ESCUELA DE INGENIERÍA DE FUENLABRADA

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

Análisis factorial exploratorio y métodos de clustering para la identificación de factores de riesgo en pacientes con hipoglucemia severa

Autora: Mounat El Jarmouni

Tutora: Cristina Soguero Ruiz

Co-tutor: Cristian David Chushig Muzo

Curso académico 2022/2023

El éxito es la consecuencia directa de la perseverancia.

– Anónimo

Agradecimientos

En primer lugar, agradecer al *Jaeb Center For Health Research* por haber permitido hacer uso de las bases de datos, sin las cuales este estudio no hubiese sido posible. También quiero agradecer a mis tutores Cristina Soguero Ruiz y Cristian David Chushig Muzo por su dedicación, disponibilidad y orientación a lo largo de estos meses. Su ayuda ha sido fundamental para la realización de este proyecto.

Además, quiero dar las gracias a mis amigas, de manera especial a las me ha dado esta etapa universitaria. Gracias por compartir lo mejor y lo peor de estos años, pero sobre todo por aprender tantas cosas juntas. Por último, y lo más importante, agradecer a mi familia por ser un apoyo constante en esta y en todas las etapas de mi vida.

Resumen

Actualmente, las enfermedades crónicas (ECs) representan el 80 % de la carga mundial de enfermedades, siendo una de las principales causas de muerte globalmente. La diabetes mellitus tipo 1 (DMT1) es una EC en la que el sistema inmunitario destruye las células beta pancreáticas, provocando una caída en la producción de insulina y requiriendo insulina exógena. Sin un correcto control, los individuos con DMT1 pueden presentar episodios de hipoglucemia severa (HS). Esta ocurre cuando el nivel de glucosa cae a un nivel en el que el individuo entra en un estado de confusión, pérdida de conocimiento y necesita ayuda de otra persona. Se estima que el 35 % de los pacientes con DMT1 presentan episodios de HS.

En los últimos años, las tecnologías de la información han permitido producir y analizar datos que son claves en la investigación clínica. En ese contexto, los modelos de Aprendizaje Automático (ML, del inglés *Machine Learning*) han sido utilizados para identificar factores de riesgo implicados en la aparición de eventos clínicos graves. En este sentido, modelos ML son prometedores para abordar la problemática de la HS y mejorar la calidad de vida de los pacientes. El Análisis Factorial Exploratorio (EFA, del inglés *Exploratory Factorial Analysis*) es un modelo ML usado para explorar relaciones entre variables de entrada y asociar estas relaciones en grupos de variables (factores), permitiendo la identificación de patrones.

El objetivo del presente Trabajo Fin de Grado (TFG) es doble. En primer lugar, se aplica EFA para identificar factores (grupos de variables) que permitan comprender las causas subyacentes de la HS. En segundo lugar, se aplican métodos de *clustering* para agrupar pacientes con DMT1 según características clínicas similares. Para la realización de este TFG, se utilizan datos de bases de datos públicas del *Jaeb Center for Health Research* formados por cuestionarios realizados a individuos con DMT1. Específicamente, se consideran cuestionarios sobre depresión, miedo y desconocimiento de síntomas de hipoglucemia, evaluación cognitiva y aptitudes de los pacientes frente a niveles de glucosa.

Los resultados experimentales usando EFA han permitido identificar patrones de comportamiento de individuos con DMT1 frente a escenarios con: bajos/altos niveles de glucosa, síntomas de hipoglucemia, emociones positivas, preocupaciones, entre otras. Por otro lado, los métodos de *clustering* han permitido identificar grupos de pacientes con patrones de respuesta similares, distinguiendo aquellos con una tendencia a sufrir HS. La combinación de ambas técnicas ha permitido revelar relaciones entre respuestas a cuestionarios, e identificación de patrones en el desarrollo de HS. Los modelos usados en este TFG han mostrado ser eficientes y válidos para la identificación de factores de riesgo de HS, siendo prometedores para apoyar la toma de decisiones y mejoramiento de calidad de vida de pacientes.

Índice general

Agradecimientos

Resumen

Índice de figuras III

Índice de tablas V

1. Introducción y objetivos **1**

1.1. Contexto y motivación 1

1.2. Objetivos 3

1.3. Metodología 5

1.4. Estructura de la memoria 7

2. Conceptos previos **9**

2.1. Diabetes mellitus e hipoglucemia severa 9

2.2. Conceptos de aprendizaje automático 14

2.3. Análisis factorial exploratorio 16

2.4. Métodos de *clustering* 23

3. Base de datos y análisis descriptivo **27**

3.1. Descripción de las bases de datos 27

3.2. Análisis descriptivo y preprocesamiento 29
3.2.1. Base de datos *BBGAttitudeScale* 30
3.2.2. Base de datos *BGeriDepressScale* 33
3.2.3. Base de datos *BHypoFearSurvey* 38
3.2.4. Base de datos *BHypoUnawareSurvey* 42
3.2.5. Base de datos *BMoCA* 46

4. Experimentos y resultados **51**
4.1. Configuración experimental 51
4.2. Análisis factor exploratorio y *clustering* usando *BBGAttitudeScale* 54
4.3. Análisis factor exploratorio y *clustering* usando *BGeriDepressScale* 60
4.4. Análisis factor exploratorio y *clustering* usando *BHypoFearSurvey* 65
4.5. Análisis factor exploratorio y *clustering* usando *BHypoUnawareSurvey* 72
4.6. Análisis factor exploratorio y *clustering* usando *BMoCA* 76
4.7. Análisis factorial exploratorio y *clustering* usando bases concatenadas 81

5. Conclusiones y líneas futuras **89**
5.1. Conclusiones 89
5.2. Líneas futuras 90

Bibliografía **93**

Índice de figuras

1.1. Diagrama de Gantt con la duración orientativa de las actividades realizadas. . . 7

3.1. Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BBGAttitudeScale*. 31

3.2. Histogramas de las variables de la base de datos *BBGAttitudeScale* diferenciando entre casos y controles. 33

3.3. Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BGeriDepressScale*. 34

3.4. Histogramas de las variables de la base de datos *BGeriDepressScale* diferenciando entre casos y controles. 37

3.5. Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BHypoFearSurvey*. 38

3.6. Histogramas de las variables de la base de datos *HypoFearSurvCompDaysFromEnroll* diferenciando entre casos y controles. 41

3.7. Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BHypoUnawareSurvey*. 42

3.8. Histogramas de las variables de la base de datos *BHypoUnawareSurvey* diferenciando entre casos y controles. 46

3.9. Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BMoCA*. 47

3.10. Histogramas de las variables de la base de datos *BHypoUnawareSurvey* diferenciando entre casos y controles. 49

4.1. Esquema de la parte experimental del estudio. 53

4.2. Resultados del EFA usando *BBGAttitudeScale*. 56
4.3. Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BBGAttitudeScale*. 58
4.4. Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BBGAttitudeScale*. 59
4.5. Resultados del EFA usando *BGeriDepressScale*. 62
4.6. Mátricas para la determinación del número de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BGeriDepressScale*. 64
4.7. Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BGeriDepressScale*. 65
4.8. Resultados del EFA usando *BHypoFearSurvey*. 68
4.9. Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BHypoFearSurvey*. 70
4.10. Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BHypoFearSurvey*. 71
4.11. Resultados del EFA usando *BHypoUnawareSurvey*. 73
4.12. Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BHypoUnawareSurvey* 75
4.13. Perfiles de los *clusters* de la base de datos *BHypoUnawareSurvey*. 76
4.14. Resultados del EFA usando *BMoCA*. 78
4.15. Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BMoCA*. 80
4.16. Perfiles de los *clusters* de la base de datos *BMoCA* 81
4.17. Resultados del EFA usando las bases de datos concatenadas. 84
4.18. Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para bases concatenadas. 86
4.19. Perfiles de los *clusters* de las bases de datos concatenadas. 87

Índice de tablas

3.1. Codificación correspondiente a la escala Likert. 30

3.2. Descripción y opciones de respuesta para las variables de la base de datos *BBG-AttitudeScale.* 32

3.3. Descripción y opciones de respuesta para las variables de la base de datos *BGeriDepressScale.* 35

3.4. Descripción y opciones de respuesta para las variables de la base de datos *BHypoFearSurvey.* 39

3.5. Descripción y opciones de respuesta para cada una de las variables de la base de datos *BHypoUnawareSurvey.* 44

3.6. Descripción e intervalos de respuesta para cada una de las variables de la base de datos *BMoCA.* 48

Capítulo 1

Introducción y objetivos

Este capítulo constituye un preámbulo al problema de la diabetes, incluyendo su relación con la aparición de episodios de hipoglucemia severa que sufren algunos pacientes que presentan diabetes tipo I (DMT1). A continuación, se definen los objetivos del presente Trabajo de Fin de Grado (TFG), la metodología seguida y una breve estructura de la memoria.

1.1. Contexto y motivación

La Organización Mundial de la salud (OMS) define las enfermedades crónicas como «aquellas patologías que presentan una larga duración (más de 6 meses) y una progresión lenta, no se transmiten de una persona a otra y son consideradas, por lo tanto, como no transmisibles» [1]. Estas enfermedades están entre las principales causas de muerte e incapacidad en el mundo y se han convertido un problema de salud pública [2]. Afectan a la mayoría de los países debido a que su desarrollo no es precedible y se manifiestan por varios factores de riesgo generalmente modificables como: consumo de tabaco y alcohol, una dieta inadecuada y una actividad física reducida. De estas enfermedades, las que mayor prevalencia presentan son las enfermedades cardiovasculares, el cáncer, las enfermedades respiratorias y la diabetes [3].

Este proyecto se centra en la diabetes mellitus (DM) que es la alteración endocrina más frecuente caracterizada por elevados niveles de glucosa en sangre debidos a una deficiencia en la secreción o acción de la insulina [4]. Además, se trata de una enfermedad crónica con un gran número de complicaciones y comorbilidades asociadas que tienen un alto impacto en el gasto sanitario [5]. La diabetes se clasifica en varios tipos, siendo las más comunes la diabetes tipo I (DMT1) que se debe a una producción insuficiente de insulina y generalmente se desarrolla en

la infancia o adolescencia [4]. La diabetes tipo II (DMT2) que es la más frecuente, se caracteriza por un uso metabólico ineficiente de la insulina producida por el páncreas y está estrechamente relacionada con el sobrepeso y la obesidad y por último la diabetes gestacional que afecta a algunas mujeres durante el embarazo [4]. La prevención y el control de esta patología son imprescindibles para evitar complicaciones a largo plazo y reducir la carga social y económica que supone para los sistemas sanitarios [6].

La prevalencia de la diabetes está en constante aumento en todos los países del mundo, lo cual es motivo de preocupación para los sistemas de salud. Esta situación es especialmente preocupante en los países en vías de desarrollo, donde el aumento de la esperanza de vida y los estilos de vida poco saludables son factores determinantes en este incremento [7]. Por consiguiente, la DM se posiciona como una enfermedad de primera importancia en términos de salud pública y es el segundo factor que más incide en la reducción de la esperanza de vida ajustada a la salud a nivel global [7]. La Federación Internacional de la Diabetes (en inglés, *International Diabetes Federation*, (IDF)) sitúa la prevalencia de esta patología a nivel mundial en el 8.3 % de la población, lo que equivale a más de 422 millones de personas y se estima que este valor se incrementará hasta alcanzar el 10.1 % de la población en 2035 superando de esta forma los 592 millones de afectados en todo el mundo [8]. En particular, la prevalencia de DM en España alcanza el 6.66 % del total de la población asignada a la atención primaria del Sistema Nacional de Salud (SNS) siendo más común en hombres que en mujeres (7.27 % vs. 6.06 %) y aumenta con la edad hasta los 80 años [8].

Es evidente que la DM es una enfermedad compleja debido a su estrecha relación con el desarrollo de comorbilidades y complicaciones crónicas que no solo aumentan la mortalidad, sino que también generan un aumento en los costes para los sistemas de salud [9]. Esto se debe a una mayor utilización de recursos sanitarios y al incremento del riesgo de hospitalizaciones. Según las estimaciones, un paciente con diabetes consume entre 2 y 6 veces más recursos sanitarios directos en comparación con pacientes del mismo sexo y edades similares pero con otras enfermedades crónicas [9]. A nivel global, se estima que los gastos sanitarios asociados a la DM en 2021 ascendieron a 960 mil millones de dólares, lo que equivale aproximadamente al 15 % del gasto sanitario total en adultos [9]. En el caso de España. la DM se posiciona como una de las enfermedades crónicas más predominantes y los costes directos asociados a ella pueden representar entre un 7-13 % del gasto total del SNS [9].

Además de un aumento en el riesgo de presentar diversas complicaciones crónicas como alteraciones cardíacas, enfermedad renal crónica y daños neurológicos, más del 50 % de los pacientes diabéticos, presentan algún episodio de hipoglucemia durante la evolución de su enfer-

medad. Estos episodios deben ser reconocidos y tratados para evitar graves consecuencias [10]. La hipoglucemia representa el principal desafío para lograr el control glucémico adecuado, especialmente en pacientes con DMT1 [11]. En numerosas ocasiones, esta condición expone a los pacientes a riesgos graves, incluyendo la posibilidad de muerte. Se estima que aproximadamente el 35 % de los pacientes con DMT1 experimenta hipoglucemia grave, y entre el 4 y el 10 % de las muertes en pacientes con DMT1 se relacionan con la presencia de hipoglucemia [11]. Además del impacto clínico, estos episodios generan una carga económica significativa tanto en términos directos (costes de atención urgente y hospitalización) como indirectos (ausentismo laboral y disminución de la productividad) [11]. Por lo tanto, la prevención y el tratamiento de la hipoglucemia son imprescindibles tanto para preservar la salud y seguridad de los pacientes, como para mitigar el impacto económico asociado.

Para abordar el análisis de factores relacionados con la hipoglucemia, es necesario el manejo de grandes cantidades de información con el fin de optimizar los resultados. Afortunadamente, las tecnologías de la información han evolucionado y permiten acceder, producir y aprender de los datos. En los últimos años, el Aprendizaje Automático (ML por sus siglas en inglés, *Machine Learning*) se ha convertido en una herramienta ideal para mejorar los sistemas de salud mediante el análisis de datos y creación de modelos [12]. Con la creciente disponibilidad de información clínica, el ML se considera una gran herramienta para identificar patrones, variables y factores de riesgo asociados a la hipoglucemia en pacientes con DMT1. Estas técnicas pueden proporcionar conocimientos valiosos para la prevención y el tratamiento de la hipoglucemia, mejorando así la calidad de vida de los pacientes.

1.2. Objetivos

El objetivo de este trabajo es reducir la dimensionalidad de los datos mediante la identificación de factores asociados a los episodios de hipoglucemia en pacientes con DMT1 y la agrupación de los pacientes en subconjuntos con características similares. Estos factores incluyen miedos y preocupaciones de los pacientes, su estado emocional, una evaluación de sus capacidades cognitivas y los síntomas que experimentan. Para llevar a cabo este estudio, se utilizan datos clínicos recopilados de cinco bases de datos pertenecientes al estudio *T1D Exchange* realizado por el por el *Jaeb Center for Health Research* en Florida, Estados Unidos. Además, se emplean dos técnicas de ML para facilitar el proceso: Análisis Factorial Exploratorio (en inglés *Exploratory Factor Analysis*, (EFA)) y *clustering*. El objetivo de este TFG es de carácter múltiple, pudiendo dividirse en los objetivos que se exponen a continuación:

- Comprender diferentes técnicas de ML utilizadas en el análisis de datos clínicos y evaluar cómo se pueden aplicar de manera efectiva en el contexto de la hipoglucemia en pacientes con DMT1. Esto implica considerar factores como la estructura de las variables en las bases de datos previamente seleccionadas y los resultados deseados.

- Ajustar los parámetros necesarios de los algoritmos de ML seleccionados, en este caso EFA y *clustering*, con el fin de obtener resultados significativos y relevantes para el análisis de los datos clínicos.

- Aplicar el EFA con el objetivo de reducir la dimensionalidad de los datos e identificar factores que agrupen variables relacionadas con la hipoglucemia, y que presenten una relación significativa entre ellas.

- Aplicar un algoritmo de *clustering* a los resultados obtenidos del EFA para agrupar a los pacientes en subconjuntos con características similares en relación a la hipoglucemia.

- Realizar un análisis cualitativo de los resultados obtenidos, identificando patrones significativos entre las variables que evalúan aspectos relacionados con la hipoglucemia.

- Realizar una interpretación clínica de los resultados, relacinando los patrones identificados con la información clínica relevante y los conocimientos existentes sobre la hipoglucemia en pacientes con DMT1.

- Generar hipótesis para futuras investigaciones basadas en los patrones y las interpretaciones obtenidas, con el objetivo de ampliar el conocimiento y la comprensión de la hipoglucemia y mejorar las estrategias de prevenciín y tratamiento.

Así pues, los objetivos principales del presente estudio son dos. El primero es utilizar el EFA para establecer una una relación cuantitativa entre diferentes variables clínicas relacionadas con los episodios de hipoglucemia en pacientes con DMT1, y agruparlas en factores. El segundo objetivo es utilizar *clustering* para agrupar a los pacientes en subconjuntos con características similares, que no son fácilmente visibles mediante un análisis cualitativo inicial. Esto permitirá establecer relaciones significativas entre las variables, además de reducir la dimensionalidad de los datos para facilitar el análisis e interpretación de los resultados.

1.3. Metodología

Para el desarrollo del TFG se han utilizado los datos recopilados en el estudio *T1D Exchange* previamente mencionado. Estos datos consisten en cinco bases de datos derivadas de encuestas realizadas a pacientes con DMT1 que presentan tendencia a sufrir episodios de hipoglucemia. Estas encuestas evalúan aspectos como el estado emocional, las capacidades cognitivas, los síntomas, los miedos y las preocupaciones de los pacientes. Cabe destacar que los datos han sido anonimizados previamente y se presentan en el Capítulo 3 de este trabajo. Con el fin de alcanzar los objetivos descritos anteriormente, se ha llevado a cabo la siguiente metodología:

- Examinar diferentes bases de datos, tanto públicas como privadas y elegir las que mejor se ajusten mejor tanto al problema que se pretende abordar como a las técnicas de ML que se desean utilizar. En el caso de este TFG se han elegido 5 bases de datos que se consideraron adecuadas para la aplicación de las herramientas de ML seleccionadas en el estudio. Son bases de datos provenientes de encuestas que evalúan el estado emocional, las capacidades cognitivas, los miedos, las preocupaciones y los síntomas que experimentan los pacientes que sufren DMT1 y una tendencia a sufrir episodios hipoglucémicos.

- Revisar la literatura para entender los conceptos médicos necesarios para abordar el estudio relacionado con los episodios hipoglucémicos en pacientes con DMT1. En este caso se ha buscado información sobre las enfermedades crónicas, haciendo hincapié en la DMT1 y en los episodios de hipoglucemia severa asociados a ella.

- Revisar el estado del arte. En este punto se han leído y analizado los artículos que se han considerado necesarios para el desarrollo del TFG. Se trata de estudios en los que se hayan aplicado previamente los algoritmos de EFA y *clustering* sobre datos clínicos.

- Entender y profundizar en el conocimiento de las técnicas de ML elegidas. Es fundamental buscar información y comprobar que se cumplen las condiciones que indica cada técnica para justificar su uso sobre los datos que se tienen. En este caso, por ejemplo, antes de aplicar el EFA, se ha realizado un análisis de las variables que componen las bases de datos para ver si existe una relación significativa entre ellas.

- Preprocesamiento de los datos. Este punto es de los más importantes porque a menudo las bases de datos incluyen datos contaminados por factores negativos como el ruido, lo que conlleva un riesgo muy alto de obtener resultados poco fiables y por lo tanto, sin ningún significado relevante. Es por ello por lo que este paso ha sido imprescindible para

poder trabajar con datos limpios que prmitan obtener resultados significativos y facilitar la interpretación de los mismos para sacar conclusiones.

- Aplicar el EFA sobre cada una de las cinco bases de datos de manera independiente. Para cada base, se ha justificado la aplicación de dicha herramienta como corresponde. Los resultados obtenidos han permitido reducir la dimensionalidad de los datos y establecer relaciones significativas entre las variables.

- Aplicar un algoritmo de *clustering* sobre los datos obtenidos en el apartado anterior. Con esto se separan a los pacientes en subconjuntos con características similares en función de variables relevantes relacionadas con la hipoglucemia. El algoritmo de *clustering* asignará a cada paciente a un *cluster* específico, lo que permtirá una mejor comprensión de los diferentes perfiles y patrones asociados a la hipoglucemia en pacientes con DMT1.

- Concatenar las bases de datos con otra que contiene las etiquetas correspondientes a casos y controles de los pacientes. Este paso proporciona un soporte para el análisis de los resultados y se realiza después de la aplicación de los algoritmos con el fin de mantener el enfoque del estudio en métodos de ML petenecientes al aprendizaje no supervisado.

- Analizar e interpretar los resultados obtenidos. Para cada base de datos se han obtenido resultados razonables tanto en el EFA como en *clustering*. Se han conseguido ver relaciones significativas entre variables y diversas diferencias entre los perfiles de cada *cluster*.

- Concatenar las cinco bases de datos con el fin de aplicar las mismas herramientas de ML. Fue una estrategia empleada para combinar variables que evaluaban diferentes aspectos de los pacientes. Esta técnica permitió analizar de manera conjunta y más completa la información de las bases de datos, buscando patrones y relaciones que podrían pasar desapercibidas en el análisis individual de cada base de datos.

- Discutir los resultados y extraer conclusiones. Se examinan los factores y subconjuntos obtenidos tras aplicar las técnicas de ML. En el caso de los factores, se evalúan las ponderaciones de cada variable con respecto a su factor correspondiente, lo que ayuda a comprender su importancia en la formación de dicho factor. Por otro lado, en el caso de los *clusters* se realizan comparaciones entre los perfiles de cada *cluster* con el fin de identificar características comunes y diferencias significativas entre ellos. Este análisis conjunto proporciona una visión más completa de los datos, permitiendo obtener conclusiones relevantes que puedan ayudar en la toma de decisiones clínicas.

En la Figura 1.1 se presenta un diagrama de Gantt que detalla las diferentes tareas realizadas durante el presente TFG. Cabe mencionar que es una representación a modo orientativa, y en todo momento se ha intentado establecer el desarrollo temporal de cada tarea con la mayor exactitud posible.

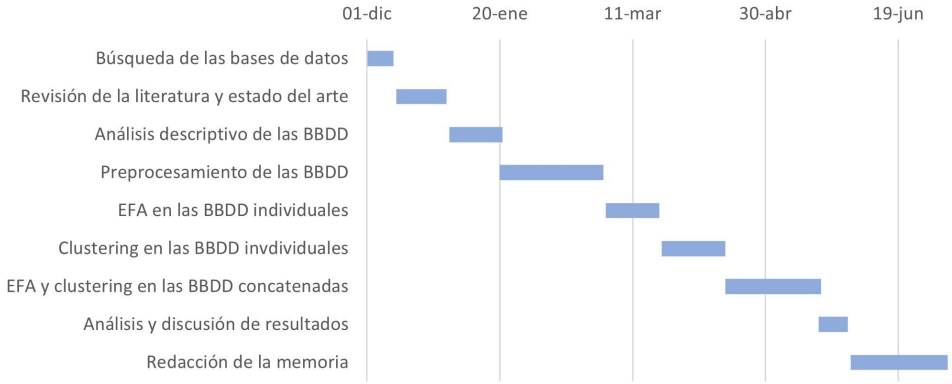


Figura 1.1: Diagrama de Gantt con la duración orientativa de las actividades realizadas.

1.4. Estructura de la memoria

A continuación se describe el contenido incluido en cada capítulo de la memoria:

- **Capítulo 1: Introducción y objetivos**. Este capítulo inicial se divide en 4 apartados. En primer lugar, se hace una contextualización del tema que se va a tratar para situar al lector, junto con una breve justificación del porqué es necesario abordar este tema. En el segundo apartado se exponen los objetivos del presente trabajo. En tercer lugar, se presentan tanto la metodología y los pasos que se han seguido para obtener los resultados finales como un diagrama de Gantt con las diferentes actividades realizadas. Por último, se expone la estructura diseñada para la elaboración de la memoria.

- **Capítulo 2: Conceptos previos**. Se presentan los conceptos necesarios para la comprensión de este TFG. En primer lugar, se exponen los términos clínicos relacionados con la diabetes y se describen sus diferentes tipos. A continuación, se aborda la hipoglucemia, un fenómeno que afecta a pacientes con DMT1 y se explican sus características y consecuencias. Por último, se describen los conceptos y herramientas de ML utilizadas en el estudio, proporcionando una visión enfocada en el análisis de los datos relacionados con los episodios de hipoglucemia.

- **Capítulo 3: Base de datos y análisis descriptivo**. En este capítulo se proporciona una descripción detallada de las bases de datos utilizadas en el estudio, incluyendo información sobre las variables que las componen y el proceso de preprocesamiento necesario antes de comenzar el análisis con dichos datos. En este capítulo se explicará en detalle cómo se llevó a cabo el preprocesamiento en cada base de datos.

- **Capítulo 4: Experimentos y resultados**. Se presentan los experimentos realizados y se detallan los pasos empleados en el análisis de los datos recopilados. Se lleva a cabo un análisis exhaustivo de los resultados obtenidos, resaltando los hallazgos más significativos y proporcionando una interpretación detallada de los mismos. Esto permite establecer las bases para conclusiones fundamentales, contribuyendo al conocimiento en el área de estudio y brindando información relevante para futuras investigaciones.

- **Capítulo 5: Conclusiones y líneas futuras**. Se realiza un resumen con los resultados y las conclusiones obtenidos gracias al presente estudio. Además, se comentan brevemente varias líneas de investigación futuras.

Capítulo 2

Conceptos previos

En este capítulo se introducen una serie de conceptos fundamentales que ayudan a entender el marco clínico en el que se desarrolla este TFG. En primer lugar, se explican los conceptos de DMT1 e hipoglucemia severa, que son los focos principales de este trabajo. A continuación, se exponen los conceptos de ML que se han considerado necesarios para la comprensión de este TFG. Se presentan EFA y métodos de *clustering* considerados, proporcionando una explicación detallada de cada una de ellas.

2.1. Diabetes mellitus e hipoglucemia severa

La DM es una enfermedad crónica que describe un desorden metabólico de múltiples etiologías y puede ser debida a una insuficiencia en la segregación de la insulina por parte del páncreas y/o a una utilización ineficaz de la misma por parte del organismo [13]. La insulina es una hormona que regula la concentración de glucosa en sangre o glucemia. Una diabetes no controlada adecuadamente puede estar asociada a diversas complicaciones, tanto agudas como crónicas [13]. En el caso de los episodios de hipoglucemia, que son el enfoque de este estudio, se tratan como una complicación aguda. Por otro lado, las complicaciones crónicas pueden ser de naturaleza micro o macrovascular y están relacionadas con daños graves en los órganos y sistemas del organismo. Estas complicaciones pueden incluir insuficiencia renal, ceguera, infarto de miocardio y amputación de extremidades inferiores [13].

La DM ha alcanzado unas cifras alarmantes, afectando al 5.1 % de las personas de 20 a 79 años de edad a nivel mundial, siendo la DMT2 responsable del 90 % de los casos [14]. En el año 2019, la DM fue la causa directa de 1.5 millones de defunciones de las cuales el 48 %

pertenecieron a personas cuya edad no superaba los 70 años [13]. Además, en el mismo año, se han registrado 460,000 fallecimientos a causa de complicaciones y enfermedades asociadas a la DM [15]. Según la OMS, entre los años 2000 y 2019, la probabilidad de fallecer debido a alguna de las enfermedades crónicas más comunes se redujo en un 22 %. Sin embargo, en el caso específico de la diabetes, las estadísticas no reflejan dicha mejora, ya que las tasas de mortalidad por esta enfermedad aumentaron en un 3 % [16]. Además, en los países menos favorecidos económicamente, este porcentaje de aumento en la mortalidad por diabetes ascendió en un 13 % [16]. Existen tres tipos principales de diabetes que son la DMT1, la DMT2 y la diabetes gestacional. En cuanto a los síntomas de esta enfermedad, estos pueden ocurrir repentinamente y en la DMT2, pueden ser ligeros y tardar muchos años en manifestarse. Los síntomas principales de la diabetes son: sensación de mucha sed, poliuria, visión borrosa y pérdida de peso involuntaria. Entre las manifestaciones clínicas más graves se encuentra la cetoacidosis o un estado hiperosmolar no cetósico que puede conducir a deshidratación, coma y, en ausencia de tratamiento eficaz, incluso puede llegar a producir la muerte [15].

La DMT1, también llamada insulinodependiente, juvenil o de inicio en la infancia, se define como una alteración producida por una reacción autoinmunitaria, es decir, el cuerpo se ataca a sí mismo por error impidiendo la producción de insulina [17]. En el contexto mundial, este tipo de DM tiene mucha menor incidencia que la DMT2, pues se estima que aproximadamente del 5 al 10 % de las personas que tienen DM presentan este tipo [17]. A diferencia de la DMT2, los síntomas de esta patología aparecen rápido y las personas más afectadas son los niños, los adolescentes y los adultos jóvenes, aunque puede aparecer a cualquier edad. Estos pacientes deben recibir insulina todos los días como tratamiento de sustitución hormonal [17]. Este tipo de DM se produce debido a la destrucción de las células beta del páncreas, lo que provoca una deficiencia de insulina que al principio puede no presentarse de forma grave, pero que evoluciona con el tiempo hasta alcanzar la carencia absoluta de esta hormona. Lo que determina la intensidad y por lo tanto, la gravedad del cuadro clínico, es la velocidad de destrucción de las células beta pancreáticas: si el cuadro se inicia a temprana edad, la intensidad será mayor [18].

