglomerative Clustering Algorithms.” *ArXiv:1109.2378 [Cs, Stat]*, September. http://arxiv.org/abs/1109.2378.

[14] Müllner, D. 2013. “Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python.” Journal of Statistical Software 53 (1): 1–18. https://doi.org/10.18637/jss.v053.i09.

[15] Konopka, BM, Lwow, F., Owczarz M., Łaczmański Ł. (2018) Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PLoS ONE 13 (8): e0201950. https://doi.org/10.1371/journal. pone.0201950

[16] Arbelaitz, O., Gurrutxaga, I., Muguerza J., Pérez, J., and Perona I. 2013. “An Extensive Comparative Study of Cluster Validity Indices.” *Pattern Recognition* 46 (1): 243–56. https://doi.org/10.1016/j.patcog.2012.07.021.

[17] Subirats, L., Gil, R., and García, R. 2019. “Personalization of Ontologies Visualization: Use Case of Diabetes.” In *Current Trends in Semantic Web Technologies: Theory and Practice*, edited by Giner Alor-Hernández, José Luis Sánchez-Cervantes, Alejandro Rodríguez-González, and Rafael Valencia-García, 3–24. Studies in Computational Intelligence. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-06149-4_1.

[18] Cook, R., and Hu, G. 2010. “Hidden Patterns: Clustering Diabetes Data.” In *CAINE*.

[19] Farran, B., AlWotayan R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A. and Thanaraj, TA. 2019. “Use of Non-Invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait.” *Frontiers in Endocrinology* 10. https://doi.org/10.3389/fendo.2019.00624.

[20] Xie, Z., Nikolayeva, O., Luo, J., LI, D. 2019. “Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques.” *Preventing Chronic Disease* 16. https://doi.org/10.5888/pcd16.190109.

[21] Detection of Outliers, https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm [Accedido: 16/11/2019]

[22] InterQuartile Range (IQR) http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html [Accedido: 16/11/2019]

[23] 2000 Census of Population and Housing. Profiles of General Demographic Characteristics. https://www2.census.gov/census_2000/datasets/demographic_profile/0_United_States/2kh00.pdf [Accedido: 17/11/2019]

[24] El índice de masa corporal para adultos, CDC Centro de estudios para el control y la prevención de enfermedades https://www.cdc.gov/healthyweight/spanish/assessing/bmi/adult_bmi/index.html [Accedido: 17/11/2019]

[25] Insulin Resistance/T2 Diabetes: Map Your Test Results https://www.thebloodcode.com/insulin-resistancet2-diabetes-map-test-results/ [Accedido: 20/11/2019]

[26] Understanding Blood Pressure Readings, https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings [Accedido: 20/11/2019]

[27] Encoding categorical features https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-categorical-features [Accedido: 01/01/2019]

[28] Sklearn preprocessing module https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing [Accedido: 01/12/2019]

[29] SciPy.org scipy.cluster.hierarchy.linkage https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage [Accedido: 04/12/2019]

[30] Euclidean Space and Metric Spaces https://www.math.uci.edu/~gpatrick/source/205b06/chapviii.pdf [Accedido: 04/12/2019]

[31] Gironés J., Casas J., Minguillón J., Caihuelas R. (2017). Minería de datos. Modelos y algoritmos, Capítulo 7, punto 7.2.

[32] Selecting the number of clusters with silhouette analysis on KMeans clustering https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html [Accedido: 05/12/2019]

[33] Cophenet https://es.mathworks.com/help/stats/cophenet.html?lang=en [Accedido: 05/12/2019]

[34] Saraçli, S., Doğan, N. and Doğan, I. 2013. "Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation." *Journal of Inequalities and Applications* 2013 (1): 203. https://doi.org/10.1186/1029-242X-2013-203.

[35] Coeficiente de variación. https://es.wikipedia.org/wiki/Coeficiente_de_variación [Accedido: 09/12/2019]

[36] Schubert, E, Sander, J., Ester, M., Peter Kriegel, H. and Xu, X. 2017. "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN." *ACM Trans. Database Syst.* 42 (3): 19:1–19:21. https://doi.org/10.1145/3068335.

[37] Davies, David, and Don Bouldin. 1979. "A Cluster Separation Measure." *Pattern Analysis and Machine Intelligence, IEEE Transactions On* PAMI-1 (May): 224–27. https://doi.org/10.1109/TPAMI.1979.4766909.

[38] Flask https://www.palletsprojects.com/p/flask/ [Accedido: 21/12/2019]

[39] What is Heroku? https://www.heroku.com/about [Accedido: 21/12/2019]

7. Anexos

Anexo 1: Consulta del código generado durante la ejecución del trabajo.

Alojado en GitHub se puede acceder al código generado a través del siguiente enlace: https://github.com/rasantem/TFM