La mayoría de los pacientes con DMT1 experimentan con facilidad episodios de hipoglucemia e hiperglucemia, ambos considerados producto de una inestabilidad metabólica. Además, algunos pacientes con esta condición pueden presentar obesidad antes o después de comenzar el tratamiento, lo que sugiere la posible influencia de factores genéticos que predisponen a la resistencia a la insulina o a la obesidad [18]. La susceptibilidad genética para desarrollar DMT1 se asocia a algunos antígenos de histocompatibilidad como por ejemplo los leucocitarios (en inglés *Human Leukocyte Antigen*, HLA) [13]. Sobre esta base genética, actúan otros factores ambientales que pueden favorecer la expresión de la enfermedad. Entre estos factores se encuen-

tran los hábitos alimentarios, el estrés, el crecimiento acelerado que ocurre en la pubertad y la contaminación, esta útlima considerada como una posible causa del incremento en el número de nuevos casos diagnosticados en los últimos años [13].

Actualmente, se considera que una persona presenta DMT1 cuando el 90 % de sus células beta pancreáticas han sido destruidas, pero la evolución de la enfermedad se da por etapas y la detección precoz es imprescindible para evitar graves consecuencias [13]. No obstante, no se sabe cómo se produce este proceso de autodestrucción ni cuál es el factor desencadenante, además de tratarse de un proceso irreversible porque las células beta no se regeneran [13]. A día de hoy, no hay cura disponible para esta patología y los pacientes que la padecen, deben depender de las inyecciones de insulina para sobrevivir. Además, deben controlar sus niveles de azúcar en sangre regularmente durante el día [19]. Sin embargo, se está avanzando tanto en la mejora de estrategias que favorecen el diagnóstico precoz de DMT1 en su fase preclínica, como en el desarrollo de terapias inmunológicas que son capaces de retrasar la progresión de este tipo de DM y reducir la prevalencia de cetoacidosis [19].
A finales del año 2022, la Administración de Alimentos y Medicamentos (en inglés *Food and Drug Administration*, FDA) ha aprobado el primer medicamento que puede retrasar la aparición de la DMT1 [20]. Se trata de un fármaco que se une a determinadas células inmunitarias del organismo con el objetivo de retrasar la progresión de la enfermedad. Este medicamento es capaz de desactivar las células del sistema inmune que atacan a las células beta pancreáticas, aumentando a la vez el número de células que ayudan a moderar la respuesta inmunitaria. No obstante, el uso de este remedio viene con precauciones y advertencias debido a los efectos secundarios que presenta [21].

En las personas que sufren DMT1, la destrucción autoinmune de las células beta pancreáticas genera la necesidad de tratamiento a través de la administración de insulina para evitar el desarrollo de complicaciones [22]. No obstante, esta terapia conlleva un alto riesgo de hipoglucemia que es una alteración cuyo diagnóstico tiene lugar cuando la concentración de glucosa en sangre es inferior a 70 mg/dl. Esta situación puede darse con o sin síntomas siendo el último caso un peligro para la vida del paciente, además de comprometer su calidad de vida [22]. Más del 50 % de los pacientes diabéticos presentan episodios de hipoglucemia en algún momento de su vida. Se estima que estos pacientes pueden llegar a tener hasta dos episodios leves por semana y uno de hipoglucemia severa al año. El principal problema de la hipoglucemia es que muchas veces pasa desapercibida, ya sea por una falta de monitorización de los niveles de glucosa en sangre o por episodios que son asintomáticos [23].

La hipoglucemia es una preocupación principal tanto para los pacientes como para los equi-

pos clínicos que utilizan insulina en el tratamiento de la diabetes. Esta condición funciona como una barrera para lograr un control óptimo de los niveles de glucosa en sangre y se considera una emergencia médica verdadera que debe ser tratada de forma rápida para evitar complicaciones, que incluyen eventos cardiovasculares, daño neurológico, traumas e incluso la muerte [24]. Más del 5 % de las muertes relacionadas con la diabetes se atribuyen a episoodios de hipoglucemia severa [10]. Según la Asociación Estadounidense de Diabetes (en inglés, *American Diabetes Association*, ADA), la hipoglucemia se puede clasificar en varios tipos atendiendo a la gravedad del cuadro clínico presentado [25]:

1. **Hipoglucemia leve**. Sin compromiso de las habilidades neurológicas, y donde el paciente puede resolver la situación sin grandes complicaciones (autónoma).
2. **Hipoglucemia moderada**. Existe un grado de afectación neurológica pero el paciente puede resolver la situación de forma autónoma.
3. **Hipoglucemia grave**. La consciencia del paciente se ve comprometida, y es necesaria la intervención de terceros para solucionar la situación.

El diagnóstico de la hipoglucemia se realiza mediante la tríada de Whipple [26] que incluye una clínica compatible, una baja concentración de glucosa plasmática y la desaparición de los síntomas después de la administración de carbohidratos. Las manifestaciones clínicas de esta alteración se pueden clasificar en distintas categorías [26]:

- **Síntomas neurogénicos o autonómicos**. Con glucemias inferiores a 60 mg/dl. Se distinguen señales como palpitaciones, temblores, palidez, sudoración y parestesias.
- **Síntomas neuroglucopénicos**. Con glucemias inferiores a 50 mg/dl. Se ven como trastornos de conducta, cefalea, confusión, convulsiones, pérdida de conocimiento y muerte cerebral o coma, en los casos más extremos.

La gravedad de un episodio está relacionado con la aparición de los síntomas neuroglucopénicos que son el resultado de un déficit cerebral de glucosa [26]. Diferentes estudios han demostrado que la hipoglucemia severa tiene una repercusión negativa sobre la calidad de vida de los pacientes, ya que el bienestar de estos puede verse afectado tanto por los episodios de hipoglucemia como por el miedo a que estos ocurran [10]. Es por ello por lo que la mayoría de las personas que tienen tendencia a experimentar estos episodios, sufren depresión y alteraciones crónicas en el estado de ánimo [10]. Por otra parte, también se ha demostrado que las

capacidades cognitivas pueden verse afectadas por la hipoglucemia e incluso, en los casos más extremos, los pacientes desarrollan demencia con el paso del tiempo [10].

Para la prevención de la hipoglucemia, se llevan a cabo diferentes estrategias que incluyen [27]: (1) la educación del paciente y su familia; (2) la monitorización continua de la glucemia; y (3) la modificación de los suministros de insulina (cuando es necesario). La educación es clave para reducir el miedo de los pacientes y de esta forma evitar los trastornos psicológicos que pueden llegar a desarrollar. El médico ha de realizar una evaluación de los pacientes para identificar aquellos que tienen una mayor probabilidad de sufrir hipoglucemia. Por su parte, el paciente debe conocer su tratamiento y la importancia de seguir las indicaciones. La monitorización en casa, por ejemplo, es una herramienta muy útil para detectar la hipoglucemia a tiempo y así evitar complicaciones [27]. En cuanto al tratamiento, depende de la gravedad de la hipoglucemia, pues para los episodios leves suele ser el consumo de azúcar o carbohidratos de acción rápida, mientras que en una hipoglucemia severa se ha de suministrar glucagón por vía subcutánea o intramuscular [27]. No obstante, en ambos casos, una vez recuperado el paciente, debe ingerir una determinada cantidad de hidratos de carbono, además de analizar la causa del episodio para tomar las medidas necesarias con el fin de que no se vuelva a dar [27].

En los últimos años, ha habido un avance importante en el desarrollo tecnológico para tratar tanto los episodios de hipoglucemia como la DM de forma general. Un ejemplo son los sensores que ayudan a tener una monitorización de los niveles de glucosa en sangre [28]. Además, existe una tecnología bastante prometedora que es un páncreas artificial cuyo objetivo es el de eliminar, o por lo menos mejorar el pronóstico de las hipoglucemias. La finalidad principal es conseguir la interrupción del suministro de insulina cuando los niveles de glucemia están por debajo de lo normal, llegando así a minimizar la duración de los episodios de hipoglucemia [29].

Dentro de las complicaciones asociadas a la DM, se ha puesto un enfoque especial en el estudio de los episodios hipoglucémicos. El objetivo principal es analizar diversos aspectos como los síntomas, las capacidades cognitivas, el estado emocional, los miedos y las preocupaciones de los pacientes que son propensos a experimentar estos espisodios. El propósito es identificar características comunes entre ellos y comprender mejor esta complicación. Lo que ha facilitado el proceso para alcanzar estos objetivos, es el uso que se ha hecho de las herramientas de ML para el análisis de datos que se presentan en el siguiente apartado.

2.2. Conceptos de aprendizaje automático

El aprendizaje automático o ML es un concepto que surgió en 1.959 y el pionero en su uso, Arthur Samuel, lo ha definido como «una tecnología que utiliza técnicas estadísticas y algoritmos computacionales para proporcionar a los ordenadores la capacidad de aprender, es decir, mejorar sus resultados en una tarea específica tras procesar datos en suficiente cantidad y sin intrucciones explícitas externas proporcionadas por el programador» [30]. Este concepto se considera una rama de la inteligencia artificial ya que aprende analizando datos y genera decisiones a través de algoritmos que son procedimientos matemáticos que describen las relaciones entre variables [30]. Los métodos de ML se clasifican principalmente en tres grandes grupos que son los que se exponen a continuación:

- **Aprendizaje supervisado**. Se entrena un modelo proporcionándole características de entrada junto con etiquetas o resultados predefinidos por humanos [30]. El objetivo es que el modelo adquiera conocimiento de las relaciones entre las características y sus etiquetas correspondientes, obteniendo experiencia a partir de los datos proporcionados y aprendiendo patrones para luego realizar predicciones. Se distinguen dos tipos [31]:
  - **Clasificación**. El conjunto de etiquetas es numerable. Pueden ser dos o más, pero siempre numerables. Un ejemplo puede ser el número de unos objetos determinados.
  - **Regresión**. El conjunto de etiquetas no es numerable. Hay infinitos valores a utilizar. Un ejemplo puede ser la temperatura.
- **Aprendizaje no supervisado**. A diferencia del aprendizaje supervisado, no se le proporciona información de etiqueta al modelo. En cambio, se introducen grandes cantidades de datos sin etiqueta y el modelo debe descubrir patrones o tendencias ocultas con el fin de separar la información en grupos de manera automática [30].
- **Aprendizaje por refuerzo**. Técnica en la que el modelo recibe datos tanto etiquetados como no etiquetados. El sistema interactúa con el entorno, recibe *feedback* y aprende a través del proceso de prueba y error. En otras palabras, en esta técnica el modelo adquiere conocimiento a partir de experiencias pasadas y comienza a adaptar su funcionamiento para producir la salida óptima [32].

La inteligencia artificial y diversas ténicas de ML se están introduciendo en el campo de la salud de forma progresiva. Existen numerosos estudios que usan ML para diagnósticos y

pronósticos, aunque actualmente siempre siguen sujetos a una validación por parte los clínicos. Estas herramientas resultan de gran utilidad sobre todo para ahorrar tiempo en las tareas que son laboriosas o repetitivas y numerosas fuentes afirman que en unos años acabarán cambiando la forma convencional en la que se ha venido practicando la medicina hasta el momento [12].

El presente trabajo se basa en dos modelos de aprendizaje no supervisado, pues no se le ha proporcionado previamente la salida o etiqueta asociadas ni a cada paciente ni a cada variable. Se está abordando una tarea en la que se van a usar dos algoritmos que son el EFA y *clustering* para alcanzar los objetivos previamente expuestos. Para aplicar estas herramientas, es necesario comenzar con una base de datos preprocesada para evitar que factores negativos como ruido o valores nulos distorsionen los resultados. A continuación, se presentan algunos conceptos de ML relacionados con las variables que servirán como base para entender todo el preprocesamiento descrito en el Capítulo 3 de este trabajo.

Tipos de variables y métodos de normalización

Los modelos de ML se emplean para extraer información de observaciones, y en el caso del aprendizaje supervisado, de etiquetas. Las bases de datos son consideradas como el mejor método para almacenar estas observaciones junto con sus etiquetas correspondientes, cuando sea aplicable. En este TFG, se trabajará con bases de datos que contienen información estructurada de pacientes, con variables que pueden ser de diversos tipos [33]:

- **Variables cualitativas o categóricas.** Son aquellas que describen características como cualidades o atributos. Pueden ser *nominales*, donde las categorías no tienen un orden específico, como el género o el estado civil, u *ordinales*, donde las categorías sí tienen un orden específico, como el nivel educativo.
- **Variables cuantitativas o numéricas.** Son las que describen una característica en términos de cantidades o valores numéricos.

La normalización de variables es una etapa común en el preprocesamiento de bases de datos y se realiza antes de aplicar técnicas de ML. Su objetivo es ajustar los valores de las columnas numéricas a una escala común, sin alterar las diferencias en los intervalos ni la información que aporta cada variable. En este trabajo se han utilizado dos tipos de normalización [34]:

- **Normalización Min-Max**. Transforma los datos a un rango específico, generalmente en-

tre 0 y 1. La fórmula utilizada para esta normalización es la siguiente:

$$\mathbf{x}_{\text{norm}} = \left(\frac{\mathbf{x} - \text{mín}(\mathbf{x})}{\text{máx}(\mathbf{x}) - \text{mín}(\mathbf{x})} \right)$$

donde x es el valor original, $\mathbf{x}_{norm}$ es el valor normalizado, $min(\mathbf{x})$ es el valor mínimo en el conjunto de datos y $max(\mathbf{x})$ es el valor máximo del mismo conjunto. En este caso, los conjuntos de datos corresponden a los datos que se recogen en las columnas de cada base de datos.

- **Normalización estándar.** Transforma los datos de manera que tengan una distribución con media 0 y desviación estándar 1. Esto permite comparar y analizar las variables en la misma escala. La fórmula correspondiente a este método de normalización es la que se expone a conitnuación:

$$\mathbf{z} = \frac{\mathbf{x} - \text{mean}(\mathbf{x})}{\text{std}(\mathbf{x})}$$

donde la $\mathbf{z}$ representa los valores normalizados, $mean(\mathbf{x})$ sirve para llamar a la media de los datos y $std(\mathbf{x})$ es la desviación estándar de las variables.

Ambos métodos de normalización, son ampliamente utilizados en el preprocesamiento de datos para asegurar una escala adecuada antes de entrenar modelos de ML. La elección entre uno u otro depende del contexto y del problema específico que se está abordando [35].

2.3. Análisis factorial exploratorio

El EFA tuvo sus comienzos a principios del siglo XX y se ha convertido en una de las técnicas estadísticas más empleadas en la actualidad. Esta herramienta es ampliamente utilizada en diversas áreas de investigación, especialmente en el ámbito de las ciencias sociales [36]. Su principal objetivo es reducir la dimensionalidad de los datos e identificar factores latentes que explican las relaciones entre las variables observadas, facilitando así la comprensión de los patrones subyacentes en los datos [36]. A diferencia de los modelos de regresión tradicionales, el EFA se caracteriza por su enfoque exploratorio e interdependiente, sin establecer hipótesis previas ni distinguir entre variables dependientes e independientes. En su lugar, se busca descubrir patrones emergentes y estructuras ocultas en los datos [37].

Este análisis se puede utilizar en diversos casos. En primer lugar, resulta ideal para identificar las variables más relevantes en una base de datos, las cuales pueden ser utilizadas posteriormente en otros tipos de análisis [38]. Además, el EFA es especialmente útil en el análisis

de cuestionarios, pues permite determinar la posibilidad de agrupar los datos recopilados en encuestas en factores que tengan una interpretación teórica coherente [38].
Muchas veces, se confunde el EFA con el Análisis de Componentes Principales (en inglés *Principal Component Analysis*, PCA), pero ambas técnicas difieren en su enfoque teórico y matemático. El EFA se basa en un modelo reflectivo, donde el objetivo es descubrir la estructura latente de los datos identificando factores que expliquen las correlaciones entre las variables observadas. Por otro lado, en PCA se busca encontrar combinaciones lineales de las variables originales que expliquen la mayor cantidad de variabilidad en los datos [39]. A diferencia del EFA, PCA no tiene como objetivo descubrir factores teóricos subyacentes, sino encontrar una representación compacta de los datos utilizando los componentes principales [39]. Otra diferencia significativa entre ambas técnicas se encuentra en la interpretación de los resultados: en PCA, los componentes son combinaciones lineales de las variables originales y no siempre tienen una interpretación teórica directa, mientras que en el EFA, los factores identificados tienen una explicación teórica más clara [39].

Antes de utilizar el EFA, es importante evaluar el número y el tipo de variables con las que se va a trabajar. Esta técnica requiere generalmente un tamaño muestral considerable, ya que las correlaciones son menos confiables cuando se usan muestras pequeñas [40]. Las recomendaciones sobre el tamaño de muestra adecuado pueden variar según la situación, pero se sugiere que una cifra entre 150 y 300 casos puede proporcionar buenos resultados [40]. Además, es necesario realizar un análisis preliminar para determinar si existen razones suficientes que justifiquen el uso de esta técnica en cada caso específico, ya que solo en determinadas condiciones se puede utilizar el EFA como herramienta de análisis. Para empezar, la primera condición que se debe cumplir es que las variables presenten un determinado grado de correlación. Para medir esta correlación se pueden utilizar varias técnicas [40]:

- Prueba de esfericidad de Bartlett [41]. Se evalúa una hipótesis nula que establece que no hay correlación entre las variables. El objetivo es demostrar que la matriz de correlaciones no es una matriz identidad y que, por lo tanto, sí existe correlación entre las variables [41]. Como resultado de esta prueba, se obtienen 2 valores que son el chi-cuadrado y el p_{valor}. En general, un valor alto de chi-cuadrado y un p-valor cercano a 0 sugieren la existencia de una correlación significativa entre las variables, lo que indica que el EFA puede ser apropiado. No hay un valor específico de chi-cuadrado que sirva como umbral para tomar la decisión de aplicar el EFA, pero es importante que el p_{valor} sea inferior a 0.05, lo que indica que hay evidencia suficiente para rechazar la hipótesis nula [42].
- Evaluación de la fuerza de la relación entre dos variables basada en las correlaciones par-

ciales entre ellas, una vez eliminada la influencia de las demás variables. Para llevar a cabo esta evaluación, se utiliza un índice llamado *Kaiser Mayer Olkin* (KMO), el cual varía entre 0 y 1 [40]. El KMO proporciona una medida de la adecuación de las correlaciones entre las variables. Si las correlaciones son lo suficientemente altas, se consideran adecuadas para la realización del análisis factorial. Sin embargo, es importante tener en cuenta que el índice KMO es sensible al tamaño de la muestra, aumentando con muestras más grandes y disminuyendo cuando las correlaciones son más bajas. Para su interpretación se utiliza la siguiente escala [43]:

  - **Valores inaceptables.** Aquellos que son inferiores a 0.5.
  - **Valores regulares.** Aquellos que van de 0.5 a 0.59.
  - **Valores pobres.** Aquellos que van de 0.6 a 0.79.
  - **Valores meritorios.** Aquellos que van de 0.8 a 1.

- Utilización del índice de la muestra individual (MSA), que se deriva del KMO, pero en este caso evalúa la relación que presenta una sola variable con las demás. La interpretación de los valores es la misma que se mencionó anteriormente, donde valores más altos indican una relación más adecuada [40].

- Calcular el determinante de la matriz de correlaciones. En este caso, se busca que el determinante tenga un valor pequeño pero no igual a 0. Si el valor del determinante se acerca a 1, indica independencia entre las variables y, por lo tanto, el EFA no sería adecuado debido a la falta de correlación entre las variables [40].

Es recomendable aplicar al menos dos de las medidas expuestas anteriormente para justificar el uso del EFA. Si se cumplen las condiciones establecidas por cada medida, entonces se puede proceder con el análisis. En caso contrario, se debe considerar la aplicación de otro tipo de análisis estadístico para los datos presentes [40]. En este TFG se han usado dos de estas medidas. La primera es la prueba de esfericidad de Bartlett, y como segunda medida se ha utilizado el índice de Kaiser Mayer Olkin (KMO). Estas dos han sido seleccionadas porque se busca evaluar la correlación entre todas las variables en vez de analizar únicamente la correlación de una variable con respecto a las demás. Por lo tanto, estas métricas permiten determinar si las variables tienen una relación significativa en conjunto.

En relación a los datos, estos deben ser de tipo numérico para poder realizar un análisis adecuado y obtener resultados más significativos. En caso de que algunas variables sean categóricas, se debe aplicar la codificación que corresponde en cada caso para convertirlas en

variables numéricas. Es importante tener en cuenta que no es apropiado aplicar esta técnica en variables que sean muy heterogéneas entre sí, es decir, si presentan una gran diversidad o no tienen una relación clara. En tales casos, se recomienda realizar un análisis por separado para cada grupo de variables con una mayor correlación entre ellas [44].

Una vez vez justificada la aplicación de este tipo de análisis a las variables disponibles, es necesario especificar el método estadístico a utilizar para extraer los factores y determinar su número [40]. Entre las diversas metodologías de extracción de factores, las más utilizadas son el análisis de componentes principales y el análisis de factores comunes [40]. Para elegir entre ambos métodos, es necesario tener una idea de la dispersión de las variables, ya que a mayor interrelación entre las variables, mayor será la varianza que comparten. La varianza total de cualquier variable puede descomponerse en tres partes que, al sumarlas, forman la varianza total [40]:

- **Varianza común o conmunalidad.** Es la varianza de una variable compartida con el resto de variables que se están considerando en el análisis.
- **Varianza específica o unicidad.** Es la varianza que solo depende de la variable en cuestión y no puede explicarse por medio de otras.
- **Error de varianza.** Es la varianza que no se puede explicar por los factores comunes ni por las características específicas de las variables. Se debe al error aleatorio.

Volviendo a la selección entre el análisis de componentes principales y el análisis de factores comunes para la extracción de factores, es importante considerar los objetivos del estudio y el conocimiento previo de las variables, especialmente en términos de su variabilidad [40]. Cuando se busca reducir la dimensionalidad de las variables y/o la unicidad y que el error de varianza sea bajos, es más apropiado utilizar el análisis de componentes principales. Este método considera la varianza total y los factores resultantes tienen una pequeña varianza única [40].
En cambio, si el EFA se utiliza con el objetivo de crear nuevos factores, es más adecuado utilizar el análisis de factores comunes, que se basa en la conmunalidad y no considera relevante el error de varianza y la unicidad. No obstante, este método puede presentar inconvenientes cuando la varianza común entre las variables es insuficiente, lo que puede requerir la eliminación de una o más variables para asegurar resultados confiables. Además, puede generar resultados no únicos para las cargas factoriales, lo que implica que diferentes configuraciones de cargas pueden llevar a resultados similares. Por lo tanto, es necesario considerar estas limitaciones al interpretar los resultados del análisis de factores comunes [40].

Después de obtener los factores mediante uno de estos métodos, es necesario determinar el número de factores y el tipo de rotación que se ha de utilizar para el análisis, ambos explicados a continuación. Los factores buscan encontrar la mejor combinación lineal que explique la mayor variabilidad entre las variables originales. El primer factor explica la mayor varianza, seguido por el segundo factor que es ortogonal al primero y así sucesivamente [45]. A la varianza para el total de variables que puede ser explicada por cada uno de los factores, se le conoce como *eign-value* o valor propio. En cuanto a la elección del número de factores, se pueden usar distintos criterios de los cuales los más comunes son [45]:

- **Gráfico *scree plot* o gráfico de autovalores.** Representación de los autovalores en el eje de ordenadas y del número de factores en el eje de abscisas. La curva de autovalores se presenta en orden descendente, de mayor a menor. En este estudio, se ha considerado óptimo el número de factores correspondiente al último autovalor que sea mayor que 1, ya que indica una mayor explicación de la varianza [46].

- **Criterio del test de pendiente.** Depende de los valores propios pero estos son graficados permitiendo así realizar un análisis visual con el fin de encontrar un punto de inflexión donde la curva cambie de sentido. Una limitación que tiene es la subjetividad porque la interpretación cambia de una persona a otra, lo que hace que el número de factores dependa básicamente del criterio del investigador.

- **Criterio del porcentaje de varianza.** Establece previamente el porcentaje de la varianza total mínima que debería ser explicada por los factores, por lo que la selección final corresponde al número de factores que cumple esta condición. Su limitación es que se pueden tomar muchos más factores de los necesarios.

- **Criterio *a priori.*** Es de los más subjetivos y se usa en casos muy concretos. Consiste en que el investigador estime *a priori* el número de factores que considere adecuado.

Una vez determinado el número de factores, uno de los pasos finales pero más importantes del EFA es la interpretación de los resultados obtenidos que depende principalmente de la experiencia y los conocimientos del investigador o experto de dominio. Para realizar una interpretación adecuada, se recomienda seguir los siguientes pasos [40]:

- **Estimar la matriz de factores.** Se calcula la matriz de factores (resultado del EFA) que contiene las ponderaciones o pesos que representan la correlación de cada variable con su factor asociado. Valores altos en términos absolutos indican una mayor contribución de

la variable al factor, mientras que los valores bajos son indicadores de una baja representatividad del factor por parte de la variable. Este paso ayuda a reducir la dimensionalidad de los datos. Sin embargo, si el objetivo es buscar nuevos factores, es necesario realizar una rotación de los ejes de los factores obtenidos en este primer paso.

- **Rotación de factores.** Hace referencia a girar los ejes factoriales a distintos grados, pero manteniendo fijo el origen. El objetivo es obtener una redistribución de la varianza de las variables originales en los factores, lo que facilita una mejor interpretación de los resultados. En la actualidad, existen dos tipos de rotaciones utilizadas: rotación ortogonal y rotación oblicua. La elección entre estos métodos depende del conocimiento del investigador y la naturaleza del problema en estudio.

  - Ortogonales. Los factores se giran simultáneamente manteniendo su independencia. Un ejemplo de rotación ortogonal es la rotación *varimax*, que tiene como objetivo maximizar las ponderaciones de los factores de manera que cada variable esté asociada principalmente a un solo factor, lo que ayuda a reducir al mínimo el número de variables con carga significativa en cada factor [43].
  - Oblicuas. El giro no mantiene la independencia entre factores. Una de las rotaciones más utilizadas en este tipo es la *promax* que va modificando los resultados de una rotación ortogonal hasta crear una matriz de pesos lo más parecida posible a una estructura ideal. Para lograr esto, se elevan las cargas o pesos de los factores a una potencia específica, generalmente entre 2 y 4. Cuanto mayor sea la potencia, mayor será el grado de oblicuidad en la solución [43].

  El tipo de rotación adecuado dependerá del problema que se está abordando. Es por ello por lo que se debe tener un conocimiento previo acerca del tema, pues si se aprecia una posible correlación entre los factores es mejor utilizar una rotación oblicua y en el caso contrario, suponiendo independencia, es mejor optar por una rotación ortogonal [43].

Para interpretar los resultados y realizar un análisis estadístico adecuado, es importante tener en cuenta la significancia de las ponderaciones en términos absolutos: valores inferiores a 0.3 se consideran no significativos, valores entre 0.3 y 0.5 se consideran de aporte mínimo, valores entre 0.5 y 0.7 se consideran de aporte significativo y por último valores superiores a 0.7 se consideran relevantes y son el objetivo del análisis [40].

En el campo de la salud, el EFA tiene muchas aplicaciones como método estadístico. Se utiliza con frecuenica en estudios de investigación para comprender la estructura subyacente

de los datos. Es especialmente útil en el análisis y validación de cuestionarios y medidas psicométricas utilizadas en la evaluación de síntomas, calidad de vida y salud mental, entre otros aspectos [47]. El EFA puede proporcionar información importante sobre la estructura de los datos y las relaciones entre variables, permitiendo la agrupación de factores que pueden estar asociados con diferentes aspectos de la salud. Esto contribuye a la reducción de la dimensionalidad de los datos y facilita la interpretación de los resultados [47]. En la revisión del estado del arte del EFA, se han encontrado varios estudios clínicos en los que se ha aplicado esta herramienta como método de análisis estadístico. Por ejemplo en el estudio titulado *Factor analysis of the COVID-19 Perceived Risk Scale: A preliminary study* de Murat Yıldırım y Abdurrahim Güler [48] se usa el EFA para examinar la estructura de la Escala de Percepción del Riesgo de COVID-19 con el objetivo de identificar los factores subyacentes que contribuyen a la percepción del riesgo en relación con la enfermedad. Como se ha llevado a cabo durante la pandemia se ha hecho uso de las redes sociales para la recopilación de los datos seleccionando de todos los participantes, en concreto 3.109 personas. Como resultado, se han encontrado dos factores que se han identificado como dimensión cognitiva y dimensión emocional. Esto ha permitido una comprensión de los factores que influyen en los comportamientos relacionados con la pandemia y la información recopilada se puede usar para la planificación de estrategias de prevención adaptadas a las necesidades de la población.

Otro estudio en el que se ha aplicado EFA en el campo de la salud es *Exploratory factor analysis determines latent factors in Guillain–Barré syndrome* de Seiichi Omura et al. [49]. Se ha utilizado esta herramienta para encontrar los factores subyacentes en el síndrome de Guillain-Barré, una enfermedad de origen neurológico. En el estudio, se han identificado 5 factores que son consistentes con los hallazgos experimentales y clínicos previos. Con estos resultados, se ha demostrado que el EFA puede aplicarse a datos médicos para extraer factores latentes en otras enfermedades con causas desconocidas. Este hallazgo tiene implicaciones importantes, ya que puede contribuir a mejorar el diagnóstico, el pronóstico y el desarrollo de tratamientos más personalizados para el síndrome de Guillain-Barré y otras enfermedades similares.

Por último, en el estudio titulado *Multidimensional symptom clusters: an exploratory factor analysis in advanced chronic kidney disease* de Hayfa Almutary et al. [50] se utiliza EFA con el objetivo de agrupar en factores los síntomas de los pacientes con enfermedad renal crónica avanzada. Los 436 participantes completaron una encuesta que evaluaba cuatro dimensiones de los síntomas: ocurrencia, malestar, gravedad y frecuencia. Se ha conseguido una agrupación significativa de los síntomas, lo que puede contribuir en el desarrollo un tratamiento más eficaz para reducir los síntomas y por tanto, mejorar la calidad de vida de los pacientes.

2.4. Métodos de *clustering*

Después de una era en la que el desafío era lidiar con la recopilación de datos, hoy en día el problema ha cambiado a cómo procesar las enormes cantidades de datos que se obtienen [51]. Para abordar este reto, es necesario utilizar herramientas de las que el *clustering* forma parte. El *clustering* es una técnica de aprendizaje no supervisado ampliamente utilizada en diversos campos, como el análisis de imágenes y la bioinformática, que permite realizar un análisis estadístico de los datos [51]. Consiste en la partición de un conjunto de datos diferentes pero con ciertas similitudes en subconjuntos homogéneos o *clusters*, de manera que los datos en cada subconjunto sean similares según alguna medida de distancia definida. Un *cluster* es, por lo tanto, una colección de elementos que son similares entre sí y diferentes a los pertenecientes al resto de *clusters* o subconjuntos [51].

Los algoritmos de *clustering* pueden ser jerárquicos o particionales. Los algoritmos jerárquicos encuentran *clusters* sucesivos utilizando *clusters* previamente establecidos mientras que los algoritmos particionales determinan todos los *clusters* a la vez [52]. Los algoritmos jerárquicos se dividen en dos tipos: aglomerativos y divisivos. Los algoritmos jerárquicos aglomerativos comienzan con tantos *clusters* como datos disponibles y los van fusionando en función de las distancias y similitudes entre ellos. Por otro lado, los algoritmos jerárquicos divisivos comienzan con un único *cluster* y lo van dividiendo en subgrupos más pequeños en función de métricas y distancias [52].

Un paso importante en el agrupamiento jerárquico utilizado en este TFG, es la selección de una medida de distancia [53]. En una revisión del análisis de *clusters* en la investigación de psicología de la salud, se encontró que la medida de distancia más comúnmente utilizada en los estudios publicados en esa área de investigación es la distancia euclídea [54].

- **Distancia de Manhattan.** Calcula la distancia que se recorrería para llegar de un punto de datos al otro siguiendo un camino cuadricular. La distancia de Manhattan entre dos elementos es la suma de las diferencias de sus componentes correspondientes. Si por ejemplo se quiere calcular la distancia entre un punto $X = (X_1, X_2, ..., X_i)$ y otro punto $Y = (Y_1, Y_2, ..., Y_i)$ se tiene que aplicar la siguiente ecuación:

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

  donde n es el número de variables y X_i e Y_i son los valores de la i-ésima variable en los puntos X e Y respectivamente.

- **Distancia euclídea.** Medir la distancia euclídea entre dos puntos de datos implica calcular la raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores correspondientes. La fórmula que se debe aplicar para calcular la distancia euclidiana entre un punto $X = (X_1, X_2, ..., X_i)$ y otro punto $Y = (Y_1, Y_2, ..., Y_i)$ es la siguiente:

$$d = \sqrt{\sum_{j=1}^{n} (X_j - Y_j)^2}$$

Si se hace una comparación entre ambas distancias, se puede afirmar que la distancia de Manhattan es una medida más restrictiva, pues solo permite el movimiento en ángulos rectos mientras que la euclídea permite el movimiento en todas las direcciones del plano [53].

Uno de los desafíos principales para el agrupamiento particional es estimar el número de grupos. Algunos de los métodos utilizados para abordar este problema son [55]:

- **Método del codo.** Consiste en calcular y graficar la suma de errores cuadráticos dentro de cada número de *clusters*. Se busca un punto en la gráfica donde haya un cambio de pendiente pronunciado a una pendiente más suave, este punto se asocia con el número óptimo de *clusters* para esos datos [56].

- **Método de la silueta.** Es una técnica utilizada para evaluar la calidad del agrupamiento y determinar el número óptimo de *clusters* en un conjunto de datos [57]. Se basa en la medida de la distancia de separación entre los grupos resultantes. Para su aplicación, se calcula un coeficiente llamado de silueta (en inglés, *silhouette coefficient*) para cada punto en función de la distancia promedio dentro del mismo *cluster* y la distancia promedio al *cluster* vecino más cercano. Este coeficiente varía de -1 a 1, donde un valor de 1 indica que los puntos están bien separados dentro de su propio *cluster* y un valor de -1 indica que los puntos están más cerca de los *clusters* vecinos que del suyo propio [57].

- **Índice de Calinski-Harabasz.** Es una métrica utilizada en la evaluación de algoritmos de agrupamiento. Es especialmente útil para evaluar la calidad de la partición realizada por un algoritmo. Se calcula como la razón entre la suma de la dispersión entre subconjuntos y la suma de la dispersión dentro de los subconjuntos. Un valor alto de este índice indica un mejor *clustering*, ya que las observaciones dentro de cada subconjunto estarían más cerca entre sí y los *clusters* más separados unos de otros [58].

- **Índice Dunn.** Se calcula como la razón entre la menor distancia *inter-cluster* y la mayor distancia *intra-cluster*. Un valor alto de este índice indica un mejor agrupamiento, ya que

las observaciones dentro de cada subconjunto identificado estarían más cerca entre sí, mientras que los suconjuntos estarían más separados unos de otros [59].

- **Índice Davies-Bouldin.** Se calcula como la similitud promedio de cada subconjunto con el subconjunto más similar a él. Cuanto menor sea el este índice, menor será la similud promedio, lo que inidca que los *clusters* están mejor estarán separados. Por lo tanto, un valor más bajo de este índice implica un mejor resultado del agrupamiento [60].

En este TFG se le da especial importancia al método de la silueta debido a su capacidad para evaluar la calidad de la partición en diferentes números de *clusters* en un conjunto de datos determinado. Sin embargo, además del método de la silueta, se emplean los tres últimos métodos mencionados anteriormente (Calinski, Dunn y Davies) para validar el número óptimo de subconjuntos. Estos métodos proporcionan una evaluación adicional y complementaria al método de la silueta, lo que permite tener una mayor confianza en la elección del número de *clusters* adecuado para la aplicación del algoritmo de *clustering* en los datos analizados.

La toma de decisiones desempeña un papel crítico en medicina. Aplicar técnicas de *clustering* sobre los datos disponibles de los pacientes, hace posible la creación de subconjuntos significativos que permiten la exploración y el análisis de características comunes de forma más eficiente. Esto ayudará a mejorar la toma de decisiones, debido a que en el análisis final se estaría trabajando una serie de datos más estructurados. El uso de *clustering* en medicina tiene una amplia gama de aplicaciones en diversas especialidades. Por ejemplo, se utiliza en la segmentación y análisis de imágenes médicas, en la clasificación de pacientes y en el descubrimiento de patrones [61]. En varios estudios se han aplicado técnicas de *clustering* en el área médica, de los cuales algunos se exponen a continuación.

Por ejemplo en el estudio *A Clustering Approach for Predicting Readmissions in Intensive Medicine* de Rui Veloso et al. [61] se utiliza un conjunto de escenarios de *clustering* para la identificación de pacientes readmitidos y no readmitidos en una unidad de cuidados intesivos. Con los resultados obtenidos, se ha podido caracterizar el perfil clínico de los pacientes que tienen una mayor probabilidad de ser readmitidos. Por otra parte, en el estudio *Segmentación de Imágenes Médicas Digitales mediante Técnicas de Clustering* de Gustavo Lorca et al. [62] se utilizan técnicas de *clustering* en la segmentación de imágenes médicas digitales con el objetivo de ser utilizadas en la reconstrucción de modelos tridimensionales anatómicos. Como resultado, se ha podido desarrollar un módulo de segmentación de imágenes médicas digitales mediante *clustering* demostrando que esta herramienta tiene ventaja sobre las otras técnicas de segmentiación.

Capítulo 3

Base de datos y análisis descriptivo

Este capítulo tiene como finalidad presentar las bases de datos utilizadas en el TFG, así como describir las etapas de preprocesamiento seguidas para posteriormente aplicar las técnicas de ML sobre datos adecuados y de calidad. Además, en este capítulo se incluye un análisis descriptivo de los datos para obtener una comprensión clínica más amplia.

3.1. Descripción de las bases de datos

Para el desarrollo de este trabajo se utilizaron datos clínicos registrados en el estudio *T1D Exchange* realizado por el *Jaeb Center for Health Research*, en Florida, Estados Unidos [63]. Este estudio consta de un conjunto de 16 bases de datos, de las cuales se seleccionaron 5 para este proyecto debido a su relevancia en el análisis utilizando el EFA. Estas bases de datos provienen de encuestas que proporcionan información detallada sobre los miedos, las preocupaciones, el estado emocional y los síntomas experimentados por los pacientes con DMT1 que tienen tendencia a sufrir episodios de hipoglucemia. Además, a través de estas encuestas se evalúan las capacidades cognitivas de estos pacientes y los posibles deterioros asociados causados por la enfermedad. El objetivo principal del estudio previamente mencionado, ha sido evaluar los posibles factores contribuyentes en la aparición de los episodios de hipoglucemia grave y colaborar en la construcción de bases de datos que impulsen la investigacion para optimizar la prevención, diagnóstico y tratamiento de la DMT1 [63].

Los datos recopilados corresponden a 201 personas diferentes, de las cuales 101 son casos y 100 son controles. Los pacientes del grupo de casos han experimentado al menos un episodio grave de hipoglucemia en los últimos 12 meses, mientras que para formar el grupo de los

controles, se han seleccionado personas que no han experimentado ningún episodio de hipoglucemia grave en los últimos 3 años. Todos los participantes tenían 60 años o más y llevaban al menos 20 años con DMT1 [63]. Con el fin de preservar la privacidad de los participantes, los datos recogidos han sido anonimizados y se le ha asignado a cada paciente un ID aleatorio para su identificación adecuada. Cabe mencionar que en todas las bases de datos existe una variable llamada *RecID* que sirve para enumerar a los participantes pero no sirve como identificador único. Se eliminará durante el preprocesamiento de cada una de las bases.

A continuación, se describen las encuestas que han dado lugar a las bases de datos utilizadas en este trabajo, las cuales han sido seleccionadas específicamente debido a que cumplen con los requisitos necesesarios establecidos por el EFA antes de su aplicación en el estudio.

- ***BBGAttitudeScale.*** Encuesta diseñada para medir los niveles de miedo que experimentan los pacientes ante la posibilidad de sufrir episiodios de hipoglucemia. La evaluación se compone de 8 preguntas cuyas posibles respuestas se corresponden con la escala Likert. Esta escala es una de las herramientas más utilizadas en las investigaciones sociales para medir actitudes y percepciones [64].

- ***BGeriDepressScale.*** Encuesta basada en la *Geriatric Depression Scale* que es una escala que se usa para evaluar y detectar la depresión en adultos mayores. Está formada por 15 preguntas que evalúan factores como los diferentes síntomas de la depresión, el estado de ánimo, la falta de energía y la pérdida de interés en actividades cotidianas. El objetivo del presente estudio es evaluar si existe una relación entre estos puntos y la DMT1 [65].

- ***BHypoFearSurvey.*** Encuesta formada por 23 preguntas que evalúa los miedos y preocupaciones que experimentan los pacientes frente al riesgo de sufrir episodios de hipoglucemia. Contiene información sobre la *Hypoglycemia Fear Survey* que es una encuesta que fue desarrollada inicialmente para medir comportamientos y preocupaciones relacionadas con el miedo a la hipoglucemia en adultos con DMT1 [66].

- ***BHypoUnawareSurvey.*** Cuestionario formado por 8 preguntas para evaluar la falta de conciencia y conocimiento en lo que respecta a la hipoglucemia en pacientes con DMT1 [67]. El objetivo principal es determinar su capacidad para reconocer los síntomas de los episodios hipoglucémicos, especialmente cuando son menos evidentes.

- ***BMoCA.*** Encuesta basada en el *Montreal Cognitive Assessment* que es una técnica de detección rápida del deterioro cognitivo. Se trata de un test de 30 puntos que se les proporciona a los pacientes en una sola página y el tiempo estimado de realización son 10

minutos [68]. El objetivo de la utilización de esta base de datos en este trabajo es relacionar la DMT1 con el deterioro de las capacidades cognitivas.

3.2. Análisis descriptivo y preprocesamiento

En el ámbito clínico se trabaja con grandes cantidades de información que se recogen en bases de datos. Estos datos a menudo se utilizan para optimizar procesos y recursos que beneficien tanto a los pacientes como a los sistemas sanitarios. En el caso de este TFG, se han utilizado datos previamente recogidos para agrupar una serie de variables en factores con el objetivo de reducir la dimensionalidad y analizar las correlaciones que existen entre estas y sus factores asignados. Además, se han agrupado los pacientes en subgrupos con características similares.

La calidad del conocimiento extraído a través de la interpretación de resultados obtenidos a partir de los experimentos, depende en gran medida de los datos iniciales que muchas veces pueden estar afectados por valores perdidos, inconsistencias y ruido [35]. Numerosos estudios han demostrado que una baja calidad de los datos conduce a una baja calidad de resultados debido a que los modelos y análisis realizados se ven afectados por la falta de precisión y confiabilidad de los datos [35]. Es por ello por lo que un previo análisis descriptivo de los datos y su preprocesamiento son dos pasos fundamentales antes de usar modelos de ML porque además de que permiten obtener un conocimiento profundo de los datos, también sirven para realizar la limpieza que corresponde de los mismos. En la parte correspondiente al preprocesamiento de datos de este TFG se ha seguido el mismo esquema para todas las bases de datos e incluye los pasos que se exponen a continuación:

1. Transformación de las variables categóricas en numéricas utilizando la técnica que corresponde para cada base de datos. Se empleó la codificación basada en la escala Likert, la codificación binaria y la codificación manual (descrita en la Tabla 3.5), dependiendo de las posibles respuestas de cada encuesta. Es importante destacar que la única base de datos que no requirió este proceso de transformación es la *BMoCA*.

2. Asegurar que no hay pacientes repetidos. Partiendo de las bases de datos anonimizadas, en este paso se ha verificado que no hubiera pacientes duplicados.

3. Eliminar las variables que forman parte de las bases de datos pero que no se corresponden con las preguntas a las que han sido sometidos los pacientes. Generalmente son datos clínicos rellenados por parte de los profesionales.

4. Analizar los pacientes cuyas respuestas contengan valores nulos (NaN) en más del 50 % de las preguntas para proceder con su eliminación (si corresponde). En caso de que el número de NaN no supere la mitad del número de preguntas, se usa una codificación en la que se sustituyen los NaN por el valor que más se repite o moda de cada variable. La moda, al ser una medida de la tendencia central que representa el valor más frecuente en un conjunto de datos, se usa cuando se busca preservar la distribución existente y evitar la introducción de sesgos o distorsiones en los datos.

5. Representar histogramas de datos para visualizar la distribución de las respuestas que han proporcionado los pacientes para cada variable, diferenciando entre los casos y los controles. Esto permitirá una mejor comprensión y análisis de los datos.

Antes de empezar a exponer el análisis y el preprocesamiento de cada una de las bases de datos utilizadas en este trabajo, cabe destacar que hay dos variables previamente mencionadas que se repiten en todas ellas. La primera es el identificador único de cada paciente, denominado *PtID* y la otra es la variable *RecID*. La variable *PtID* tendrá utilidad en aquellos casos en los que se requiera la concatenación de información perteneciente a varias bases de datos. No obstante, para la aplicación del EFA y el *clustering*, esta variable no se va a tener en cuenta, por lo que se eliminará después de comprobar que no está repetida en ninguna base de datos.

3.2.1. Base de datos *BBGAttitudeScale*

Esta base de datos está inicialmente formada por 10 variables de las cuales 6 son categóricas y 4 numéricas. En el preprocesamiento, se ha empezado por la sustitución de las variables categóricas por otras numéricas, proceso en el que se ha hecho uso de la escala Likert cuya forma de codificación se define en la Tabla 3.1.

Strongly Disagree	***Disagree***	***Neutral***	***Agree***	**Strongly agree**
(0)	(1)	(2)	(3)	(4)

Tabla 3.1: Codificación correspondiente a la escala Likert.

El siguiente paso consistió en realizar un análisis que aparece en la Figura 3.1 de valores nulos en todas las variables. Como resultado, se identificó que una de las variables categóricas (*BBGAttitudeScaleNotDone*) presentaba un alto porcentaje de valores nulos, lo que justifica su eliminación como parte del preprocesamiento de los datos.

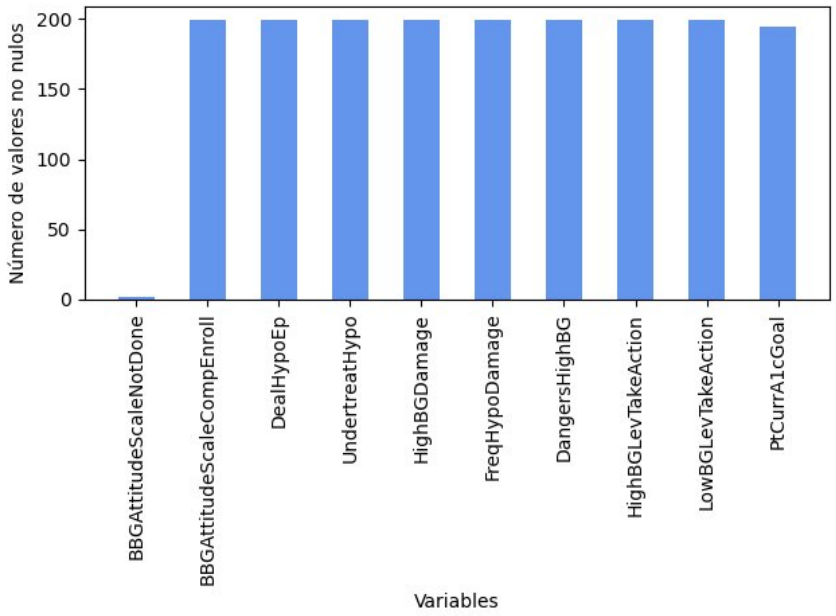


Figura 3.1: Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BBGAttitudeScale*.

Además, se identificaron variables en la base de datos que recopilan datos clínicos y medidas relacionadas con la toma de decisiones objetiva de los pacientes (*HighBGLevTakeAction*, *LowBGLevTakeAction* y *PtCurrA1cGoal*), las cuales no son objetivo del análisis que se pretende abordar en este trabajo, por lo que también han sido eliminadas. Finalmente, se obtuvo una base de datos compuesta por 5 variables que inicialmente eran categóricas pero que han sido transformadas en numéricas utilizando la escala Likert. Estas variables se han registrado como resultado de la encuesta descrita en el Tabla 3.2 a la cual han sido sometidos los pacientes para la construcción de esta primera base de datos y son las que se van a utilizar en las herramientas de ML que se van a aplicar.

VARIABLE	DESCRIPCIÓN	RESPUESTAS
DealHypoEp	Preferencia de lidiar con episodios de hipoglucemia antes que con niveles de glucosa en sangre demasiado altos	*Strongly disagree (0)* *Disagree (1)* *Neutral (2)* *Agree (3)* *Strongly agree (4)*
UnderTreatHypo	Preferencia de tratar episodios de hipoglucemia a tomar el riesgo de sufrir unos niveles de glucosa en sangre bastante altos	*Strongly disagree (0)* *Disagree (1)* *Neutral (2)* *Agree (3)* *Strongly agree (4)*
HighBGDamage	Preocupación por que una o dos lecturas que indiquen niveles de glucosa en sangre bastante altos puedan causar daños en el organismo	*Strongly disagree (0)* *Disagree (1)* *Neutral (2)* *Agree (3)* *Strongly agree (4)*
FreqHypoDamage	Preocupación por que los episodios de hipoglucemia frecuentes puedan causar daños en el organismo	*Strongly disagree (0)* *Disagree (1)* *Neutral (2)* *Agree (3)* *Strongly agree (4)*
DangersHighBG	Mayor preocupación por los peligros asociados a una glucemia alta que aquellos asociados a una glucemia baja	*Strongly disagree (0)* *Disagree (1)* *Neutral (2)* *Agree (3)* *Strongly agree (4)*

Tabla 3.2: Descripción y opciones de respuesta para las variables de la base de datos *BBGAttitudeScale.*

Una vez seleccionadas las variables con las que se va a trabajar, el siguiente paso es analizar la presencia valores nulos (NaN) en cada una de ellas. Durante el análisis, se han identificado dos pacientes que no han proporcionado respuestas para ninguna pregunta de la encuesta. Esto permite su eliminación del estudio, pues la falta de respuestas puede afectar tanto la integridad y validez de los resultados obtenidos como su posterior interpretación. Por lo tanto, después de realizar el preprocesamiento de la presente base de datos, que originalmente constaba de 10 variables y 201 pacientes, se ha conseguido limpiar los datos y la dimensionalidad se redujo a 199 pacientes y 5 variables que son los que se van a utilizar para aplicar las técnicas de ML.

El histograma de datos que muestra la disposición de cada una de las variables, se presenta en la Figura 3.2 con el objetivo de visualizar y comprender la distribución de las respuestas que han proporcionado los pacientes. En el eje de ordenadas se representan las categorías de

respuestas, mientras que en el eje de abscisas se muestra la frecuencia de cada respuesta. Al realizar un análisis descriptivo, se observa una clara preocupación tanto en los casos como en los controles por las posibles consecuencias de los episodios hipoglucémicos.

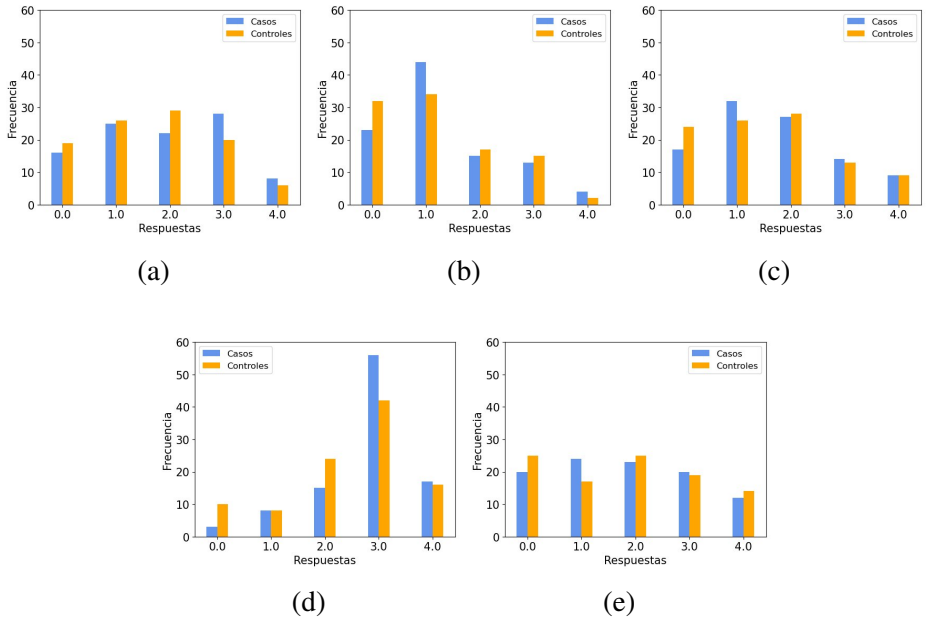


Figura 3.2: Histogramas de las variables de la base de datos *BBGAttitudeScale* diferenciando entre casos y controles. (a) *DealHypoEp* (b) *UndertreatHypo* (c) *HighBGDamage* (d) *FreqHypoDamage* (e) *DangersHighBG*.

3.2.2. Base de datos *BGeriDepressScale*

Base de datos obtenida a partir de un formulario basado en la Escala de Depresión Geriátrica que consta de 16 preguntas para evaluar y detectar la depresión en adultos mayores con DMT1 [69]. Inicialmente se cuenta con 201 pacientes y 17 variables, de las cuales una es numérica y el resto son categóricas. En el primer paso del preprocesamiento, se ha realizado un análisis de los valores nulos que pueden estar presentes en todas las variables, como se muestra en la Figura 3.3. Gracias a este análisis, se identificó una variable categórica (*GeriDepressScaleNotDone*) que presenta el 100 % de sus datos como valores nulos. Estos valores nulos no aportan información relevante para el análisis y podrían distorsionar los resultados. Por lo tanto, la opción más conveniente es eliminar esta variable de la base de datos.

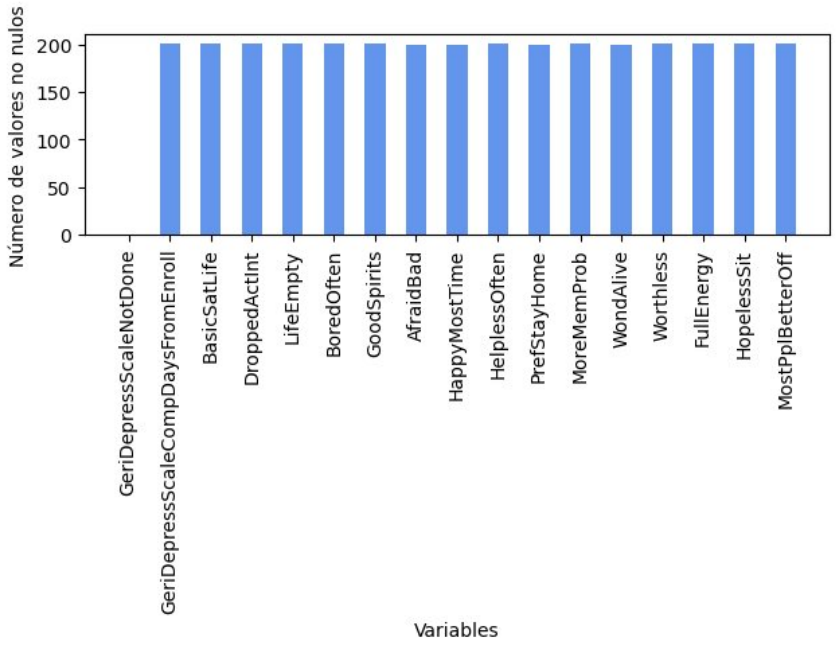


Figura 3.3: Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BGeriDepressScale*.

Por otra parte, también se encontró una variable que registra la fecha de finalización de la encuesta y en la base de datos se incluye como los días transcurridos desde que los pacientes comenzaron a completarla (*GeriDepressScaleCompDaysFromEnroll*). Sin embargo, dado que esta variable no está directamente relacionada con los aspectos que se están abordando en este análisis y no aportará información relevante, se ha considerado que su eliminación de la base de datos es la mejor opción. Después de realizar esta eliminación, se obtiene una base de datos con 15 variables categóricas resultantes de la encuesta que se muestra en el Tabla 3.3.

VARIABLE	DESCRIPCIÓN	RESPUESTAS
BasicSatLife	Satisfacción con su vida	*No (0)* *Yes (1)*
DroppedActInt	Abandono de muchas actividades e intereses	*No (0)* *Yes (1)*
LifeEmpty	Sentimiento de vida vacía	*No (0)* *Yes (1)*
BoredOften	Aburrimiento a menudo	*No (0)* *Yes (1)*
GoodSpirits	Buen humor la mayor parte del tiempo	*No (0)* *Yes (1)*
AfraidBad	Miedo a que algo malo le pase	*No (0)* *Yes (1)*
HappyMostTime	Sentimiento de felicidad la mayor parte del tiempo	*No (0)* *Yes (1)*
HelplessOften	Sentimiento de impotencia a menudo	*No (0)* *Yes (1)*
PrefStayHome	Preferencia de quedarse en casa en vez de salir	*No (0)* *Yes (1)*
MoreMemProb	Más problemas de memoria que el resto	*No (0)* *Yes (1)*
WondAlive	Sentimiento de que es maravilloso vivir	*No (0)* *Yes (1)*
Worthless	Sentimiento de inutilidad	*No (0)* *Yes (1)*
FullEnergy	Sentimiento de mucha energía	*No (0)* *Yes (1)*
HopelessSit	Sentimiento de situación desesperante	*No (0)* *Yes (1)*
MostPplBetterOff	Pensamiento de que el resto de personas se siente mejor	*No (0)* *Yes (1)*

Tabla 3.3: Descripción y opciones de respuesta para las variables de la base de datos *BGeriDepressScale.*

Como se puede ver, las posibles respuestas a cada pregunta son *Yes* o *No* por lo que para convertir las variables categóricas en variables numéricas se ha hecho uso de la codificación binaria, se asigna el valor 1 a las respuestas que son *Yes* y el valor 0 a las respuestas que son *No*. Hecho esto, se procede al análisis de los valores nulos restantes en cada variable de la base de datos. Durante el análisis, se ha identificado un paciente que no ha proporcionado respuestas para ninguna pregunta de la encuesta. Dado que este paciente no aporta ninguna información

relevante y los valores nulos se consideran factores negativos para el análisis de bases de datos, se justifica su eliminación del estudio. Por lo tanto, como resultado del preprocesamiento de esta base de datos se obtienen 200 pacientes y 15 variables cuya distribución se puede ver en los histogramas que se muestran en la Figura 3.4.

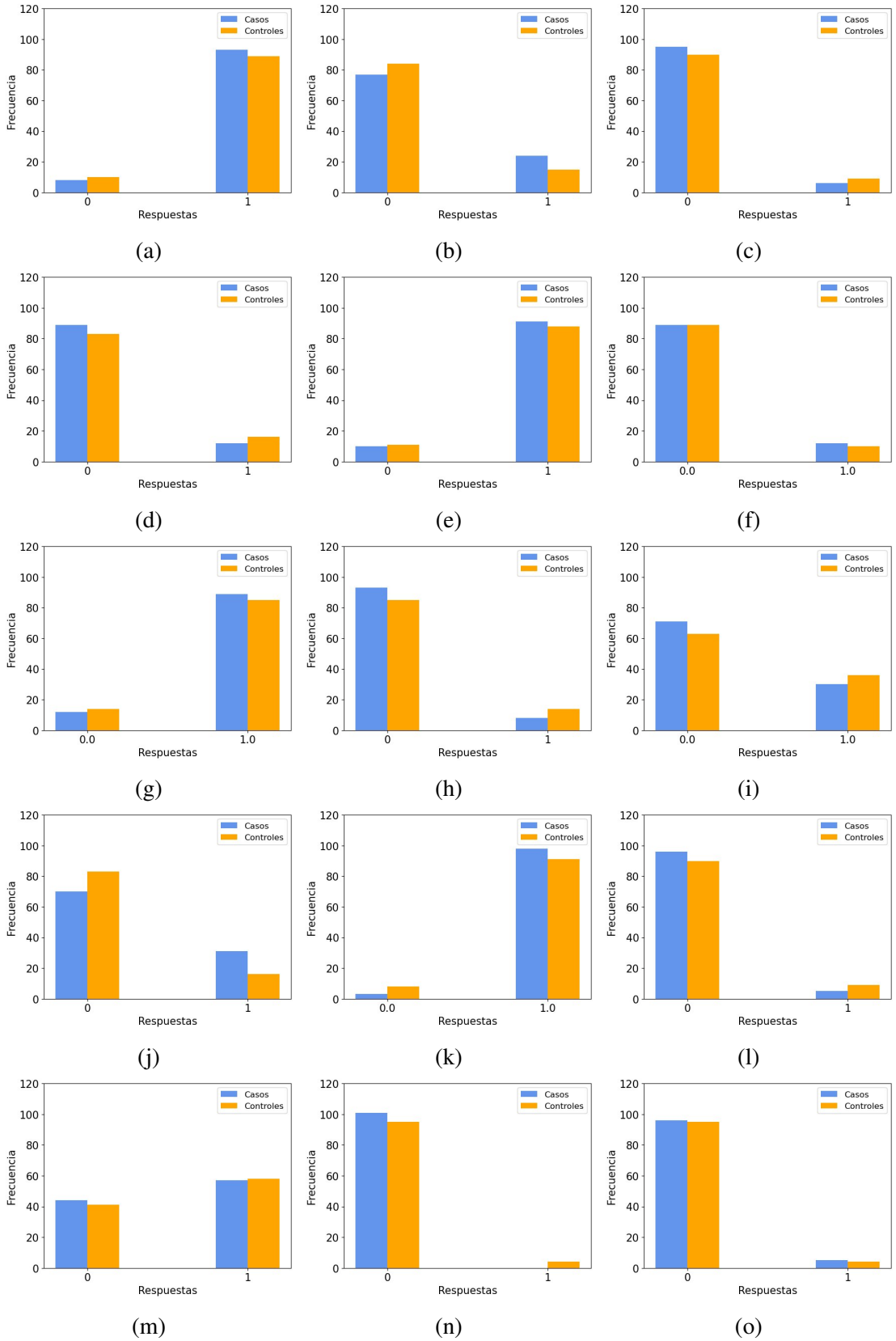


Figura 3.4: Histogramas de las variables de la base de datos *BGeriDepressScale* diferenciando entre casos y controles. (a) *BasicSatLife* (b) *DroppedActInt* (c) *LifeEmpty* (d) *BoredOften* (e) *GoodSpirits* (f) *AfraidBad* (g) *HappyMostTime* (h) *HelplessOften* (i) *PrefStayHome* (j) *MoreMemProb* (k) *WondAlive* (l) *Worthless* (m) *FullEnergy* (n) *HopelessSit* (o) *MostPplBetterOff*.

3.2.3. Base de datos *BHypoFearSurvey*

La base de datos se compone de las variables obtenidas a partir de una encuesta compuesta por 23 preguntas diseñadas para evaluar los miedos y las preocupaciones de los pacientes. En total, consta de 25 variables, donde una es categórica y las restantes son numéricas. Las variables numéricas se encentran en un rango de valores discretos de 0 a 4, los cuales se corresponden con la escala Likert previamente presentada en la Tabla 3.1.

Para el preprocesamiento de los datos, se ha comenzado realizando un análisis de los valores nulos en todas las variables, como se muestra en la Figura 3.5. Gracias a este análisis, se ha identificado que la variable categórica *HypoFearSurvNotDone* contiene un alto porcentaje de valores nulos. Debido a esto, se ha decidido eliminar como parte del preprocesamiento de datos.

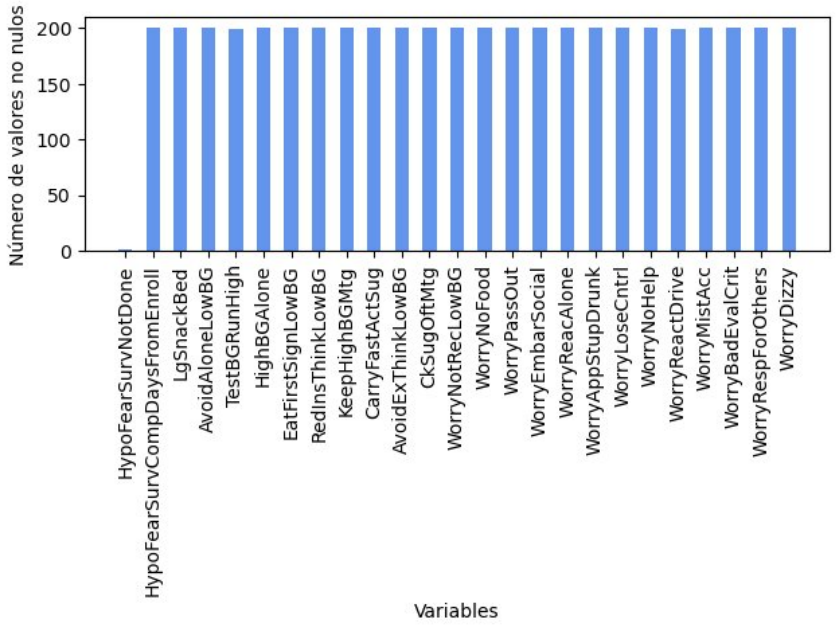


Figura 3.5: Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BHypoFearSurvey*.

Adicionalmente, se ha encontrado una variable (*HypoFearSurvCompDaysFromEnroll*) que registra la fecha de finalización de la encuesta y en la base de datos se incluyen los días transcurridos desde que los pacientes comenzaron a completarla. Sin embargo, dado que esta variable no está directamente relacionada con los aspectos que se están abordando en este análisis y no aportará información relevante, se ha considerado que su eliminación del estudio es la mejor opción. Finalmente, se ha obtenido una base de datos compuesta por 23 variables numéricas, resultado de la encuesta presentada en la Tabla 3.4.

VARIABLE	DESCRIPCIÓN	RESPUESTAS
LgSnackBed	Ingesta de grandes cantidades de comida antes de acostarse	[0-4]
AvoidAloneLowBG	Evitar estar solo cuando los niveles de azúcar en sangre pueden estar bajos	[0-4]
TestBGRunHigh	Preferencia por niveles de azúcar en sangre más altos para mayor seguridad	[0-4]
HighBGAlone	Mantener altos los niveles de azúcar en sangre altos cuando se va a estar un tiempo solo	[0-4]
EatFirstSignLowBG	Ingesta de alimentos apenas se note el primer síntoma de hipoglucemia	[0-4]
RedInsThinkLowBG	Reducir la cantidad de insulina cuando los niveles de azúcar en sangre son bajos	[0-4]
KeepHighBGMtg	Mantener altos los niveles de azúcar en sangre cuando se planea estar en eventos de larga duración	[0-4]
CarryFastActSug	Llevar consigo siempre azúcar de rápido efecto	[0-4]
AvoidExThinkLowBG	Evitar hacer ejercicio cuando se cree que los niveles de azúcar en sangre son bajos	[0-4]
CkSugOftMtg	Comprobación de los niveles de azúcar en sangre cuando se planea asistir a eventos de larga duración	[0-4]
WorryNotRecLowBG	Preocupación por no darse cuenta de que los niveles de azúcar en sangre están bajos	[0-4]
WorryNoFood	Preocupación por tener comida a mano, en caso de que se necesite	[0-4]
WorryPassOut	Preocupación por sufrir un desmayo en público	[0-4]
WorryEmbarSocial	Preocupación por sentir vergüenza o avergonzar a sus amigos en eventos sociales	[0-4]
WorryReacAlone	Preocupación por tener un episodio hipoglucémico estando solo	[0-4]
WorryAppStupDrunk	Preocupación por parecer en estado de ebriedad	[0-4]
WorryLoseCntrl	Preocupación por perder el control	[0-4]
WorryNoHelp	Preocupación por no recibir ayuda durante un episodio de hipoglucemia que lo requiera	[0-4]
WorryReactDrive	Preocupación por sufrir un episodio de hipoglucemia mientras se está conduciendo	[0-4]
WorryMistAcc	Preocupación por cometer errores o tener un accidente	[0-4]
WorryBadEvalCrit	Preocupación por recibir críticas	[0-4]
WorryRespForOthers	Preocupación por la dificultad para pensar con claridad cuando se tiene una responsabilidad	[0-4]
WorryDizzy	Preocupación por la sensación de mareo	[0-4]

Tabla 3.4: Descripción y opciones de respuesta para las variables de la base de datos *BHypoFearSurvey.*

Una vez seleccionadas las variables con las que se llevará a cabo el análisis posterior utilizando técnicas de ML, el siguiente paso consiste en identificar los valores nulos presentes en dichas variables. Durante este proceso, se ha detectado la presencia de tres pacientes que no han respondido la mayoría de las preguntas de la encuesta, lo cual indica que no aportan información relevante para el estudio. Por lo tanto, se ha tomado la decisión de la eliminación de estos tres pacientes del análisis. Después de realizar el preprocesamiento de la presente base de datos, que originalmente constaba de 25 variables y 201 pacientes, se ha conseguido limpiar los datos y la dimensionalidad se redujo a 198 pacientes y 23 variables que son los que se van a utilizar para aplicar las técnicas de ML previamente mencionadas.

Se han incluido los histogramas que se muestran en la Figura 3.6 con el objetivo de visualizar la representación de las respuestas proporcionadas por los pacientes en las variables analizadas. En el eje de ordenadas se presentan las categorías de respuestas, mientras que en el eje de abscisas se muestra la frecuencia de cada respuesta. Analizando la distribución de las respuestas que se muestra en la Figura 3.6, se observa que tanto los casos como los controles adoptan medidas de precaución frente a la posibilidad de experimentar un episodio grave de hipoglucemia. No obstante, se nota una mayor preocupación por parte de los casos en situciones donde estos episodios puedan ocurrir, ya sea en público o cuando están solos. Esto podría ser resultado de haber experimentado estos eventos de manera más reciente que los controles.

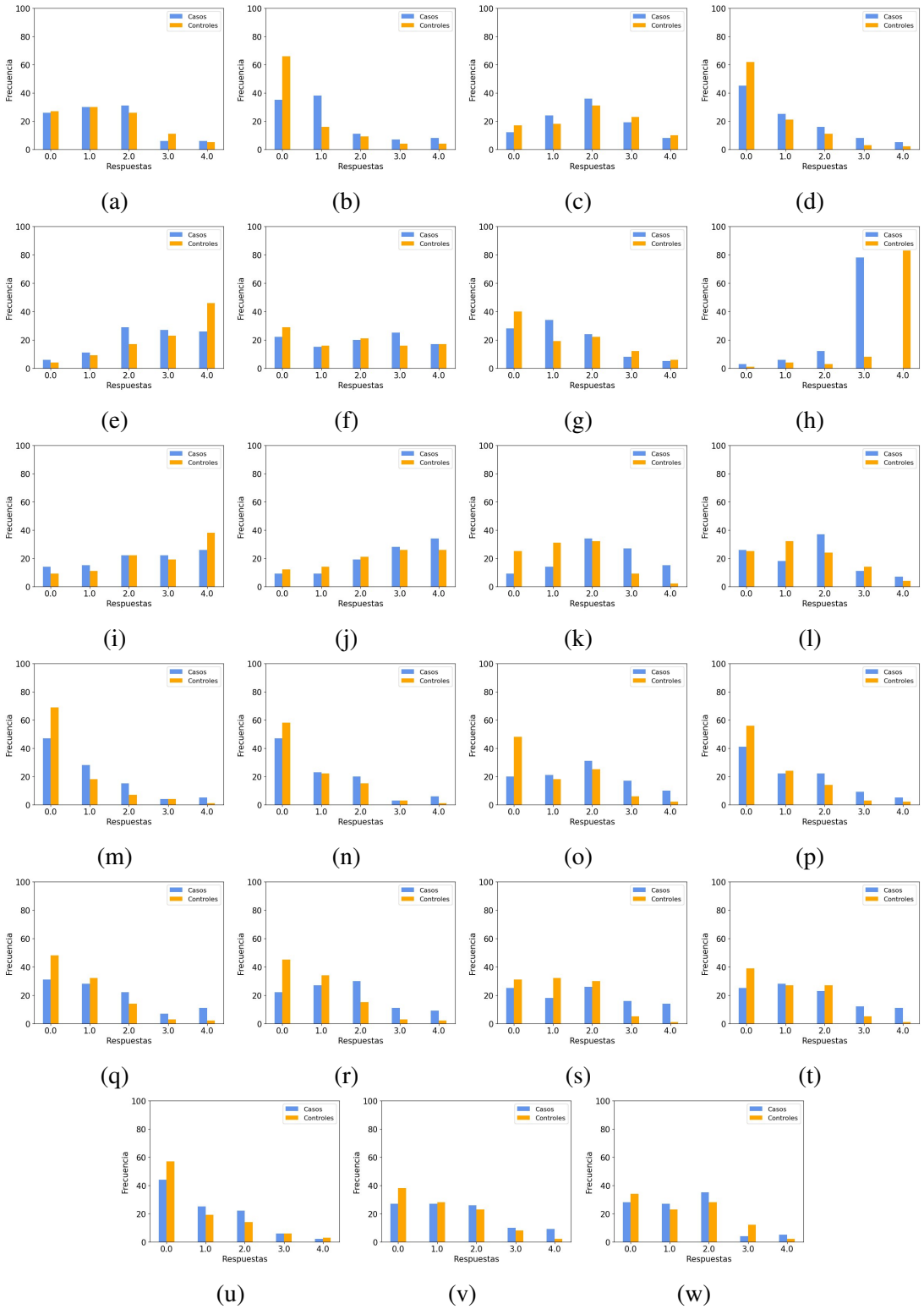


Figura 3.6: Histogramas de las variables de la base de datos *HypoFearSurvCompDaysFromEnroll* diferenciando entre casos y controles. (a) *LgSnackBed* (b) *AvoidAloneLowBG* (c) *TestBGRunHigh* (d) *HighBGAlone* (e) *EatFirstSignLowBG* (f) *RedInsThinkLowBG* (g) *KeepHighBGMtg* (h) *CarryFastActSug* (i) *AvoidExThinkLowBG* (j) *CkSugOftMtg* (k) *WorryNotRecLowBG* (l) *WorryNoFood* (m) *WorryPassOut* (n) *WorryEmbarSocial* (o) *WorryReacAlone* (p) *WorryAppStupDrunk* (q) *WorryLoseCntrl* (r) *WorryNoHelp* (s) *WorryReactDrive* (t) *WorryMistAcc* (u) *WorryBadEvalCrit* (v) *WorryRespForOthers* (v) *WorryDizzy*.

3.2.4. Base de datos *BHypoUnawareSurvey*

La base de datos se compone de variables obtenidas a partir de un cuestionario que consta de 8 preguntas diseñadas para evaluar la falta de conciencia sobre los episodios hipoglucémicos en pacientes con DMT1. Incialmente, la base de datos constaba de 10 variables de las cuales 8 eran categóricas y 2 numéricas. Como primer paso del preprocesamiento, se ha procedido a sustituir las variables categóricas por otras numéricas sin utilizar una codificación predeterminada. En lugar de ello, se ha realizado esta transformación manualmente, variable por variable, aplicando un enfoque específico para cada una de ellas. El número que se le ha asignado a cada posible respuesta se puede ver en la Tabla 3.5.

En el siguiente paso, se ha realizado un análisis de la presencia de los valores nulos en todas las variables, como se muestra en la Figura 3.7. Se ha identificado que una de las variables categóricas (*HypoUnawareSurvNotDone*) presenta el 100 % de sus datos como valores nulos. Dado que esta variable no aporta información relevante para el análisis y podría distorsionar los resultados, se ha decidido eliminarla de la base de datos.

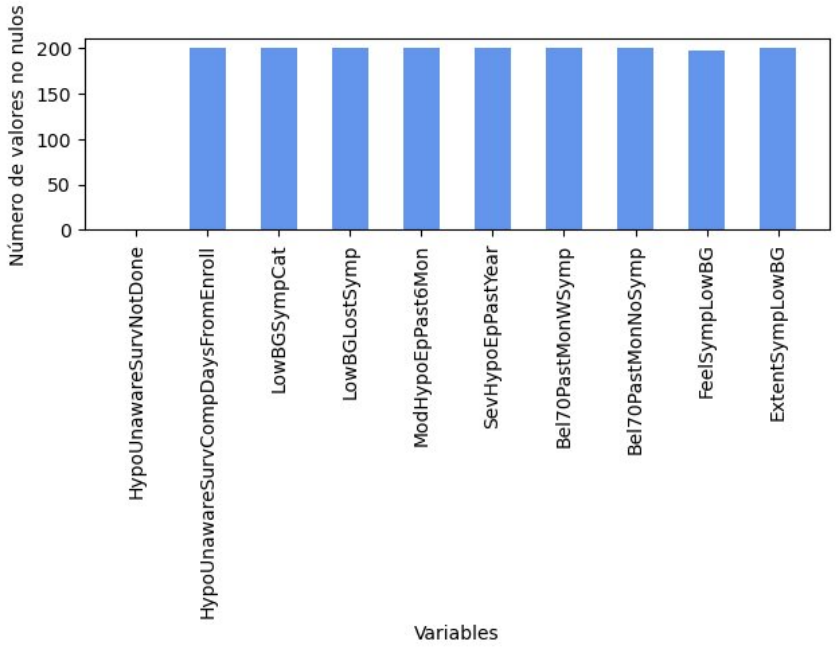


Figura 3.7: Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BHypoUnawareSurvey*.

Por otro lado, también se ha identificado una variable que registra la fecha de finalización de la encuesta (*HypoUnawareSurvCompDaysFromEnroll*), la cual representa los días transcurridos

desde que los pacientes comenzaron a completarla. Sin embargo, dado que esta variable no está directamente relacionada con los aspectos que se están abordando en este análisis no aportará información relevante, se ha considerado que su eliminación de la base de datos es la mejor opción. Después de realizar esta eliminación, se obtiene una base de datos con 8 variables categóricas resultantes de la encuesta que se muestra en la Tabla 3.5.

Variable	Descripción	Respuestas
LowBGSympCat	Seleccionar la mejor categoría que describe al paciente	*When blood sugar is low:* *No longer have symptoms (0)* *Sometimes have symtpoms (1)* *Always have symptoms (2)*
LowBGLostSymp	No darse cuenta de alguno de los síntomas propios de los episodios hipoglucémicos	*Yes (1)* *No (0)*
ModHypoEpPast6Mon	Frecuencia de episodios hipoglucémicos moderados durante los últimos 6 meses	*Never (0) - Once or twice (1) -* *Every other month (2) -* *Once a month (3) -* *More than once a month (4)*
SevHypoEpPastYear	Frecuencia de episodios hipoglucémicos severos durante el último año	*Never (0) - 1 time (1) -* *2 times (2) - 3 times (3) -* *4 times (4) - 5 times (5) -* *6 times (6) - 7 times (7) -* *8 times (8) - 9 times (9) -* *10 times (10) - 11 times (11) -* *12 or more times (12)*
Bel70PastMonWSymp	Frecuencia de lecturas inferiores a 70 mg/dL con síntomas durante el último mes	*Never (0) - 1 to 3 times (1) -* *1 time/week (2) -* *2 to 3 times/week (3) -* *4 to 5 times/week (4) -* *Almost daily (5)*
Bel70PastMonNoSymp	Frecuencia de lecturas inferiores a 70 mg/dL sin síntomas durante el último mes	*Never (0) - 1 to 3 times (1) -* *1 time/week (2) -* *2 to 3 times/week (3) -* *4 to 5 times/week (4) -* *Almost daily (5)*
FeelSympLowBG	Niveles de azúcar en sangre para la experimentación de síntomas	*60-69 mg/dL (0)* *50-59 mg/dL (1)* *40-49 mg/dL (2)* *<40 mg/dL (3)*
ExtentSympLowBG	¿Hasta qué punto se puede saber que los niveles de azúcar en sangre son bajos por los síntomas que se experimentan?	*Never (0) - Rarely (1) -* *Sometimes (2) -* *Often (3) - Always (4)*

Tabla 3.5: Descripción y opciones de respuesta para cada una de las variables de la base de datos *BHypoUnawareSurvey.*

Una vez seleccionadas las variables con las que se va a trabajar, el siguiente paso consiste en realizar un análisis de los valores nulos presentes en dichas variables. Durante este análisis,

se han identificado tres pacientes que tienen NaN en algunas de sus respuestas proporcionadas. Dado que el porcentaje de respuestas nulas es inferior a la mitad del número de preguntas, se ha decidido sustituir estos valores nulos por la moda de la variable a la que corresponden. Como resultado del preprocesamiento de la base de datos inicial, que constaba originalmente de 10 variables y 201 pacientes, se ha obtenido una base de datos procesada que tiene 8 variables y 201 pacientes. Esta es la que se va a utilizar para aplicar las técnicas de ML.

Adicionalmente, se ha realizado una representación de los histogramas de cada una de las variables, como se muestra en la Figura 3.8, con el propósito de facilitar el análisis y la interptretación de la distribución de las respuestas proporcionadas por los pacientes. Al examinar estos resultados, se puede observar que los controles muestran una mayor capacidad para reconocer los síntomas, mientra que los casos tienen una mayor incidencia de estos episodios.

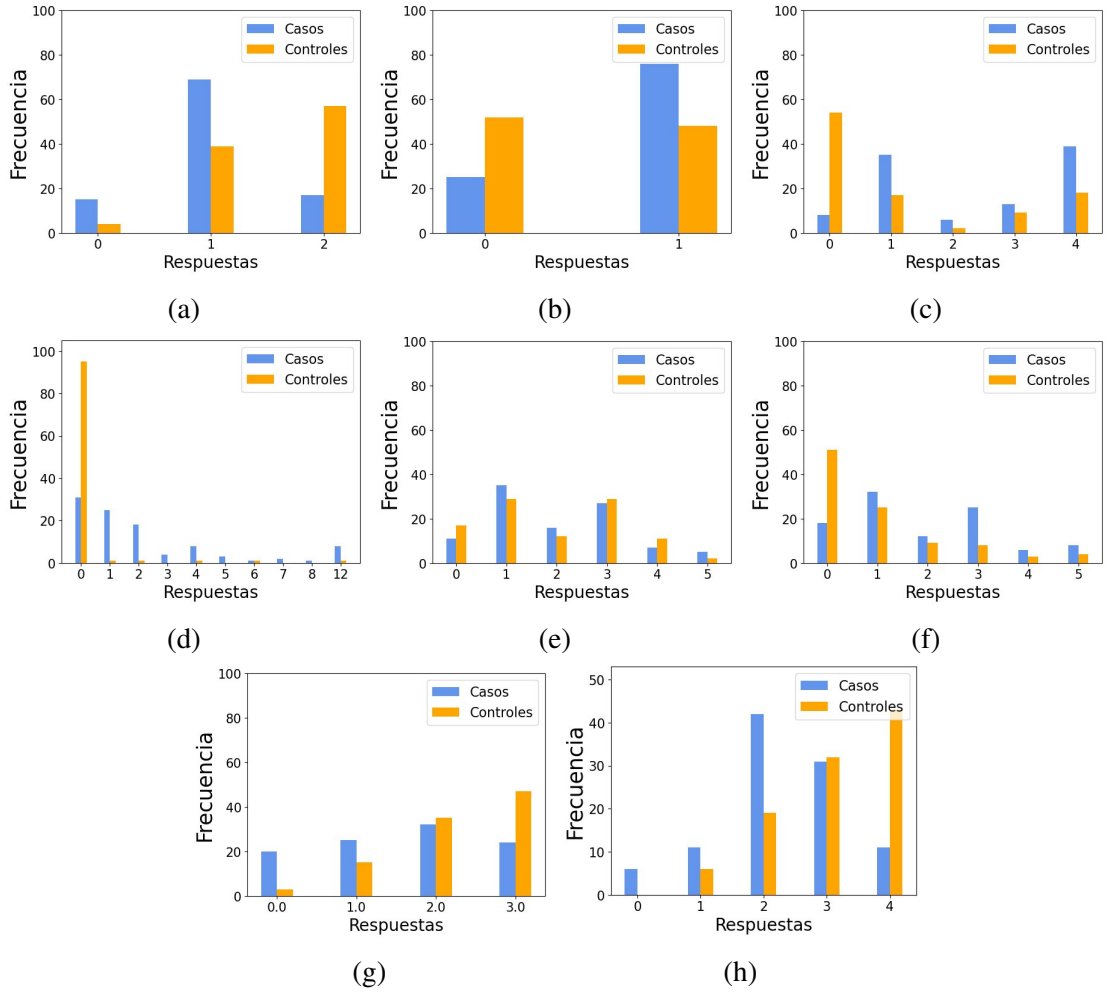


Figura 3.8: Histogramas de las variables de la base de datos *BHypoUnawareSurvey* diferenciando entre casos y controles. (a) *BasicSatLife* (b) *LowBGSympCat* (c) *LowBGLostSymp* (d) *ModHypoEpPast6Mon* (e) *SevHypoEpPastYear* (f) *Bel70PastMonWSymp* (g) *Bel70PastMonNoSymp* (h) *FeelSympLowBG* (i) *ExtentSympLowBG*.

3.2.5. Base de datos *BMoCA*

Esta base de datos está compuesta por variables obtenidas a partir de una encuesta basada en el *Montreal Cognitive Assesment*, que se usa para la detección rápida del deterioro cognitivo. Inicialmente, estaba formada por 201 pacientes y 15 variables, de las cuales 3 eran categóricas y 12 eran numéricas. Para el preprocesamiento de los datos, se ha iniciado realizando una revisión de los valores nulos en todas las variables, como se muestra en la Figura 3.9. A través de este análisis se han identificado dos variables, específicamente *MoCANotDone* y *MoCANotDoneReas*, que tienen la mayoría de sus datos como valores nulos. Dado que estos valores no

aportan información relevante y podrían distorsionar los resultados, se ha decidido eliminar estas variables.

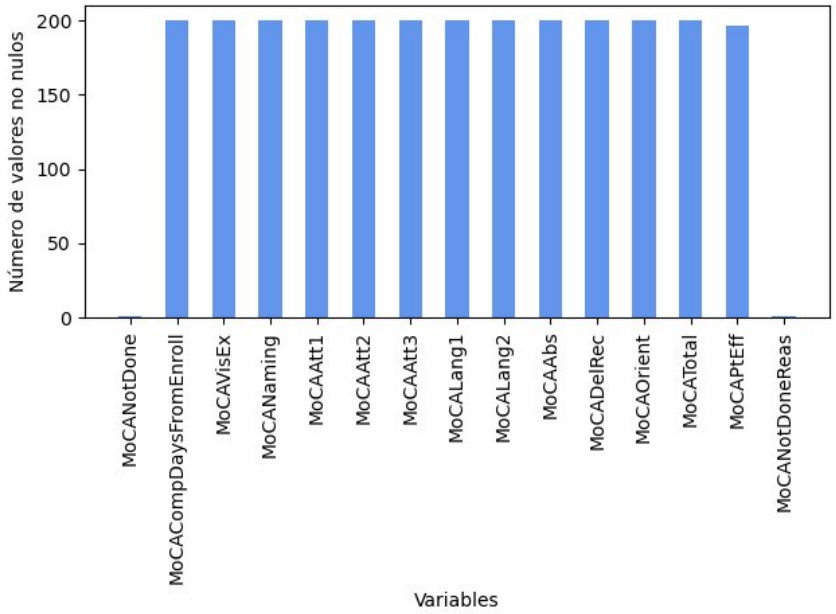


Figura 3.9: Diagrama de barras que muestra el conteo de los valores no nulos por variable en la base de datos *BMoCA*.

Por otra parte, se identificó una variable en la base de datos que registra la fecha de finalización de la encuesta (*MoCACompDaysFromEnroll*), representada como los días transcurridos desde que los pacientes comenzaron a completarla. Sin embargo, dado que esta variable no está directamente relacionada con los aspectos que se están abordando en este análisis y no aportará información relevante, se ha considerado que su eliminación es la mejor opción.

Adicionalmente, se encontraron dos variables (*MoCATotal* y *MoCAPtEff*) que tampoco aportan información relevante para el análisis que se está abordando en este trabajo. La variable *MoCATotal* registra el total de puntos obtenidos por cada paciente en el test, mientras que la variable *MoCAPtEff* representa una evaluación realizada por los clínicos para determinar el compromiso de cada paciente con la prueba. Por lo tanto, también se han eliminado como parte del preprocesamiento de los datos.
En resumen, la variable de fecha de finalización (*MoCACompDaysFromEnroll*), así como las variables *MoCATotal* y *MoCAPtEff* han sido eliminadas de la base de datos durante el preprocesamiento, debido a que no aportan información relevante para el análisis en curso. En este punto, la base de datos procesada consta de 10 variables, que se detallan en la Tabla 3.6.

VARIABLE	DESCRIPCIÓN	RESPUESTAS
MoCAVisEx	Puntuación obtenida en la prueba visoespacial	[0-5]
MoCANaming	Puntuación obtenida en la prueba de nombres	[0-3]
MoCAAtt1	Puntuación obtenida en la prueba de atención I	[0-2]
MoCAAtt2	Puntuación obtenida en la prueba de atención II	[0-1]
MoCAAtt3	Puntuación obtenida en la prueba de atención III	[0-3]
MoCALang1	Puntuación obtenida en la prueba de lenguaje I	[0-2]
MoCALang2	Puntuación obtenida en la prueba de lenguaje II	[0-1]
MoCAAbs	Puntuación obtenida en la prueba de abtracción	[0-2]
MoCADelRec	Puntuación obtenida en la prueba de recuerdo diferido	[0-5]
MoCAOrient	Puntuación obtenida en la prueba de orientación	[0-6]

Tabla 3.6: Descripción e intervalos de respuesta para cada una de las variables de la base de datos *BMoCA*.

Una vez establecidas las variables con las que se va a trabajar, el siguiente paso es analizar los valores nulos (NaN) presentes en las variables seleccionadas. Durante el análisis, se ha identificado un paciente del que no se ha registrado ninguna puntuación. Dado que este paciente no aporta información relevante y los valores nulos son considerados factores negativos en el análisis de las bases de datos, se justifica su eliminación del estudio.

Como resultado del preprocesamiento de la base de datos inicial, que constaba originalmente de 17 variables y 201 pacientes, se ha obtenido una base de datos procesada que tiene 10 variables y 200 pacientes. Esta es la que se va a utilizar para aplicar técnicas de ML y realizar las interpretaciones necesarias.

Con el objetivo de facilitar la interpretación de las puntuaciones obtenidas por los pacientes en cada prueba correspondiente al *Montreal Cognitive Assesment*, se han representado los histogramas de las variables separando entre casos y controles, tal como se muestra en la Figura 3.10. Al analizar las gráficas, se observa que las puntuaciones están bastante equilibradas, aunque se puede apreciar una ligera diferencia a favor de los controles.

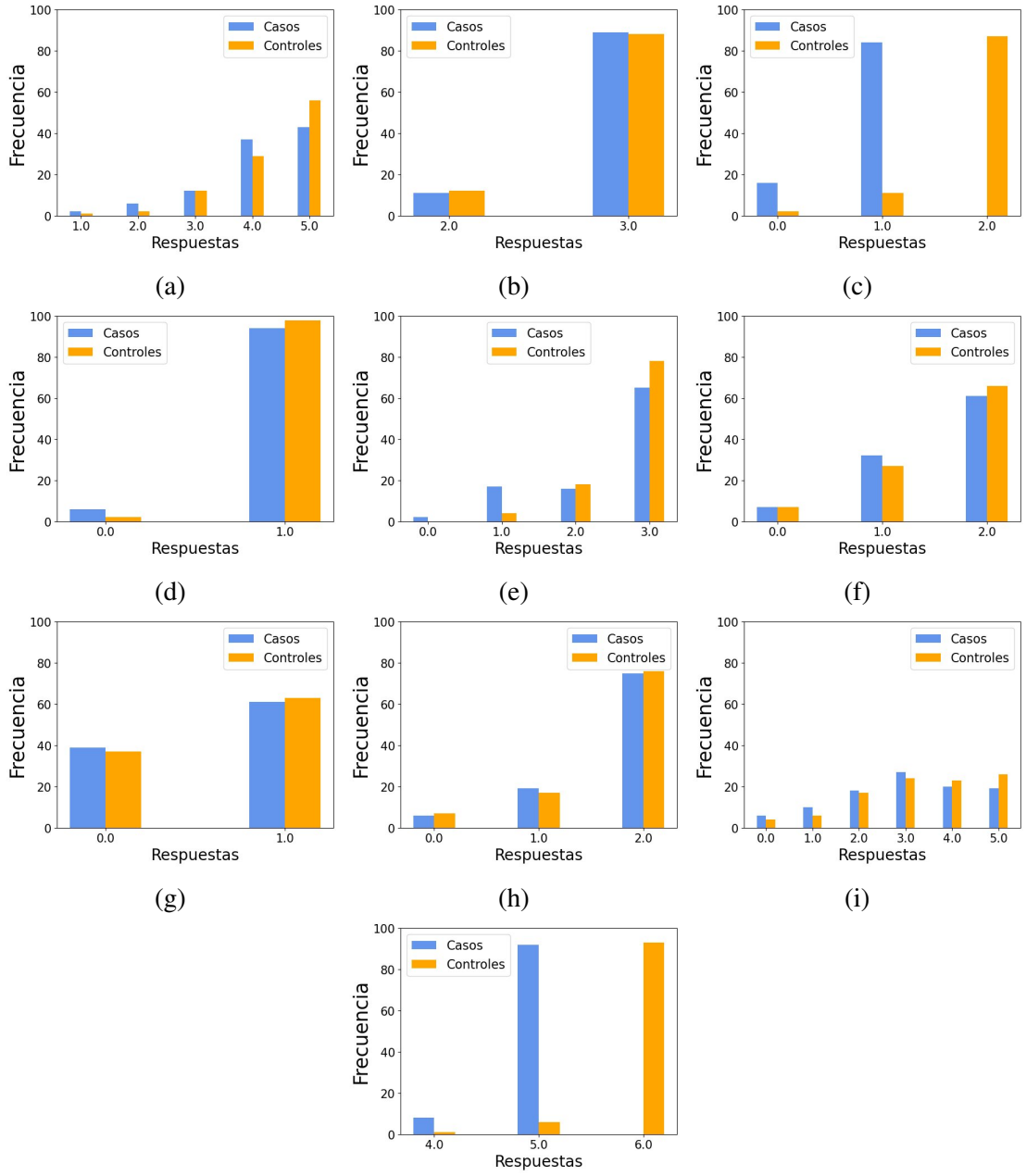


Figura 3.10: Histogramas de las variables de la base de datos *BHypoUnawareSurvey* diferenciando entre casos y controles. (a) *MoCAVisEx* (b) *MoCANaming* (c) *MoCAAtt1* (d) *MoCAAtt2* (e) *MoCAAtt3* (f) *MoCALang1* (g) *MoCALang2* (h) *MoCAAbs* (i) *MoCA-DelRec* (j) *MoCAOrient*.

Capítulo 4

Experimentos y resultados

En este capítulo se presentan los experimentos realizados y los resultados obtenidos a partir del conjunto de datos y los métodos descritos a lo largo del TFG. El objetivo principal ha sido identificar factores que relacionan variables asociadas a los episodios de hipoglucemia severa, así como agrupar a los pacientes en subconjuntos con características similares de manera automática. Estos hallazgos contribuyen a la investigación clínica sobre episodios de hipoglucemia.

4.1. Configuración experimental

El objetivo principal de este trabajo es realizar el agrupamiento de pacientes utilizando variables obtenidas de encuestas relacionadas con los episodios de hipoglucemia. Además, se busca identificar los factores que se relacionen con estas variables. Las bases de datos utilizadas pertenecen al estudio *T1D Exchange*, siendo destacable la base de datos *BPtRoster* que incluye etiquetas para diferenciar entre casos y controles a los pacientes que participaron en las encuestas. Estas etiquetas, utilizadas previamente en los histogramas del Capítulo 3, se vuelven a emplear en la parte experimental de este trabajo, concretamente en la parte de *clustering*.

Después de realizar el respectivo preprocesamiento en cada base de datos, se aplicó el EFA en cada una de ellas. Los hallazgos obtenidos de este análisis se utilizaron como entrada para la segunda herramienta de ML mencionada, *clustering*. Una vez obtenidos los resultados finales de esta primera parte y realizada su interpretación, se concatenaron todas las bases de datos utilizando el identificador correspondiente a cada paciente. Esto condujo a la creación de una base de datos final que incluye todas las variables previamente analizadas. Posteriormente, se aplicó el EFA y *clustering* en esta base de datos final.

Al acabar todo el proceso y previo a realizar la interpretación final de los resultados, se utilizó la base de datos BPtRoster para visualizar el porcentaje de casos y controles asignados a cada uno de los *clusters* obtenidos. Esta evaluación proporcionó una visión general de la distribución de los casos y controles en los diferentes subconjuntos identificados en cada base de datos. Posteriormente, se llevó a cabo la interpretacoón final de los resultados, extrayendo conclusiones relevantes. Esto permitió obtener conocimientos sobre los factores relacionados con los episodios de hipoglucemia y perfiles distintivos para cada subconjunto de pacientes, utilizando técnicas de aprendizaje automático y los datos obtenidos en las encuestas.
Con el fin de esquematizar el proceso general seguido para la parte experimental del presente trabajo, se ha implementado la Figura 4.1.

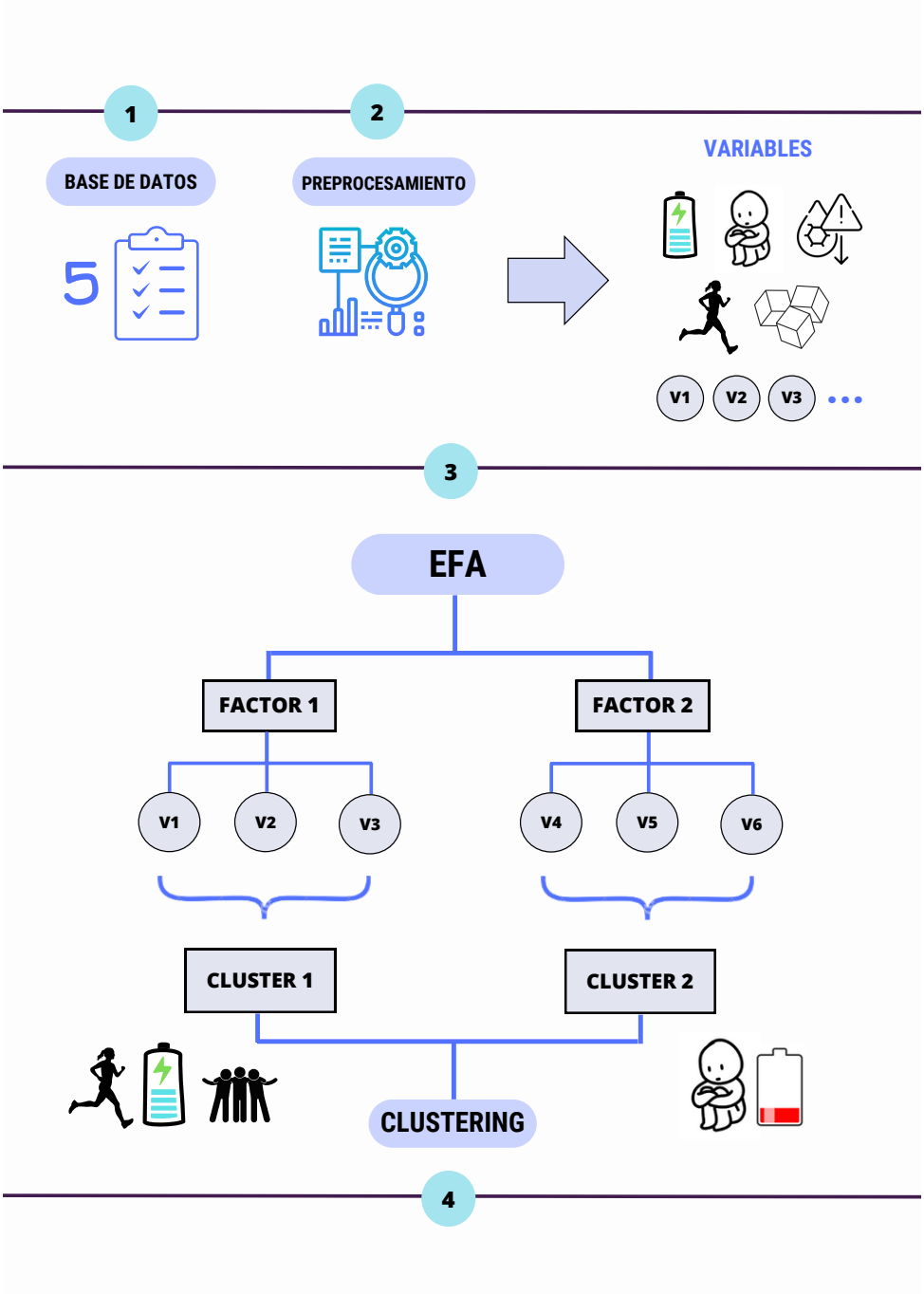


Figura 4.1: Esquema de la parte experimental del estudio.

Para la realización de este trabajo se ha utilizado como lenguaje de programación prinicipal *Python 3*, aprovechando diversas bibliotecas y métodos de ML [70]. Además, se ha utilizado R como lenguaje de programación complementario para la representación de gráficas específicas. El uso de R ha sido especialmente útil para verificar la consistencia de los resultados obtenidos

en Python. Gracias a esta comparación, se ha validado la confiabilidad de los análisis realizados, demostrando coherencia entre los resultados obtenidos en ambos lenguajes de programación.

4.2. Análisis factor exploratorio y *clustering* usando *BBG-AttitudeScale*

Partiendo de la base de datos obtenida tras el preprocesamiento descrito en el capítulo anterior, se han llevado a cabo las dos técnicas de análisis previamente expuestas: EFA y *clustering*.

El EFA se ha utilizado con el propósito de identificar patrones ocultos en el conjunto inicial de las variables seleccionadas. Se aplica con el fin de extraer factores latentes que expliquen la estructura subyacente de los datos. En el contexto de esta primera base de datos, se empleó el EFA para identificar las relaciones entre variables que reflejan las preocupaciones de los pacientes respecto a los episodios de hipoglucemia y los niveles de glucosa en sangre con los que se sienten más seguros. Esto permite tener una mejor comprensión de los factores que influyen en estos episodios y cómo se relacionan entre sí.

El primer paso previo a la aplicación del EFA, ha sido realizar la prueba de esfericidad de Bartlett para evaluar si las correlaciones entre variables son adecuadas para esta técnica explicada en la sección 2.3. Esta prueba se basa en la hipótesis nula de que las variables no están correlacionadas en la muestra y propociona dos valores: el chi-cuadrado y el p_{valor}. En general y como ya se mencionó, un valor alto de chi-cuadrado y un p_{valor} cercano a 0 sugieren la existencia de una correlación significativa entre las variables, lo que justifica el uso del EFA. En este caso particular, se obtuvo un chi-cuadrado de 156.20 y un p_{valor} inferior a 0.05. Estos resultados indican que las variables presentan correlaciones significativas y que el EFA es apropiado para explorar la estructura subyacente de los datos con los que se está trabajando.

Además, antes de aplicar el EFA, se calcula el índice KMO para evaluar la idoneidad de los datos para esta herramienta. Un valor de KMO cercano a 1 indica que los datos son adecuados para el EFA, mientras que un valor más bajo sugiere que los datos pueden no ser apropiados (véase sección 2.3). En este caso, se ha obtenido un valor de KMO de 0.641, lo cual indica que los datos presentan una adecuada idoneidad para llevar a cabo el EFA.

Una vez comprobado y justificado el uso del EFA, se lleva a cabo la determinación del número de factores. Para ello, se utilizan los autovalores (*eigenvalues*) que representan la varianza explicada por cada factor en el análisis. Estos autovalores se utilizan para determinar el número de factores mediante diferentes criterios. En este caso particular, se ha utilizado el método del

gráfico *scree plot* que muestra los autovalores en orden descendente y se identifica el último punto en el que los autovalores son superiores a 1. En la Figura 4.2 (a) se puede observar que el número ideal de factores es 2. Después de determinar el número óptimo de factores, se procede a utilizarlo junto con la rotación adecuada para ajustar el modelo del EFA. En este caso, se ha utilizado un número de factores igual a 2 y la rotación *promax*. Este enfoque ha permitido obtener los resultados que se muestran en las Figura 4.2 (b) y 4.2 (c). De esta forma, las variables quedan agrupadas en dos factores y se observa la ponderación de cada una de ellas con respecto a su factor correspondiente en la Figura 4.2 (c). Estos valores indican la importancia relativa de cada variable en la formación de los factores. Cuanto mayor sea el peso de una variable en un factor, mayor será su contribución en la explicación de la variabilidad asociada a ese factor.

- **Factor 1.** Agrupa aquellas variables que están relacionadas con la sensación de mayor seguridad de los pacientes al experimentar hipoglucemia en comparación con los posibles daños asociados a la hiperglucemia. Estas variables son *DealHypoEp*, *UndertreatHypo* y *DangersHighBG*.
- **Factor 2.** Incluye las dos variables que evalúan las preocupaciones de los pacientes sobre los efectos dañinos de los niveles desequilibrados de azúcar en sangre. Estas variables son *HighBGDamage* y *FreqHypoDamage*.

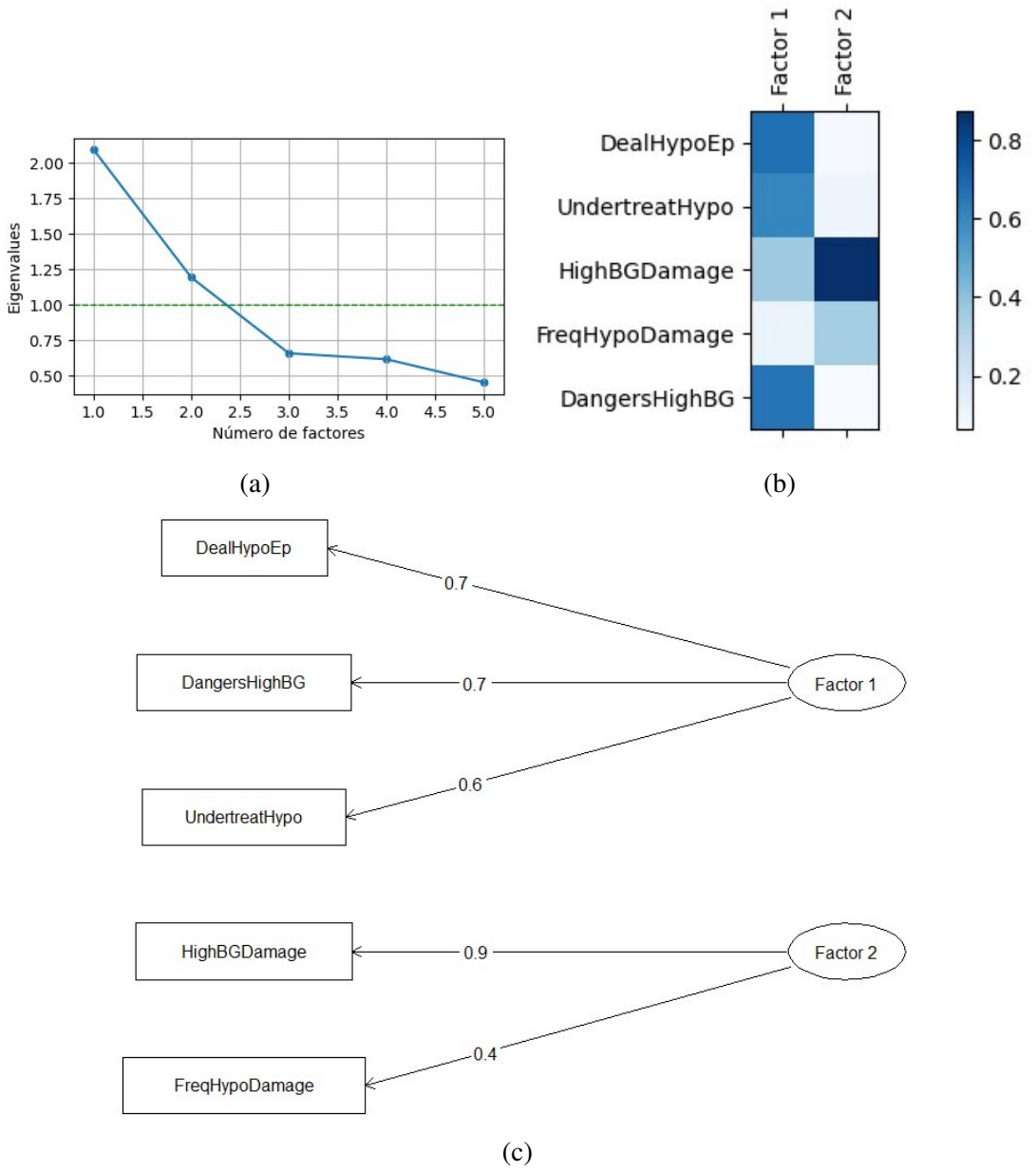


Figura 4.2: Resultados del EFA usando *BBGAttitudeScale*. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Una vez analizados los resultados del EFA, el siguiente paso es aplicar el algoritmo de *clustering* sobre los resultados obtenidos de este primer análisis. Antes de aplicar el algoritmo de *clustering*, es necesario determinar el número óptimo de *clusters* en los que se agruparán los datos. Para ello, y como ya se comentó en la sección 2.4, se han utilizado métodos jerárquicos y se ha dado especial énfasis al método de la silueta. Este método maximiza la coherencia interna

y la separación entre subconjuntos en el análisis de *clustering*. El coeficiente de la silueta se usa como métrica para evaluar la calidad de la asignación de elementos (en este caso, pacientes) a los *clusters*. Un valor alto de este coeficiente indica una asignación precisa de los pacientes a los *clusters* a los que han sido asignados, lo que implica que están agrupados de manera adecuada según sus características. Por lo tanto, se busca maximizar este valor. Para el conjunto de datos con los que se está trabajando, el número adecuado de *clusters* es 2 como se puede ver en las Figuras 4.3 (a) y 4.3 (b). En estas figuras se muestra el número óptimo de *clusters* determinado por diferentes métricas previamente definidas en la sección 2.4. Para seleccionar un número, se ha optado por aquel que se repite con mayor frecuencia.

Una vez establecido el número óptimo de *clusters*, se aplica el algoritmo de agrupamiento y se obtiene la distribución de los pacientes en cada subconjunto, como se muestra en la Figura 4.3 (c). Esta representación visual proporciona información sobre cómo se agrupan y se relacionan entre sí los pacientes en función de las características consideradas en el análisis. Por otra parte, se ha implementado la Figura 4.3 (d) para ver el porcentaje de casos y controles que hay dentro de cada *cluster*, lo cual puede resultar de gran ayuda en el análisis y comparación de la composición de los *clusters* en términos de casos y controles.

En el caso particular de esta base de datos, se han separado a los pacientes en dos *clusters*. En el *cluster* 1, se agrupa el 77.4 % de los pacientes mientras que en el *cluster* 2 se encuentra el 22.6 % de los pacientes. Dentro del primer *cluster*, el porcentaje de casos es del 49.35 %, mientras que el porcentaje de controles es del 50.65 % . En cuanto al segundo cluster, se observa que el 51.11 % de sus pacientes son casos y el 48.89 % restante son controles.

Estos resultados indican que la distribución de los pacientes en los dos subconjuntos es homogénea. Cabe mencionar que el algoritmo de agrupamiento ha logrado identificar estos subconjuntos sin tener acceso previo a las etiquetas, lo que sugiere que ha identificado patrones basados en las características o perfiles de los pacientes.

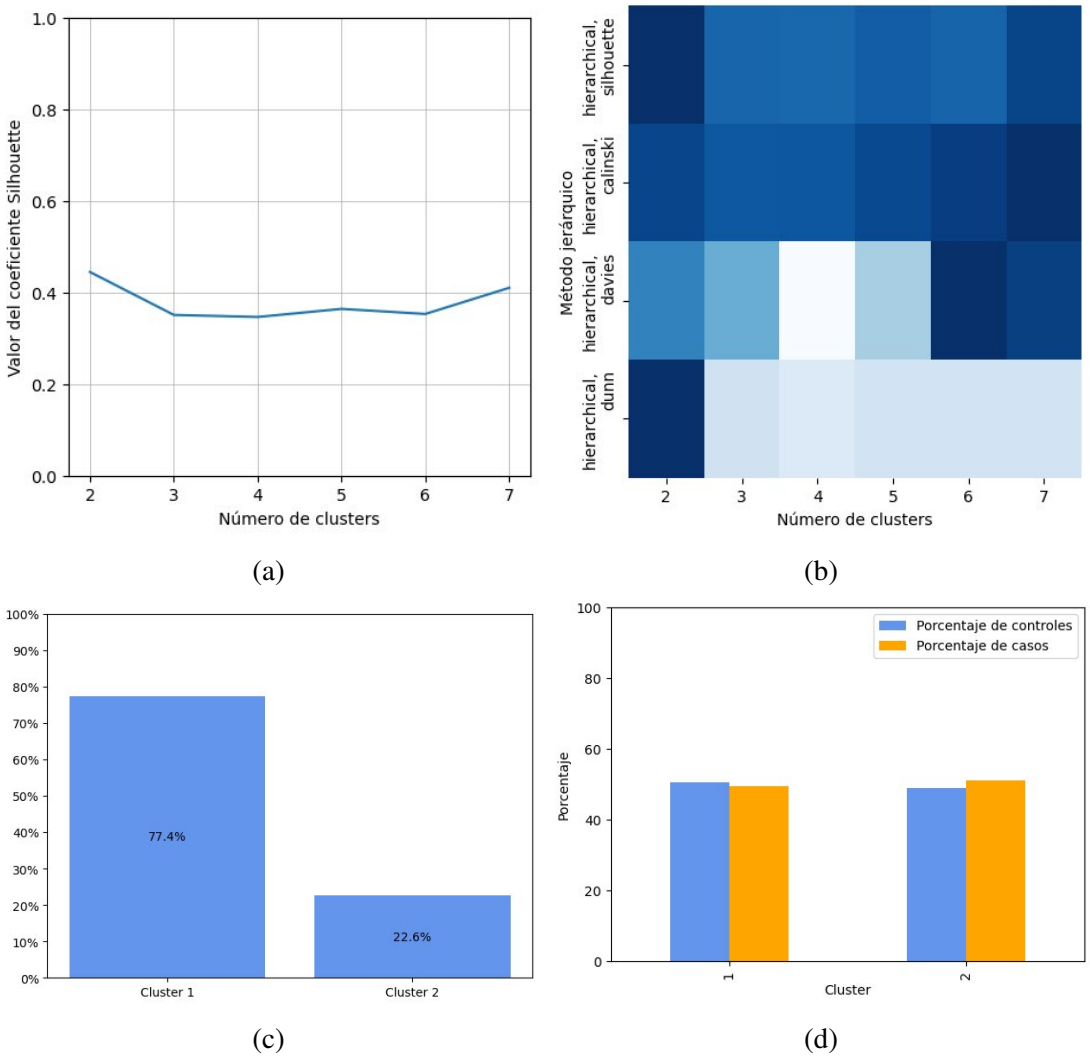


Figura 4.3: Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BBGAttitudeScale*. (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondiente a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

Como apoyo en el análisis de resultados y en la realización de comparaciones, la Figura 4.4 muestra la distribución correspondiente a los porcentajes de las respuestas proporcionadas por los pacientes en cada *cluster*, así como la distribución total en toda la base de datos. Además, en la Figura 4.4 (a) se muestran los perfiles de los subconjuntos para facilitar la comprensión de las diferencias entre ellos. Combinando la información de los perfiles y la distribuciíon de respuestas, se obtiene una visión más completa de cómo se agrupan y relacionan los pacientes en función de las respuestas proporcionadas en la encuesta sobre los episodios de hipoglucemia.

Estos resultados revelan diferencias significativas entre los *clusters*. Como se muestra en la Figura 4.4 (a), los subconjuntos se distinguen principalmente por las vairables *DealHypoEp*, *HighBGDamage* y *DangersHighBG*, cuyas respuestas más frecuentes para cada *clsuter* se observan en las Figuras 4.4 (c) y 4.4 (d). Los pacientes asignados al primer *cluster* no muestran preferencia por experimentar episodios hipolucémicos (*DealHypoEp*) en comparación con niveles altos de azúcar en sangre y su nivel de preocupación general (*HighBGDamage* y *DangersHighBG*) en comparación con niveles altos de azúcar en sangre es más neutro. En cambio, los pacientes del segundo *cluster* muestran una preferencia por experimentar niveles bajos de glucosa antes que episodios de hiperglucemia. Además, demuestran una mayor preocupación por las consecuencias de los niveles altos de glucosa en sangre.

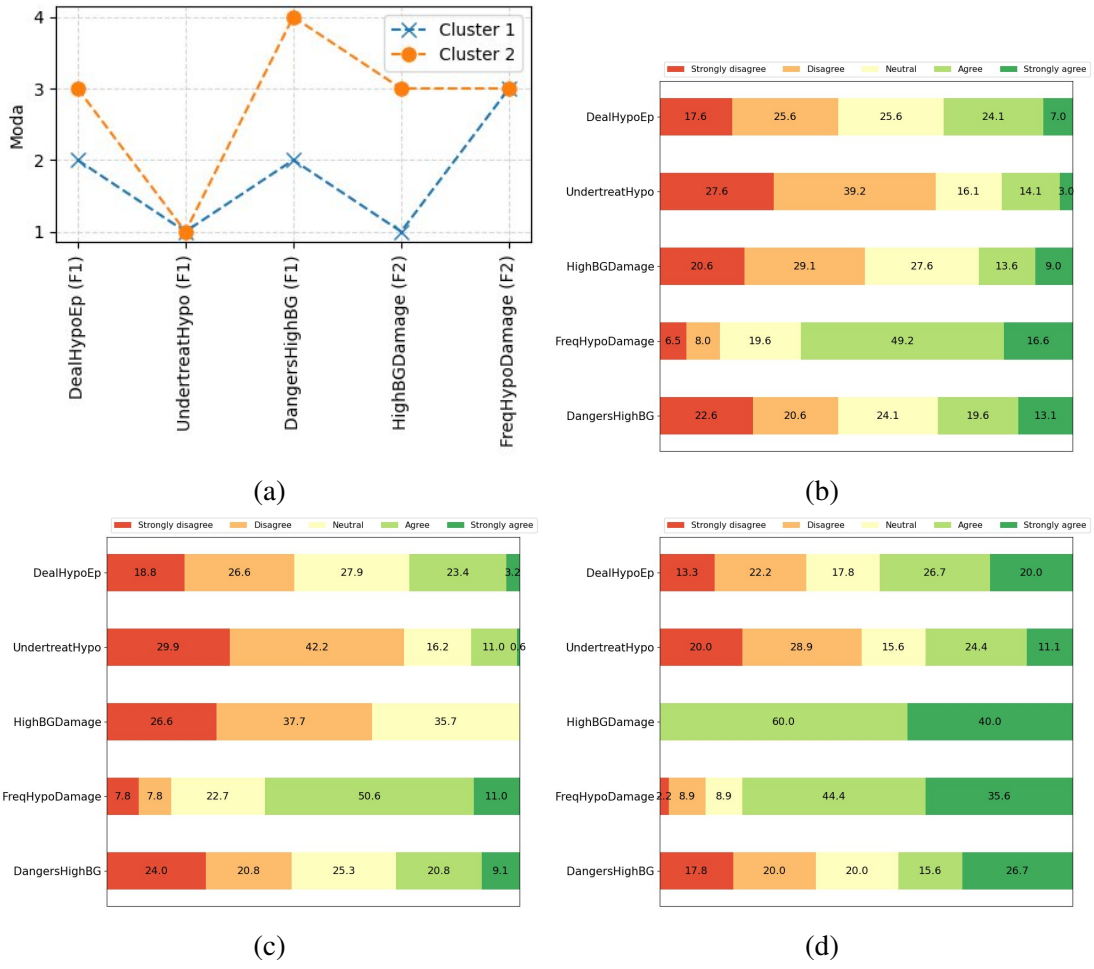


Figura 4.4: Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BBGAttitudeScale*. (a) Perfiles de los *clusters*; (b) Distribución de las respuestas de los pacientes en general; (c) Distribución de las respuestas de los pacientes en el *cluster 1*; (d) Distribución de las respuestas de los pacientes en el *cluster 2*.

4.3. Análisis factor exploratorio y *clustering* usando *BGeri-DepressScale*

Tras completar el preprocesamiento de la base de datos según se ha descrito en el capítulo anterior, se procedió a aplicar las técnicas de EFA y *clustering* para realizar un análisis detallado de los datos. Estos datos estaban asociados a una encuesta basada en la *Geriatric Depression Scale* que es una escala que se se usa para evaluar y detectar la depresión en adultos mayores. En esta base de datos, se utilizó el EFA con el propósito de descubrir patrones ocultos en un conjunto inicial de variables relacionadas con el grado de depresión en pacientes con DMT1 que presentan un riesgo de sufrir hipoglucemia.

Antes de aplicar el EFA, se realizó la prueba de esfericidad de Bartlett (véase sección 2.3) para evaluar la adecuación de las correlaciones entre variables. Los resultados de esta prueba indicaron un valor de chi-cuadrado de 1,038.47 con un p_{valor} inferior a 0.05. Estas cifras indican que existe evidencia suficiente para rechazar la hipótesis nula de que las variables no están correlacionadas, lo que respalda el uso del EFA en este conjunto de datos. El segundo paso, consistió en calcular el índice KMO para evaluar la adecuación de los datos. Se obtuvo un valor de 0.794 para este índice, lo cual indica que los datos son adecuados para este análisis. Es importante tener en cuenta que un valor de KMO cercano a 1 sugiere que las variables seleccionadas presentan una correlación adecuada y que el conjunto de datos es apropiado para llevar a cabo el EFA (véase sección 2.3).

Una vez comprobado y justificado el uso del EFA, se procedió a determinar el número óptimo de factores. Para ello, se ha utilizado el gráfico *scree plot* de los autovalores. En la Figura 4.5 (a), se puede observar que el número apropiado de factores para este conjunto de datos es 5, debido a que es el último número en el eje de abscisas que se corresponde con un valor de los autovalores superior a 1 en el eje de ordenadas. Posteriormente, se llevó a cabo el EFA utilizando el número óptimo de factores determinado y aplicando la rotación adecuada. Para esta base de datos en particular, se seleccionaron 5 factores y se utilizó la rotación *promax*. Los resultados obtenidos que proporcionan información sobre las relaciones entre las variables y los factores identificados en el análisis, se presentan en las Figura 4.5 (b) y Figura 4.5 (c).

- **Factor 1.** Agrupa aquellas variables cuyas preguntas correspondientes en la encuesta, evalúan aspectos positivos en la vida de los pacientes con DMT1 que presentan tendencia a experimentar hipoglucemia. Estas variables son *GoodSpirits* y *HappyMostTime*.
- **Factor 2.** Agrupa variables que hacen referencia a sentimientos negativos que pueden

experimentar los pacientes. Estas variables son *LifeEmpty*, *Worthless* y *HopelessSit*.

- **Factor 3.** Incluye las variables que evalúan el grado de actividad y energía que presentan los pacientes diabéticos tipo I coon tendencia a sufrir episodios de hipiglucemia. Estas variables son *DroppedActInt*, *PrefStayHome* y *FullEnergy*.

- **Factor 4.** En este factor se presentan las variables cuyas preguntas en la encuesta, permiten hacerle al paciente una evaluación general sobre el grado de satisfacción con su vida. Estas variables son *BasicSatLife*, *WondAlive* y *MostPplBetterOff*.

- **Factor 5.** Contrario al primer factor, en este se agrupan las variables que permiten realizar una evaluación de los aspectos negativos en la vida de los pacientes. Estas variables son *BoredOften*, *AfraidBad*, *HelplessOften* y *MoreMemProb*.

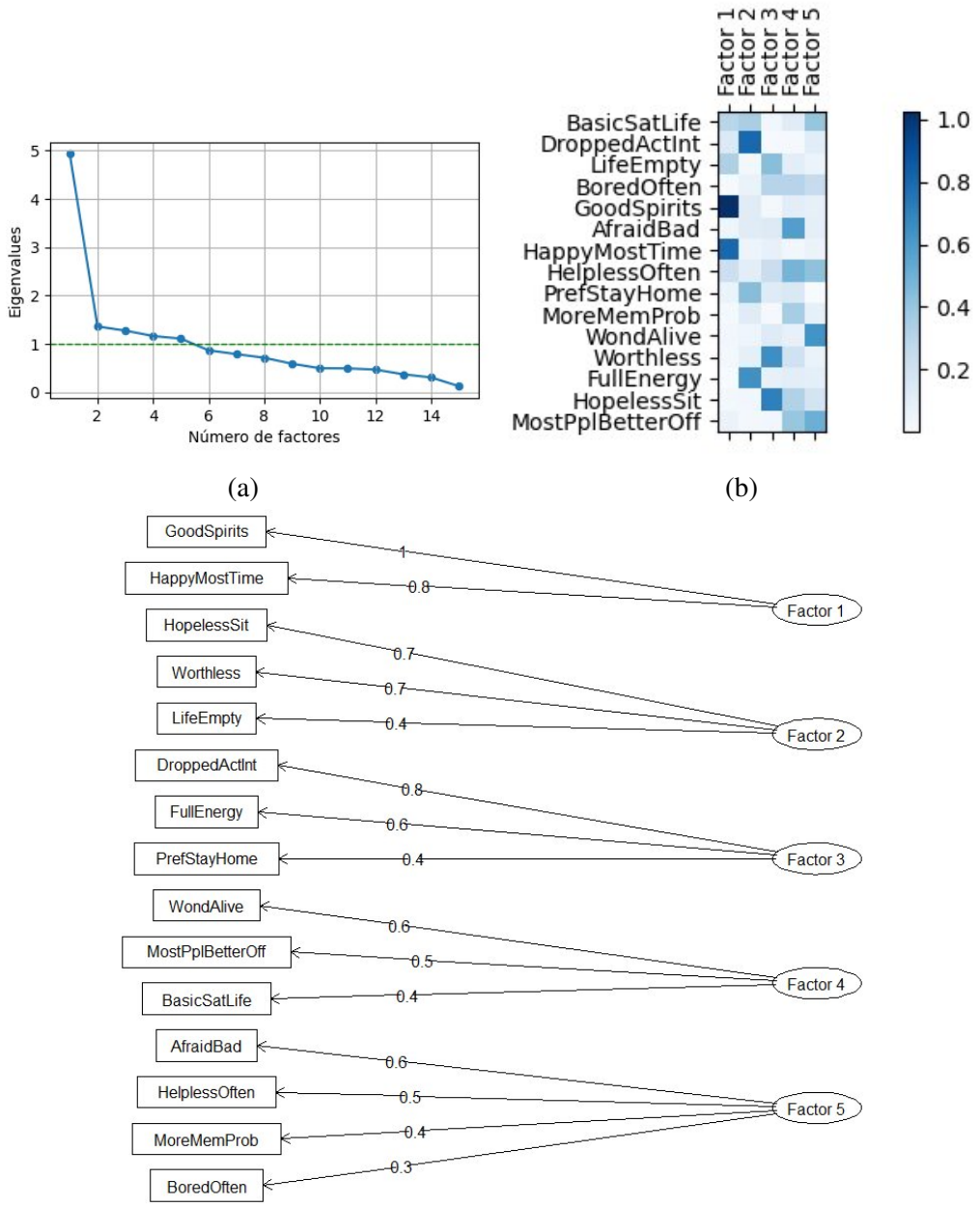


Figura 4.5: Resultados del EFA usando *BGeriDepressScale*. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Tras analizar los resultados del EFA, se procede a aplicar el algoritmo de agrupamiento sobre los resultados obtenidos en este primer paso. El paso previo a esta aplicación, es determinar el número de subconjuntos en los que se agruparán los datos. Para ello, se han utilizado el coeficiente de silueta y otros índices como métodos de validación (véase sección 2.4) mostrados en

las Figuras 4.6 (a) y 4.6 (b). Estos han permitido identificar el número óptimo de subconjuntos para este conjunto de datos, el cual es 2. A continuación, se procede a aplicar el algoritmo de agrupamiento para asignar cada paciente a un *cluster* epecífico.

Como resultado, se obtiene la distribución mostrada en la Figura 4.6 (c) que permite analizar la estructura de los subconjuntos identificados. En el caso específico de esta base de datos, se ha observado que el primer *cluster* agrupa al 87.5 % de los pacientes, mientras que el segundo *cluster* agrupa al 12.5 % restante. Adicionalmente, se ha implementado la Figura 4.6 (d) para observar la distribución de casos y controles dentro de cada subconjunto. Se puede apreciar que en el primero, el 48.57 % de los pacientes son controles y el 51.43 % son casos. En contraste, en el segundo *cluster* la mayoría de los pacientes asignados (el 56 %) son controles, mientras que el 44 % son casos. Estas proporciones pueden indicar que los subconjuntos representan diferentes perfiles de pacientes con respecto a las variables evaluadas en la encuesta. Para respaldar esta idea y facilitar las interpretaciones, se incluye la Figura 4.7 (a), que presenta los perfiles de ambos *clusters* de manera conjunta, lo que permite una comprensión más profunda de las diferencias y similitudes entre ellos. Además, en las Figuras 4.7 (c) y 4.7 (d) se muestra la distribución de las respuestas proporcionadas por los pacientes en cada *cluster*, y de forma general para el total de los pacientes en la Figura 4.7 (b).

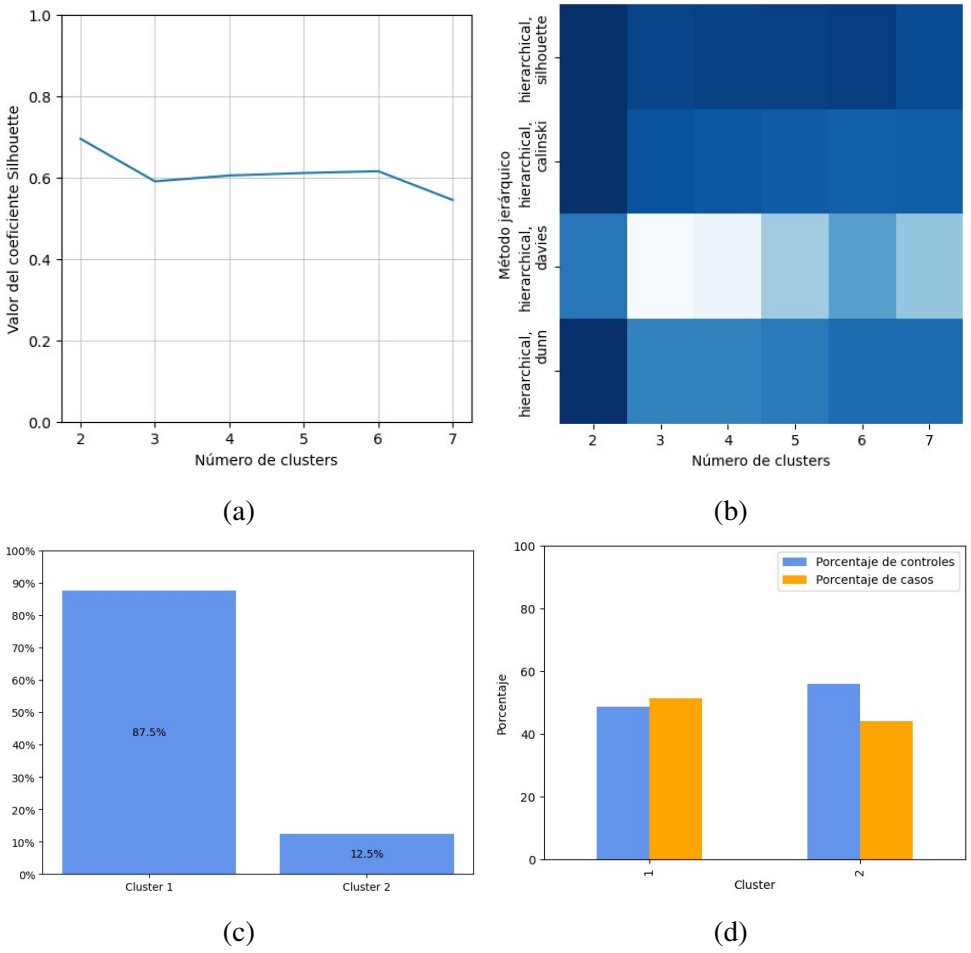


Figura 4.6: Mátricas para la determinación del número de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BGeriDepressScale*. (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondientes a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

Los resultados obtenidos muestran diferencias muy significativas entre los *clusters*. Como se puede observar en la Figura 4.7 (a), estos subconjuntos se distinguen principalmente por las diferencias entre las proporciones de respuestas de los pacientes a las preguntas asociadas a las variables *BasicSatLife*, *DroppedActInt*, *GoodSpirits*, *HappyMostTime*, *PrefStayHome* y *FullEnergy* que se muestran en las Figuras 4.7 (c) y 4.7 (d). Al interpretar estos resultados, se ha llegado a la conclusión de que los pacientes en el *cluster* 1 presentan una actitud bastante positiva en comparación con aquellos pertenecientes al segundo *cluster*. Esto demuestra que mediante el uso de esta técnica de ML se ha logrado agrupar a los pacientes en diferentes *clusters* de manera significativa.

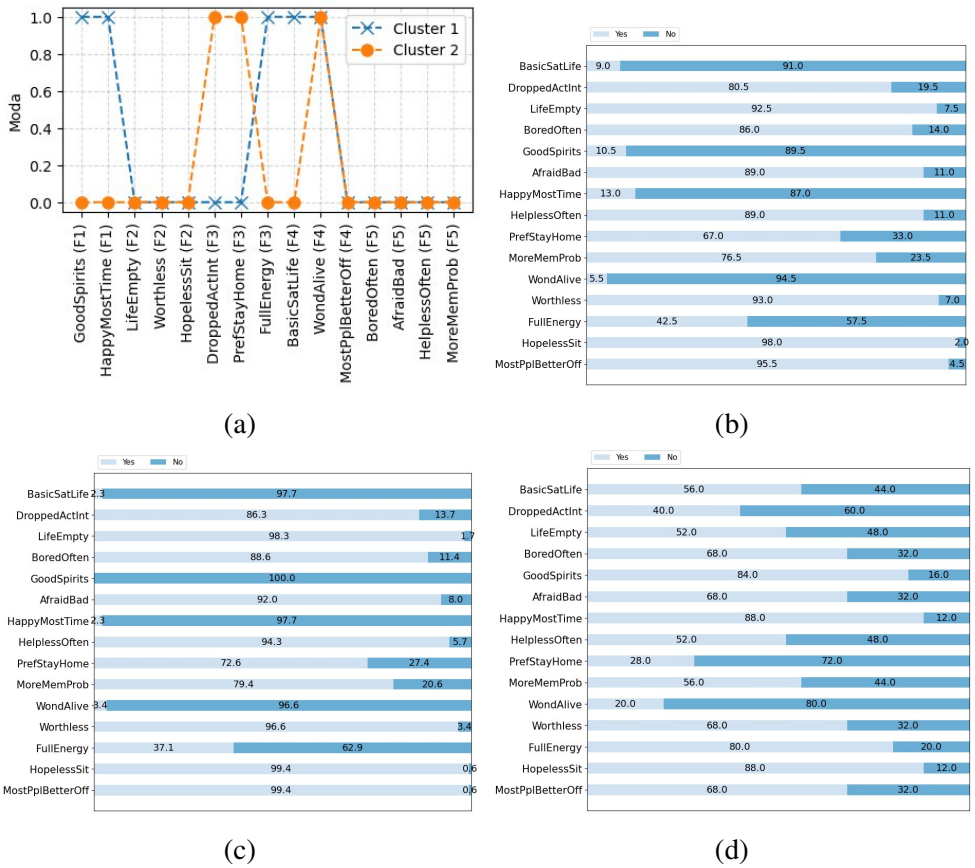


Figura 4.7: Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BGeriDepressScale*. (a) Perfiles de los *clusters*; (b) Distribución de las respuestas de los pacientes en general; (c) Distribución de las respuestas de los pacientes en el *cluster 1*; (d) Distribución de las respuestas de los pacientes en el *cluster 2*.

4.4. Análisis factor exploratorio y *clustering* usando *BHypoFearSurvey*

Se parte de la base de datos preprocesada en el capítulo anterior que contiene información sobre la *Hypoglycemia Fear Survey*, una encuesta que se usa para medir comportamientos y preocupaciones relacionadas la hipoglucemia en adultos con DMT1 (véase Tabla 3.4). Como herramientas de ML, se han empleado EFA y *clustering*.

El objetivo del EFA es revelar factores latentes que expliquen la estructura subyacente de las variables. En el caso de esta base de datos, se ha aplicado para identificar las relaciones entre las variables relacionadas con: (1) las preocupaciones de los pacientes acerca de los episodios

de hipoglucemia; (2) las medidas de prevención para evitarlos; y (3) los niveles de glucosa en sangre que les generan mayor seguridad. En primer lugar, se llevaron a cabo dos pasos previos esenciales antes de aplicar el EFA que consistieron en realizar la prueba esfericidad de Bartlett y obtener el índice KMO. Los resultados obtenidos en estos pasos justificaron el uso del EFA (véase sección 2.3). La prueba de efericidad de Bartlett mostró un valor de chi-cuadrado igual a 1,704.42 y un p_{valor} inferior a 0.05. En cuanto al índice KMO, se obtuvo un valor de 0.868. Estos resultados indican suficiente evidencia para rechazar la hipótesis nula de ausencia de correlación entre variables y confirman la idoneidad de los datos para este análisis, lo cual respalda la aplicación del EFA. A continuación, se procede a determinar el número óptimo de factores a retener en el modelo. Utilizando el método gráfico *scree plot*, se busca identificar el último punto en el eje de abscisas para el que los autovalores son superiores a 1. En este caso, al observar la Figura 4.8 (a), se puede apreciar que el número ideal de factores es 6.

Tras determinar el número óptimo de factores, se procede a utilizarlo junto con la rotación adecuada para ajustar el modelo del EFA. En este caso, se ha seleccionado un número de factores igual a 6 y se ha aplicado la rotación *promax*. Al aplicar el algoritmo con esta configuración, se obtienen los resultados que se presentan en las Figuras 4.8 (b) y 4.8 (c), puediendo observar en esta última las ponderaciones de cada variable con respecto a su factor correspondiente.

- **Factor 1.** Agrupa las variables cuyas preguntas en la encuesta evalúan las preocupaciones que experimentan los pacientes. Estas variables son *WorryNoFood*, *WorryPassOut*, *WorryEmbarSocial*, *WorryAppStupDrunk*, *WorryLoseCntrl*, *WorryNoHelp*, *WorryReactDrive*, *WorryMistAcc*, *WorryBadEvalCrit*, *WorryRespForOthers* y *WorryDizzy*.

- **Factor 2.** Incluye aquellas variables reflejan la preferencia de los pacientes por niveles altos de glucosa en sangre, en lugar de experimentar episodios de hipoglucemia. Estas variables son *LgSnackBed*, *TestBGRunHigh*, *RedInsThinkLowBG* y *KeepHighBGMtg*.

- **Factor 3.** Recoge aquellas variables que evalúan las preocupaciones y precauciones que toman los pacientes cuando están solos. Estas variables son *AvoidAloneLowBG*, *HighBGAlone* y *WorryReacAlone*.

- **Factor 4.** En este factor se encuentran las variables que evalúan la preocupación de los pacientes por no darse cuenta de que están experimentando un episodio hipoglucémico y el grado de control que realizan sobre sus niveles de azúcar. Estas variables son *CkSugOftMtg* y *WorryNotRecLowBG*.

- **Factor 5.** Representa a una única variable y es la que evalúa si los pacientes dejan de

hacer ejercicio cuando creen que los niveles de azúcar que tienen en sangre son bajos. Esta variable es *AvoidExThinkLowBG*.

- **Factor 6.** Agrupa aquellas variables que hacen referencia a las medidas que toman los pacientes para prevenir la hipoglucemia y actuar tan pronto como empiezan a experimentar los primeros síntomas. Estas variables son *EatFirstSignLowBG* y *CarryFastActSug*.

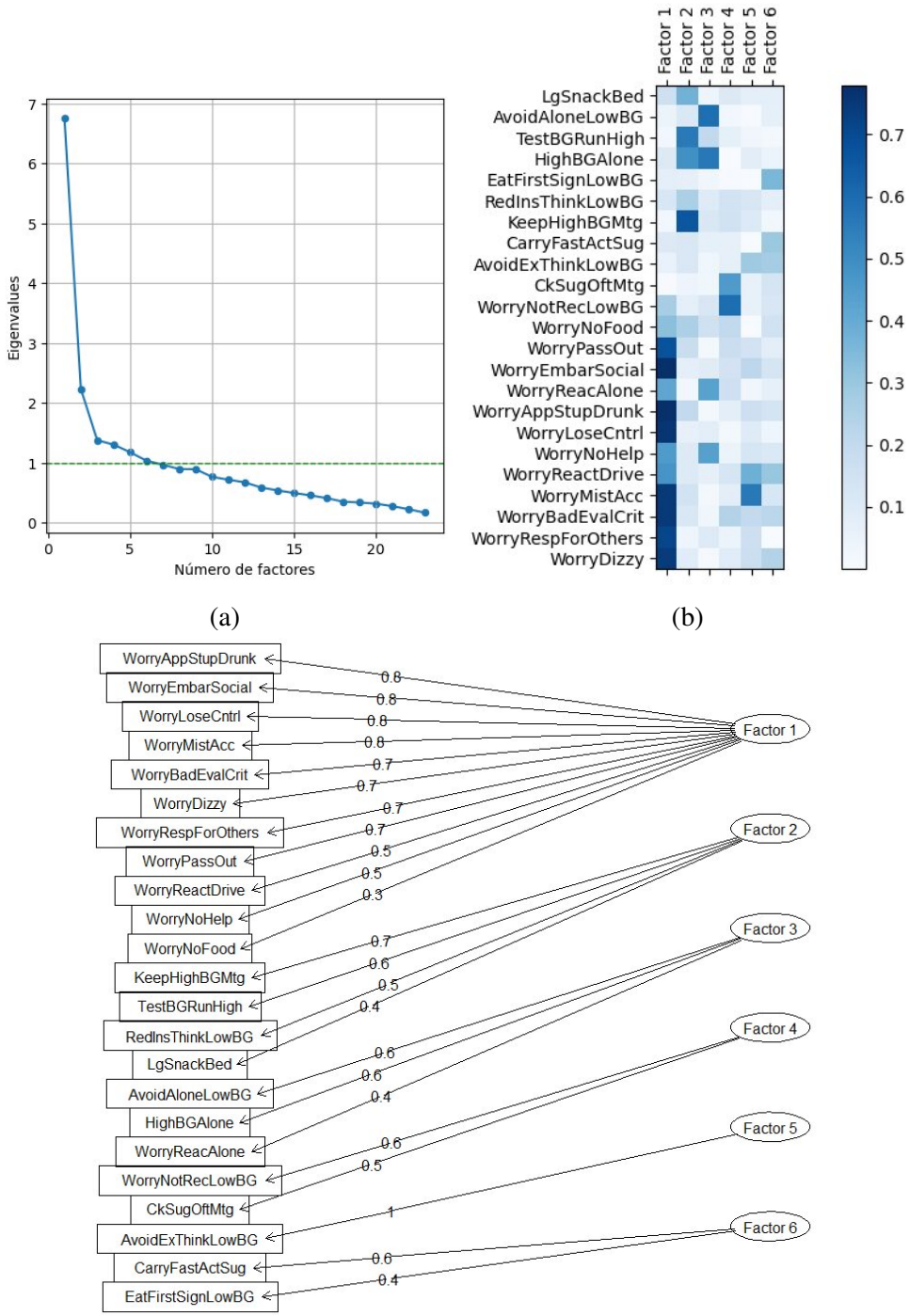


Figura 4.8: Resultados del EFA usando *BHypoFearSurvey*. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Una vez analizados los resultados del EFA, se procede a aplicar el algoritmo de *clustering*

sobre los resultados obtenidos. El primer paso es determinar el número óptimo de *clusters* en los que se agruparán los datos. Para ello, se han utilizado el coeficiente de silueta y algunos índices como métodos de validación (véase sección 2.4) mostrados en las Figuras 4.9 (a) y 4.9 (b), respectivamente. En el caso de esta base de datos, se ha determinado que el número de *clusters* adecuado es 2. A continuación, se procede a aplicar el algoritmo de *clustering* para la asignación de cada paciente a un *cluster* específico.

Como resultado, se obtuvo la distribución mostrada en la Figura 4.9 (c) que permite analizar la estructura de los subconjutos identificados. Esta representación visual facilita la observación de cómo se agrupan los pacientes en función de las características consideradas en el análisis. En este caso particular, el primer *cluster* ha sido asignado al 68.2 % de los pacientes, mientras que el segundo *cluster* está compuesto por el 31.8 % restante. Por otro lado, la Figura 4.9 (d) muestra el porcentaje de casos y controles presentes en cada *cluster*. Dentro del primer subconjunto, el 57 % de los pacientes son controles y el 43 % son casos, mientras que el segundo subconjunto está compuesto por el 34.9 % de casos y el 65.1 % de controles. Estas cifras podrían indicar que los subconjuntos son representativos de diferentes perfiles de pacientes con respecto a las variables evaluadas en la encuesta. Para respaldar esta idea y facilitar las interpretaciones, se incluye la Figura 4.10 (a), que presenta los perfiles de ambos *clusters* de forma conjunta, lo que permite una comprensión más profunda de las diferencias y similitudes entre ellos. Además, en las Figura 4.10 (c) y 4.10 (d) se muestra la distribución de las respuestas proporcionadas por los pacientes en cada cluster, y de forma general para el total de los pacientes en la Figura 4.10 (b).

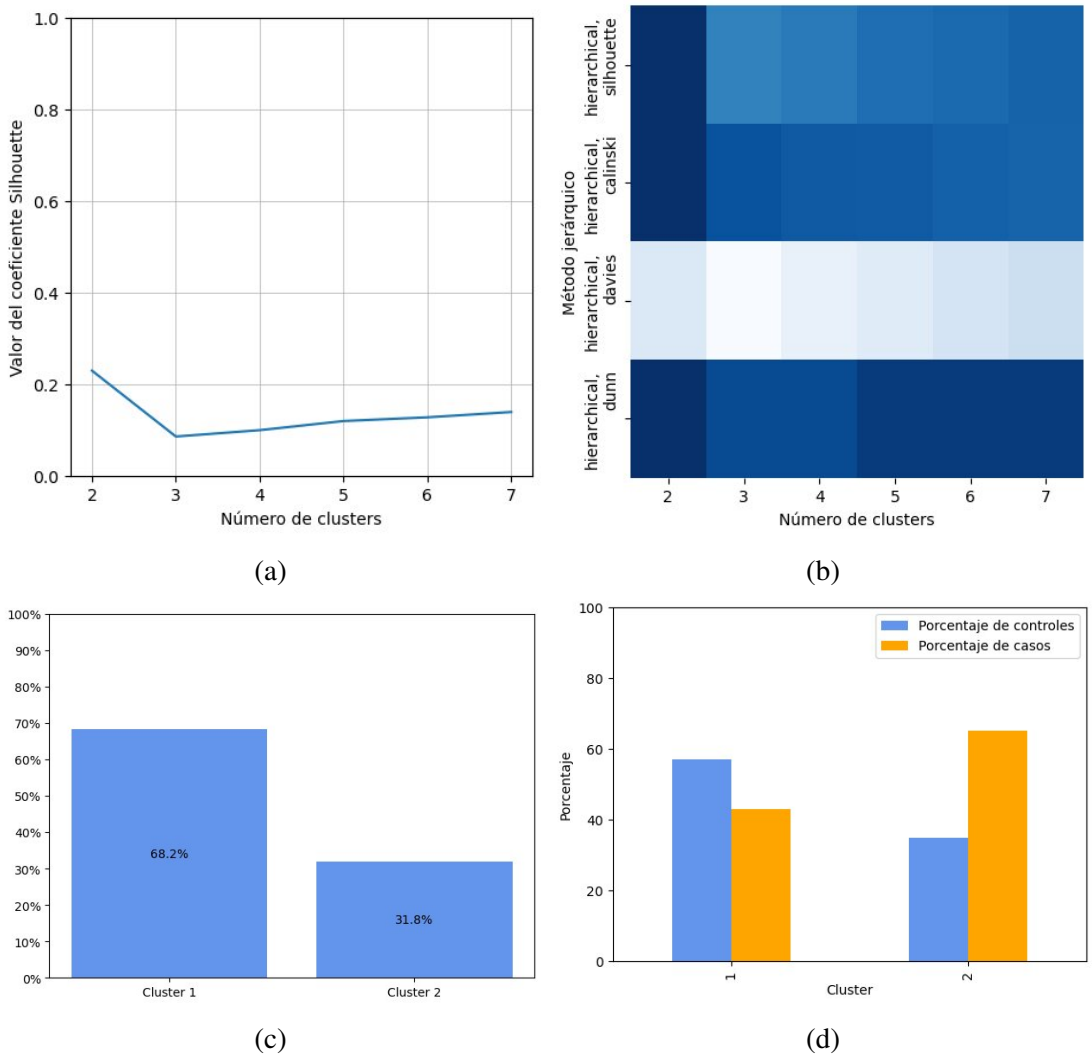


Figura 4.9: Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BHypoFearSurvey*. (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondiente a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

Los resultados obtenidos muestran diferencias significativas entre los *clusters*. Como se puede observar en la Figura 4.10 (a) estos subconjuntos se distinguen principalmente por las diferentes proporciones de respuestas de los pacientes a las preguntas asociadas con las variables *LgSnackBed*, *TestBGRunHigh*, *HighBGAlone*, *RedInsThinkLowBG*, *KeepHighBGMtg*, *CkSugOftMtg*, *WorryNoFood*, *WorryEmbarSocial*, *WorryReacAlone*, *WorryAppStupDrunk*, *WorryLoseCntrl*, *WorryNoHelp*, *WorryReactDrive*, *WorryMistAcc*, *WorryBadEvalCrit*, *WorryRespForOthers* y *WorryDizzy* que se muestran en las Figuras 4.10 (c) y 4.10 (d). Al interpretar los

resultados, se llega a la conclusión de que los pacientes en el primer *cluster* experimentan una mayor preocupación por los episodios de hipoglucemia, en comparación con aquellos pacientes asociados al segundo *cluster*. Como consecuencia, se observa que toman más precauciones y se esfuerzan más para evitar dichos episodios.

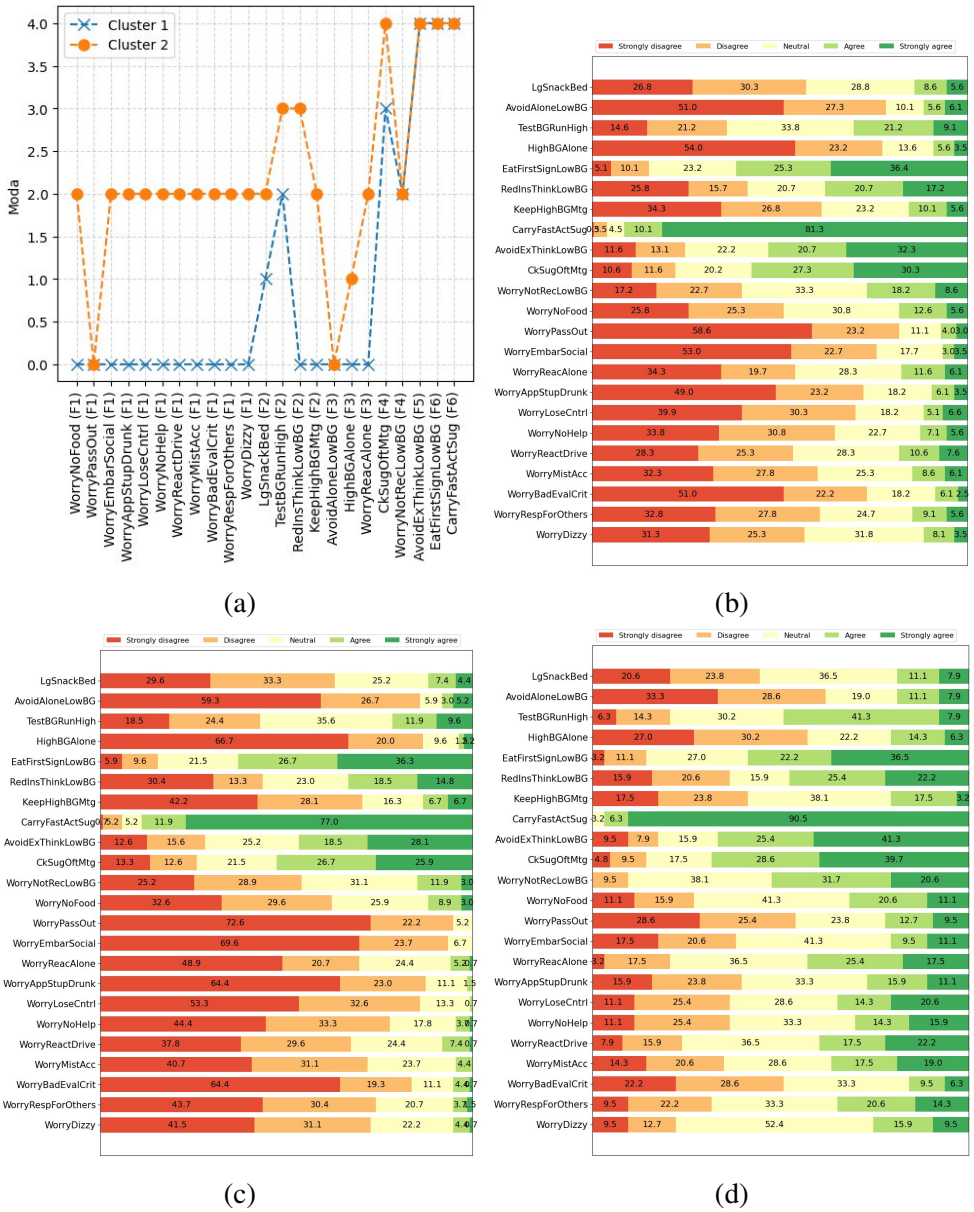


Figura 4.10: Representación conjunta de los perfiles de cada *cluster* y distribución de las respuestas de los pacientes de forma general y en cada *cluster* de la base de datos *BHypoFearSurvey*. (a) Perfiles de los *clusters*; (b) Distribución de las respuestas de los pacientes en general; (c) Distribución de las respuestas de los pacientes en el *cluster 1*; (d) Distribución de las respuestas de los pacientes en el *cluster 2*.

4.5. Análisis factor exploratorio y *clustering* usando *BHypo-UnawareSurvey*

Se va a utilizar el EFA y el *clustering* para explorar los datos de la base de datos cuyo preprocesamiento se ha presentado en el capítulo anterior. Estas técnicas se emplearán con el objetivo de extraer información relevante y descubrir patrones subyacentes en las variables que evalúan principalmente la capacidad de los pacientes para reconocer los síntomas asociados a los episodios de hipoglucemia.

El primer paso previo a la aplicación del EFA consiste en justificar su uso mediante la prueba de esfericidad de Bartlett y el índice KMO. En la prueba de esferificad de Bartlett, se obtuvo un valor de chi-cuadrado igual a 484.52 y un p_{valor} inferior a 0.05, lo que indica que existe suficiente evidencia para rechazar la hipótesis nula que establece que las variables son independientes entre sí. En cuanto al índice KMO, se obtuvo un valor de 0.806, lo que sugiere que las variables tienen adecuada correlación y son apropiadas para la aplicación del EFA. A continuación, se busca identificar el número adecuado de factores a retener en el modelo. Para ello, se utiliza el gráfico de valores *scree plot*, que muestra los autovalores en orden descendente en el eje de ordenadas y los números de factor en el eje de abscisas. Se busca identificar el último número mayor que 1 en el eje de ordenadas porque son considerados más significativos (véase sección 2.3). En este caso particular, en la Figura 4.11 (a) se observa que el número óptimo de factores es 2. Por lo tanto, se ajustará el algoritmo con un número de factores igual a 2 y la rotación correspondiente, *promax*. Tras aplicar el EFA se presentan los resultados en las Figuras 4.11 (b) y 4.11 (c), donde se observan las cargas factoriales de las variables en cada factor. Estas ponderaciones determinan el grado de correlación que hay entre cada factor y las variables que se le han sido asignadas por el modelo.

- **Factor 1**. Agrupa las variables que se centran en la evaluación de los síntomas que experimentan los pacientes cuando sufren episodios de hipoglucemia. Estas variables son *LowBGSympCat*, *LowBGLostSymp*, *FeelSympLowBG* y *ExtentSympLowBG*.
- **Factor 2**. Incluye las variables que evalúan la frecuencia de los episodios de hipoglucemia que sufren los pacientes en un periodo de tiempo determinado y sus síntomas asociados. Estas variables son *ModHypoEpPast6Mon*, *SevHypoEpPastYear*, *Bel70PastMonWSymp* y *Bel70PastMonNoSymp*.

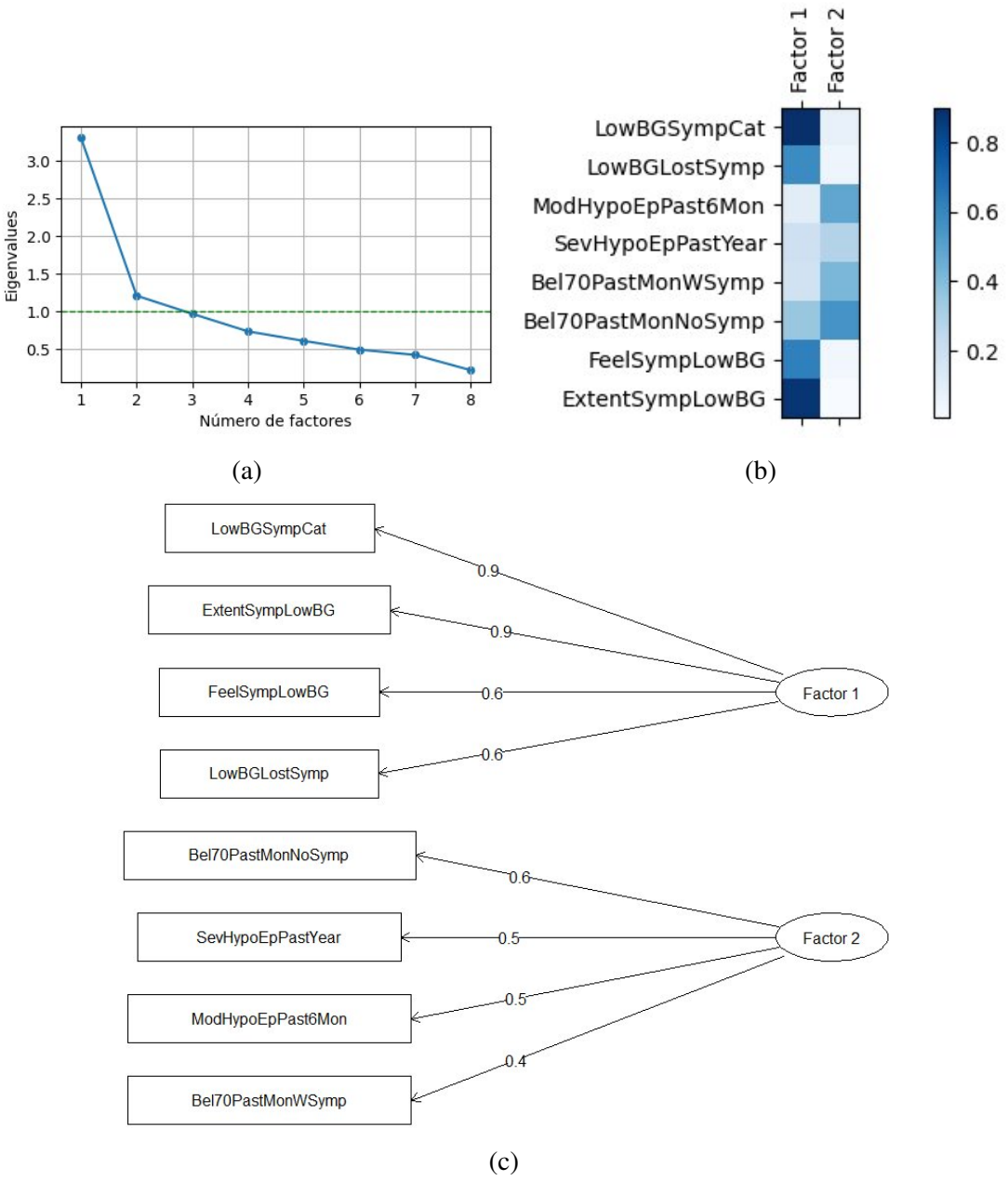


Figura 4.11: Resultados del EFA usando *BHypoUnawareSurvey*. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Una vez analizados los resultados obtenidos en el EFA, estos se utilizan para aplicar el algoritmo de *clustering*. En primer lugar, es necesario determinar el número óptimo de *clusters* en los que se agruparán los datos. Para ello, se han utilizado métodos jerárquicos y se ha dado especial énfasis al método de silueta, que evalúa la calidad de la asignación de los pacientes a

los subconjutos. Para el conjunto de datos con los que se está trabajando, el número óptimo de *clusters* es 2, como se muestra en las Figura 4.12 (a) y 4.12 (b). A continuación, se aplica el algoritmo de agrupameinto y se obtiene la distribución de los pacientes en cada subconjunto que se muestra en la Figura 4.12 (c). Esta representación visual proporciona información sobre cómo se agrupan y se relacionan entre sí los pacientes en función de las características consideradas en el análisis. En este caso, el *cluster* 1 ha sido asignado al 67.2 % de los pacientes, mientras que el *cluster* 2 está compuesto por el 32.8 % restante.

Por otra parte, se ha implementado la Figura 4.12 (d) para ver el porcentaje de casos y controles que hay en cada *cluster*. Dentro del primer *cluster*, el porcentaje de casos es del 37.77 %, mientras que el porcentaje de controles es del 62.23 %. En cuanto al segundo *cluster*, se observa que el 24.25 % de sus pacientes son controles y el 75.75 % son casos.

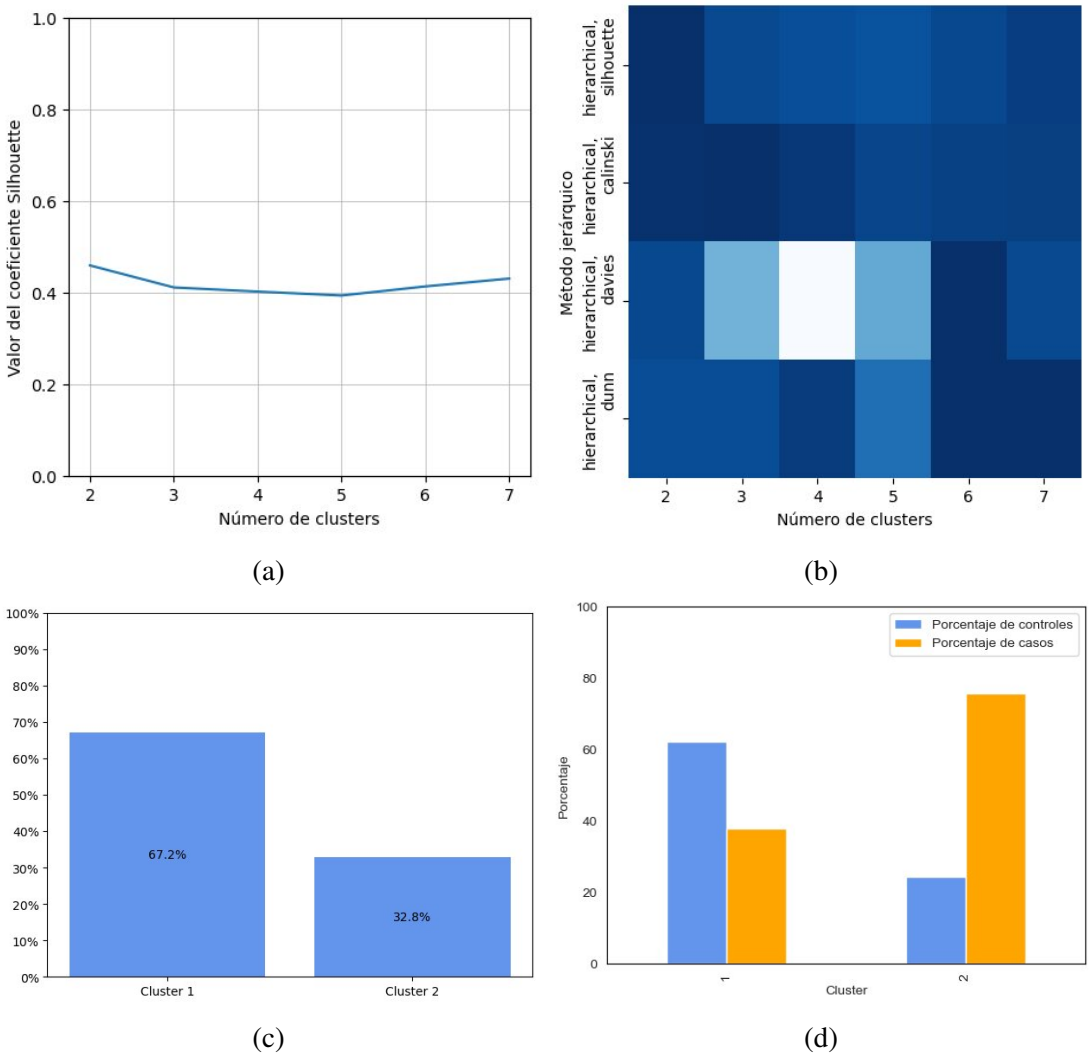


Figura 4.12: Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BHypoUnawareSurvey* . (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondiente a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

Como apoyo en el análisis de resultados y en la realización de comparaciones entre subconjuntos, la Figura 4.13 muestra los perfiles de cada *cluster* de manera conjunta. Esto permite visualizar y comparar las características y patrones asociados a cada subconjunto en términos de las variables consideradas en el estudio. Además, la información prporcionada en esta Figura, es útil para hallar y analizar las diferencias y similitudes entre los subconjuntos identificados de forma automática por el algoritmo. Se observa que las principales diferencias entre los *clusters* se dan en las respuestas de los pacientes a las preguntas asociadas a todas las variables, excepto

en la variable denominada *SevHypoEpPastYear* en la cual coinicde la moda de las respuestas en ambos *clusters*. Al interpretar los resultados, se observa que la distinción entre ambos subconjuntos se ha realizado en base a la frecuencia de los episodios de hipoglucemia y los síntomas que experimentan los pacientes. Los perfiles asociados a cada *cluster* reflejan que los pacientes del primer subconjunto, tienen una mayor tendencia a experimentar síntomas cuando se producen los episodios de hipoglucemia, aunque los sufran menos en comparación con los pacientes del segundo subconjunto.

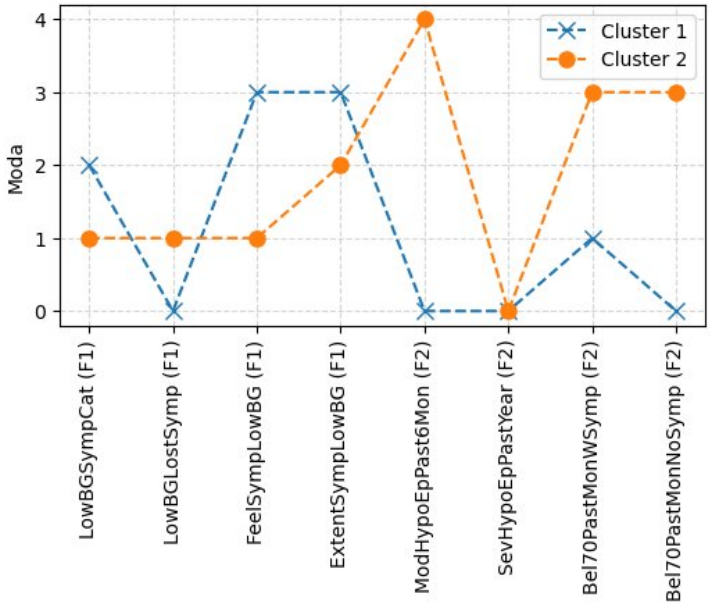


Figura 4.13: Perfiles de los *clusters* de la base de datos *BHypoUnawareSurvey*.

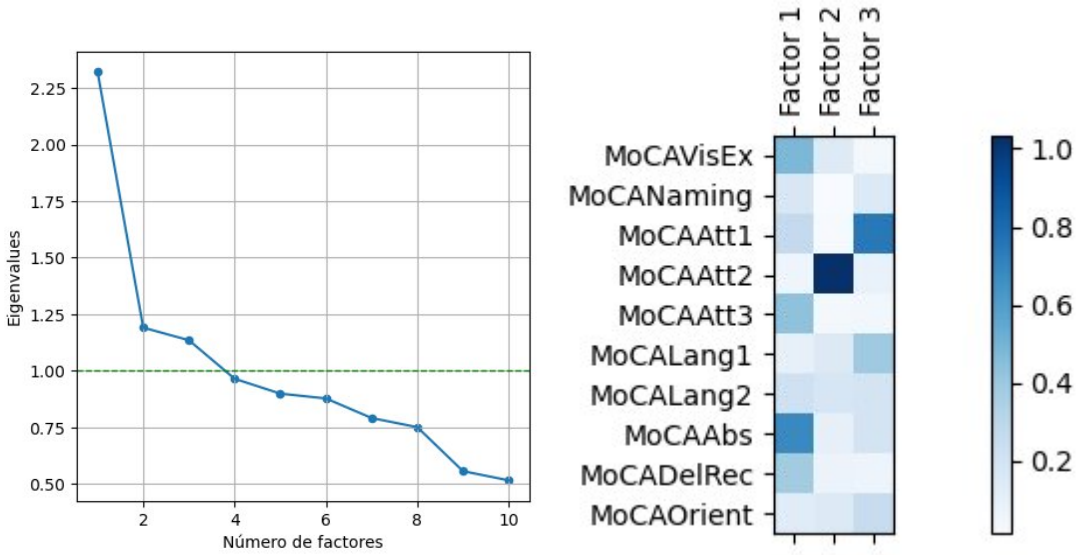
4.6. Análisis factor exploratorio y *clustering* usando *BMoCA*

Partiendo de la base de datos obtenida tras el preprocesamiento descrito en el capítulo anterior, se han llevado a cabo las dos técnicas de análisis previamente expuestas: EFA y *clustering*.

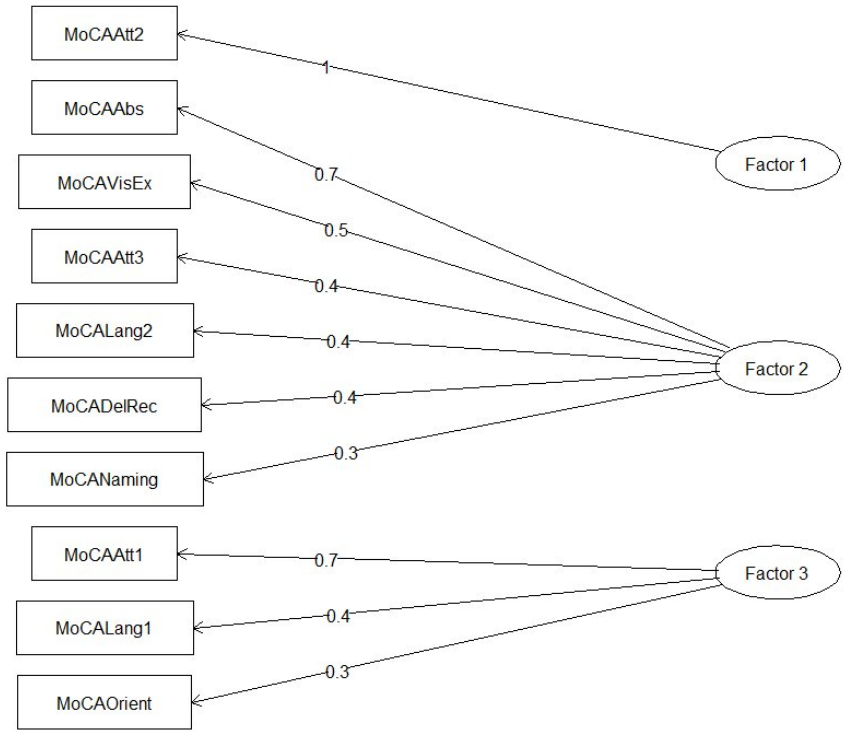
En el contexto de la base de datos *BMoCA*, el EFA se ha utilizado para identificar patrones ocultos en un conjunto inicial de variables relacionadas con las capacidades cognitivas de los pacientes con DMT1 que presentan un riesgo de sufrir episodios de hipoglucemia. Antes de aplicar el algoritmo del EFA, se ha realizado la prueba de esfericidad de Barlett para evaluar la adecuación de las correlaciones entre variables. Los resultados de esta prueba, indicaron un valor de chi-cuadrado de 175.85 y un p_{valor} inferior a 0.05. Esto indica que existen correlaciones entre las variables que pemiten rechazar la hipótesis nula y justifica el uso del EFA. Adicional-

mente, se ha calculado el índice KMO para evaluar la idoneidad de los datos para la aplicación del EFA. Se ha obtenido un valor de KMO igual a 0.666, lo cual confirma que los datos son adecuados para el análisis que se quiere abordar. A continuación, se busca identificar el número adecuado de factores a retener en el modelo. Para ello, se utiliza el gráfico de autovalores mostrado en la Figura 4.14 (a) para elegir el último número en el eje de abscisas que se corresponde con un autovalor superior a 1 en el eje de ordenadas. En este caso, se observa que el número óptimo de factores es 3. Por lo tanto, se procede a aplicar el algoritmo de agrupamiento con este número de factores y la rotación correspondiente, la *promax*. Los resultados obtenidos se muestran en las Figuras 4.14 (b) y 4.14 (c), donde se pueden observar las correlaciones entre factores y variables a través de las ponderaciones o cargas factoriales.

- **Factor 1.** Está formado por una variable que en el test sobre las capacidades cognitivas de los pacientes evalúa la atención. Esta variable es *MoCAAtt2*.
- **Factor 2.** Recoge las variables que se corresponden con preguntas relacionadas con las pruebas visoespacial, de atención, de nombres, de lenguaje y abstracción. Estas variables son *MoCAVisEx*, *MoCANaming*, *MoCAAtt3*, *MoCALang2*, *MoCAAbs* y *MoCADelRec*.
- **Factor 3.** Incluye aquellas variables cuyas preguntas en la encuesta basada en el *Montreal Cognitive Assessment* evalúan la atención, las capacidades de lenguaje y de orientación. Estas variables son *MoCAAtt1*, *MoCALang1* y *MoCAOrient*.



(a) (b)



(c)

Figura 4.14: Resultados del EFA usando *BMoCA*. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Una vez analizados los resultados obtenidos en el EFA, estos se utilizan para aplicar el

algoritmo de agrupamiento. El primer paso es determinar el número óptimo de subconjuntos en los que se agruparán los datos. Para este propósito, se han utilizado métodos jerárquicos y se ha determinado que el número adecuado de *clusters* para los datos con los que se está trabajando es 2, como se muestra en las Figuras 4.15 (a) y 4.15 (b). Después, se procede a aplicar el algoritmo de *clustering* y se obtiene la distribución de los pacientes en cada subconjunto, como se muestra en la Figura 4.15 (c). Esta representación visual permite observar la agrupación de los pacientes en función de las características consideradas en el análisis. En este caso particular, el primer subconjunto ha sido asignado al 96 % de los pacientes, mientras que el segundo subconjunto está formado por el 4 % restante. Adicionalmente, se ha graficado la Figura 4.15 (d) donde se muestra el porcentaje de casos y controles presentes en cada subconjunto. Dentro del primer *cluster*, el 51.04 % de los pacientes son controles y el 48.96 % son casos, mientras que el segundo subconjunto está compuesto por el 75 % de casos y el 25 % de controles.

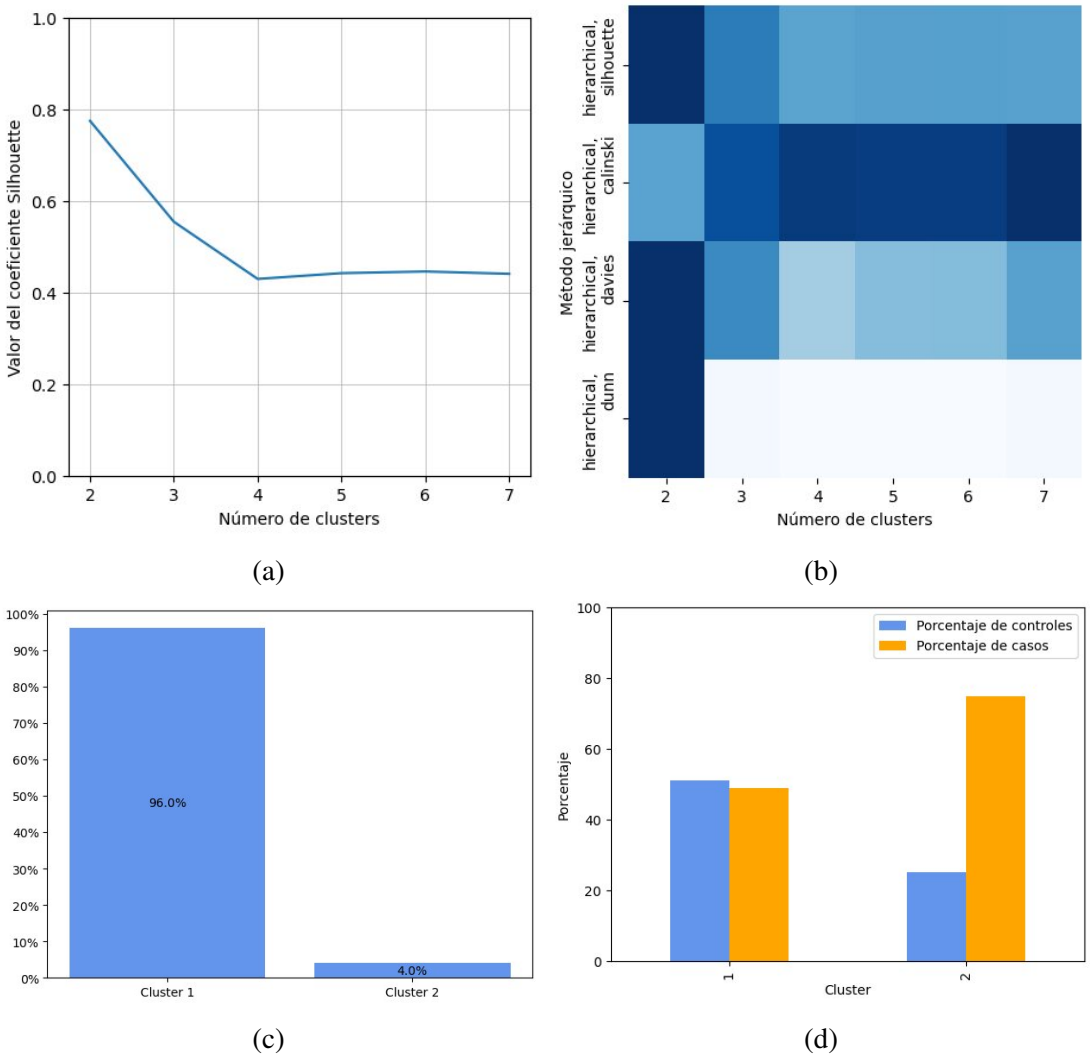


Figura 4.15: Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para la base de datos *BMoCA*. (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondiente a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

La Figura 4.16 proporciona un apoyo en el análisis de resultados y en la realización de comparaciones, ya que muestra de manera conjunta los perfiles de los *clusters* identificados. Estos perfiles permiten una comprensión más profunda de las similitudes y diferencias entre los *clusters* formados por el algoritmo. Al examinar los perfiles, se pueden identificar características y patrones distintivos para la interpretación de las agrupaciones. En este caso, se ve una clara diferencia entre la moda de algunas variables en ambos *clusters*. En aquellas variables distintivas, se observa que la moda es superior en el primer *cluster*, lo cual indica que los pacientes de este

cluster han obtenido puntuaciones más altas en las pruebas realizadas. Por lo tanto, se puede afirmar que el algoritmo de agrupamiento se ha basado en las puntuaciones obtenidas por cada paciente para formar los *clusters*. En consecuencia, se deduce que los pacientes pertenecientes al primer *cluster* presentan mejores capacidades cognitivas en comparación con aquellos pacientes pertenecientes al segundo *cluster*.

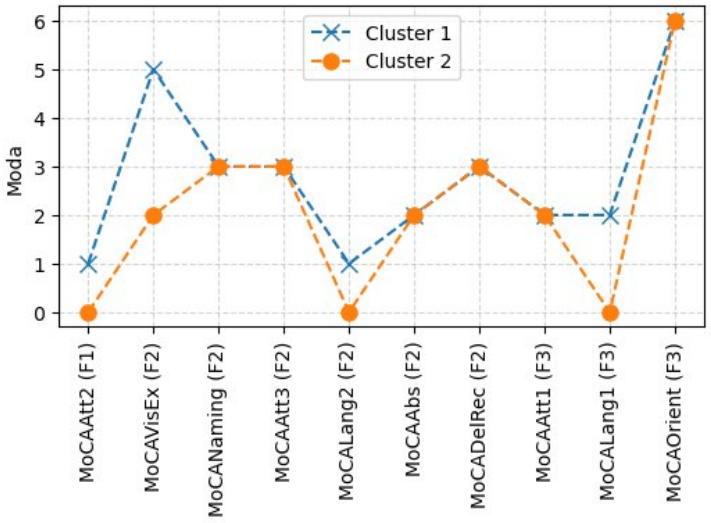


Figura 4.16: Perfiles de los *clusters* de la base de datos *BMoCA* .

4.7. Análisis factorial exploratorio y *clustering* usando bases concatenadas

Esta base de datos se obtiene como resultado de combinar todas las variables de las 5 bases de datos previamente analizadas. La concatenación se ha realizado utilizando el identificador de cada paciente. En cuanto al preprocesamiento, se siguió un proceso muy similar al descrito anteriormente, pues las bases de datos se concatenaron una vez que estaban preprocesadas, sin eliminar ningún paciente. El único paso adicional realizado en este caso fue analizar los valores nulos que formaban parte de la base de datos concatenada. Dado que estos valores correspondían a pacientes que habían respondido a más del 50 % de las preguntas, estos no podían ser eliminados. En su lugar, los valores NaN fueron reemplazados por la moda de cada variable. La moda es una medida de tendencia central que representa el valor más frecuente en esa variable. La base de datos obtenida en este preprocesamiento, es la que se va a utilizar para la aplicación del EFA y *clustering*.

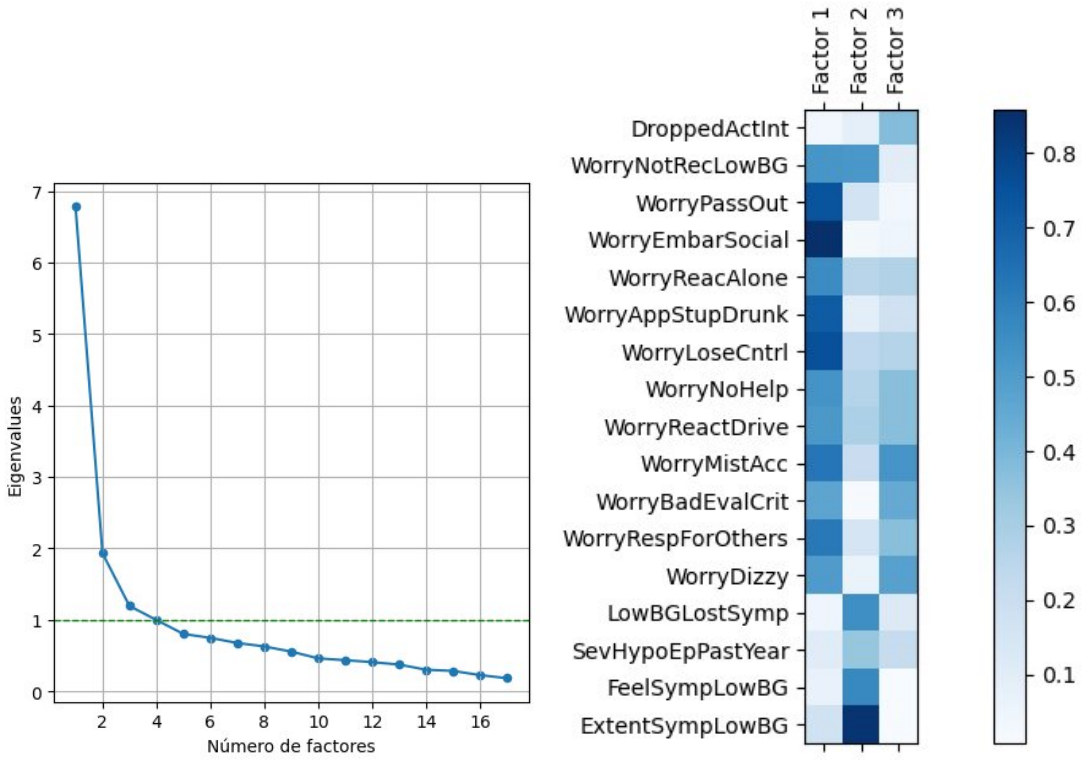
El EFA se utiliza con el objetivo de identificar patrones ocultos en los conjuntos de datos y determinar factores latentes que relacionen todas las variables asociadas con encuestas previamente analizadas de manera individual. En este caso, se ha partido con 201 pacientes y un número de variables igual a 62. El paso previo a la aplicación del EFA es realizar la prueba de esfericidad de Bartlett y calcular el índice KMO. En el caso de la prueba de esfericidad de Bartlett, se obtuvo un valor de chi-cuadrado igual a 5,161.216 y un p_{valor} inferior a 0.05, lo que respalda el uso del EFA. En cuanto al índice KMO, en este caso se ha abordado de manera un tanto distinta. Como se mencionó en la sección 2.3 de este trabajo, un índice KMO se considera meritorio cuando es superior a 0.8. Hasta el momento, se había considerado un umbral inferior debido a que se trabajaba con conjuntos de variables de baja dimensionalidad y se conocía previamente su relación basada en la literatura y el análisis descriptivo correspondiente a la base de datos de la que provenían. No obstante, en este caso se ha aplicado un filtro adicional para descartar las variables cuyo KMO es inferior a 0.75. Inicialmente, con todas las variables incluidas, se obtuvo un KMO igual a 0.728 y después de aplicar el filtro, el número de variables se redujo a 17 y se obtuvo un nuevo valor de KMO de 0.894. Estas 17 variables son las que se utilizarán para aplicar los algoritmos de ML. Con estos resultados del test de esfericidad de Bartlett y del índice KMO, se puede justificar el uso del EFA.

A continuación, se ha llevado a cabo la determinación del número de factores. Para ello se ha utilizado el gráfico *scree plot* que muestra los autovalores en orden descendente y se identifica el último punto en el que estos son superiores a 1. En la figura 4.17 (a) se puede observar que para esta base de datos el número de factores óptimo es 3. Este número se utilizará junto con la rotación *promax* para ajustar el algoritmo del EFA que una vez aplicado, se han obtenido los resultados que se muestran en las Figuras 4.17 (b) y 4.17 (c). De esta forma, las variables quedan agrupadas en factores y se puede observar la ponderacióon de cada una de ellas con respecto a su factor correspondiente en la Figura 4.17 (c).

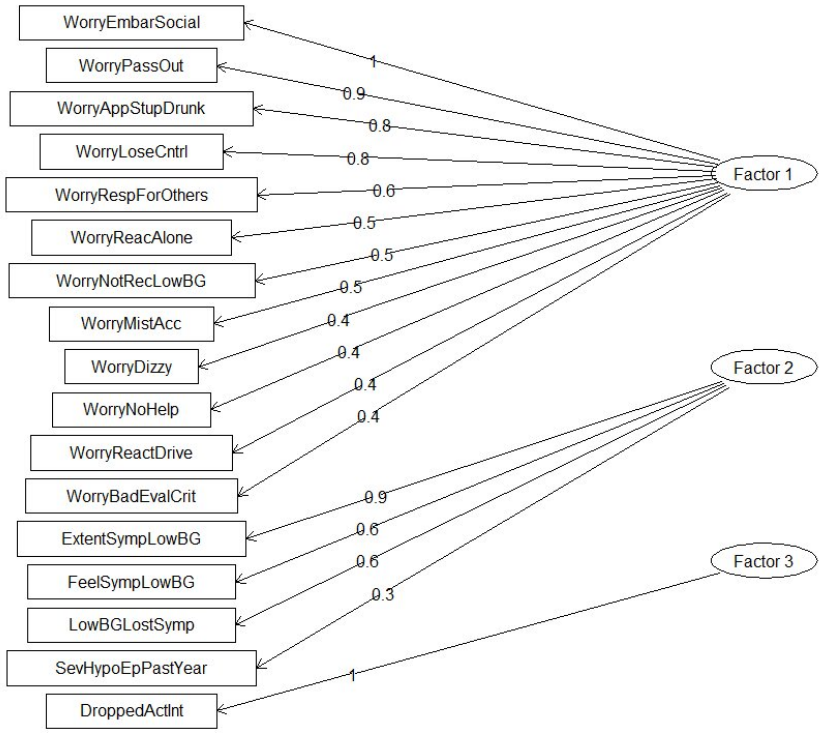
- **Factor 1**. Agrupa aquellas variables asociadas a las preocupaciones de los pacientes. Estas son *WorryNotRecLowBG*, *WorryPassOut*, *WorryEmbarSocial*, *WorryReacAlone*, *WorryAppStupDrunk*, *WorryLoseCntrl*, *WorryNoHelp*, *WorryReactDrive*, *WorryMistAcc*, *WorryBadEvalCrit*, *WorryRespForOthers*, *WorryDizzy*.

- **Factor 2**. Se incluyen aquellas variables cuyas preguntas asociadas en las encuestas hacen referencia a los síntomas que experimentan los pacientes que sufren DMT1 con tendencia a sufrir hipoglucemia. Estas variables son *LowBGLostSymp*, *SevHypoEpPastYear*, *FeelSympLowBG* y *ExtentSympLowBG*.

- **Factor 3**. Este factor representa una única variable relacionada con el las actividades e intereses de los pacientes. Esta variable es *DroppedActInt*.



(a) (b)



(c)

Figura 4.17: Resultados del EFA usando las bases de datos concatenadas. (a) Gráfico de los autovalores para determinar el número de factores; (b) Resultado del EFA que asigna cada variable a un factor; (c) Diagrama de pesos que relaciona cada variable con su factor correspondiente.

Una vez analizados los resultados del EFA, el siguiente paso es aplicar el algoritmo de *clustering* sobre los resultados obtenidos de este primer análisis. El paso previo a la aplicación de esta herramienta, es determinar el número óptimo de *clusters* en los que se agruparán los datos. Para ello, se han utilizado métodos jerárquicos dándole especial énfasis al método que maximiza la coherencia interna y la separación entre subconjuntos, el método silueta. Para el conjunto de datos con los que se está trabajando, el número adecuado de *clusters* es 4, como se puede ver en las Figuras 4.18 (a) y 4.18 (b).

A continuación, se procede con la aplicación del algoritmo de agrupamiento. Esta técnica permite obtener la distribución de los pacientes en cada *cluster* como se muestra en la Figura 4.18 (c). Esta representación propociona información sobre cómo se agrupan los pacientes de manera automática en función de las características considerdas. Por otra parte, se ha implementado la Figura 4.18 (d) para ver el porcentaje de casos y controles que hay dentro de cada *cluster*, lo que ayuda en el análisis y comparación de la composición de los clusters en términos de casos y controles. En el caso particular de esta base de datos, se han separado a los pacientes en 4 *clusters*. Al *cluster* 1 pertenece el 51.7 % del total de los pacientes. El *cluster* 2 está formado por el 10.4 % de los pacientes. El *cluster* 3 representa el 15.4 % del total de los pacientes y por el último al *cluster* 4 se le ha asociado el 22.4 % restante. Como se puede observar en la Figura 4.18 (d) el primer *cluster*, con un 70.19 % de controles y un 29.81 % de casos, podría ser representativo de un subconjunto característico de controles, mientras que los tres *clusters* restantes presentan un porcentaje mayor de casos. En el segundo *cluster* el porcentaje de los casos es del 80.95 % y el de los controles el 19.05 % del total de los pacientes asignados a este *cluster*. En el tercer *cluster* se tiene que el 70.96 % de los pacientes son casos y el 29.04 % restante pertenece a los controles. Por último, el cuarto *cluster* está formado por un 31.11 % de controles y un 68.88 % de casos.

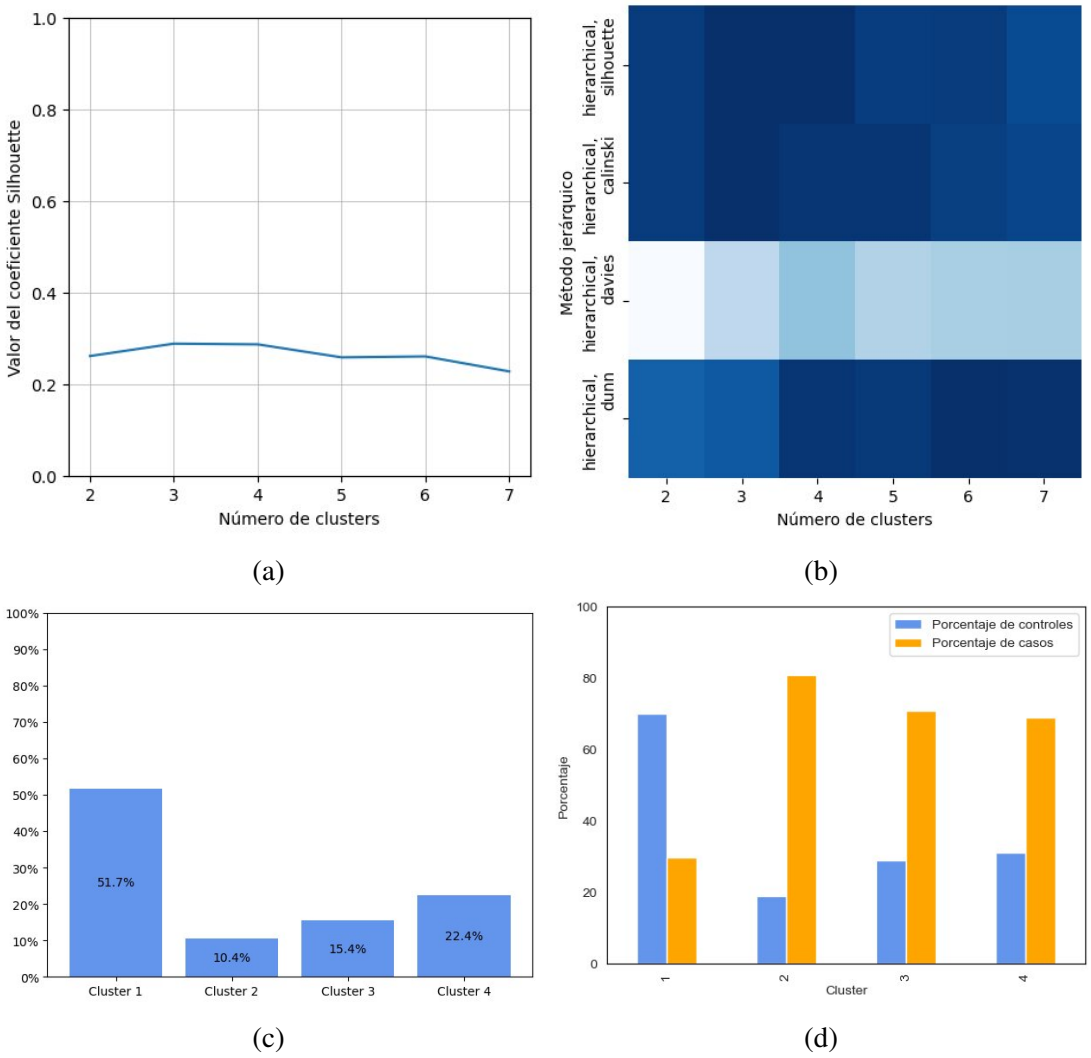


Figura 4.18: Métricas para la determinación del número óptimo de *clusters* y distribución de los pacientes en cada *cluster* para bases concatenadas. (a) Coeficiente *Silhouette*; (b) Otras métricas para la validación del número de *clusters* - Nota: El número óptimo de *clusters* para cada método viene dado por el azul más intenso; (c) Porcentaje de pacientes correspondiente a cada *cluster*; (d) Distribución de casos y controles en cada *cluster*.

La Figura 4.19 proporciona un apoyo en el análisis de resultados y en la realización de comparaciones, ya que muestra los perfiles de los subconjuntos identificados. Estos perfiles permiten una comprensión más profunda de las similitudes y diferencias entre los *clusters* formados por el algoritmo. Al examinar los perfiles, se pueden identificar características y patrones distintivos para la interpretación de las agrupaciones. Por ejemplo, se observa una diferenciación en la variable *DroppedActInt* que hace referencia al abandono de las actividades e intereses por parte de los pacientes. Se puede ver que los pacientes pertenecientes al tercer *cluster* son aquellos que

han contestado de manera positiva a esta pregunta, lo que sugiere una relación con la variable *FullEnergy* que indica el nivel de energía de los pacientes. Al analizar esta última variable, se llega a la conclusión de que los pacientes del tercer *cluster* han respondido de manera negativa en cuanto a su nivel de energía. Al unificar la información de ambas variables se puede deducir que los pacientes pertenecientes al tercer *cluster* sienten que su nivel de energía es bajo y en consecuencia, tienden a abandonar gran parte de sus actividades e intereses. Este ejemplo ilustra cómo este análisis permite relacionar variables de manera eficiente, incluso cuando se trabaja con una gran cantidad de ellas.

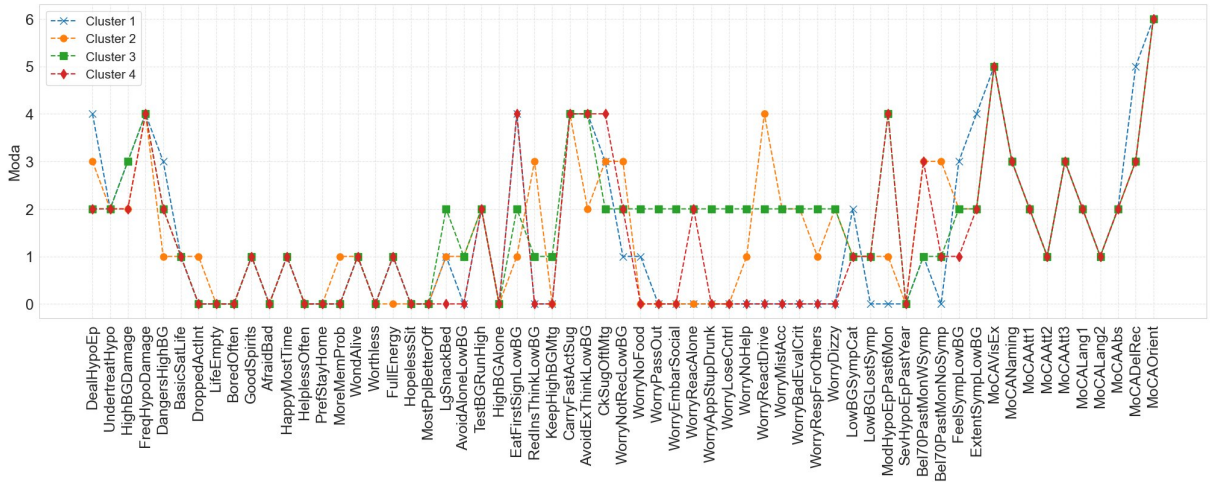


Figura 4.19: Perfiles de los *clusters* de las bases de datos concatenadas.

Capítulo 5

Conclusiones y líneas futuras

En este capítulo se presentan las conclusiones obtenidas tras la realización del presente TFG y además se incluyen varias líneas futuras de trabajo identificadas.

5.1. Conclusiones

En este TFG se ha realizado un estudio sobre la incidencia de los episodios de hipoglucemia severa en pacientes con DMT1. Para abordar esta cuestión, se ha realizado un análisis exhaustivo de la literatura clínica y se han aplicado técnicas de ML con dos objetivos principales. En primer lugar, se ha usado el EFA para identificar patrones subyacentes y crear factores que agrupen variables con correlación entre sí. En segundo lugar, se ha empleado *clustering* para la partición automática de un grupo de pacientes en subconjuntos con características similares.

Los datos utilizados en este estudio han sido recopilados a través del *T1D Exchange* por el *Jaeb Center for Health Research*. Para llevar a cabo este trabajo, se emplearon cinco bases de datos que contenían información diversa sobre los miedos y preocupaciones experimentados por los pacientes, su estado emocional, sus capacidades cognitivas y los síntomas que presentaban en relación a los episodios de hipoglucemia.

Para comenzar con el estudio, fue necesario comprender las bases de datos y realizar un análisis descriptivo de las mismas utilizando los conceptos adquiridos y analizados durante la revisión de la literatura y el estado del arte de las técnicas de ML aplicadas en el campo de la salud. A continuación, se seleccionaron las bases de datos consideradas adecuadas y se procedió al análisis mediante su preprocesamiento, lo que permitió obtener datos limpios y de calidad con el fin de aumentar la confiabilidad de los resultados.

Después del preprocesamiento, se llevó a cabo el EFA en cada base de datos para identificar factores latentes. Este análisis permitió descubrir patrones y relaciones signitifcativas entre las variables de cada conjunto de datos de manera individual. Una vez identificados los factores latentes, se les aplicó un algoritmo de *clustering* para agrupar automáticamente a los pacientes en *clusters* según características similares identificadas por el algoritmo. Gracias a estos dos pasos, se han conseguido identificar similitudes y diferencias entre factores y entre los perfiles correspondientes a los *clusters*. Esto ha permitido tener una visión más clara y precisa de los datos y, en consecuencia, obtener información valiosa para la interpretación de resultados.

Una vez analizados los resultados obtenidos en el análisis individual de cada base de datos, se ha llevado a cabo un experimento adicional utilizando una base de datos resultado de la combinación de todas las variables previamente utilizadas. Esto ha permitido relacionar aspectos relativos a los episodios de hipoglucemia de una forma mucho más general, poniendo a prueba la capacidad de estos algoritmos para alcanzar los objetivos de este TFG con datos más heterogéneos. Como resultado de este análisis, se han conseguido identificar factores latentes que agrupan variables en función de su relación significativa, además de determinar similitudes y diferencias entre los perfiles correspondientes a cada subconjunto de pacientes. Esto ha permitido relacionar aspectos relativos a los episodios de hipoglucemia ampliando el alcance de las conclusiones obtenidas anteriormente y otorga mayor validez a los resultados.

En conclusión, se ha logrado identificar de manera efectiva patrones y similitudes en una gran cantidad de datos obtenidos a partir de múltiples encuestas. Estos resultados podrían sentar las bases para futuras investigaciones en el ámbito de la diabetes y la hipoglucemia, abriendo posibilidades para mejorar la calidad de vida de las personas que padecen esta enfermedad.

5.2. Líneas futuras

Este TFG supone una pequeña aportación al estado del arte del uso de técnicas de ML en el análisis de datos clínicos. A partir de los resultados obtenidos, se pueden plantear diversas líneas futuras de investigación con el objetivo de profundizar en el análisis de los episodios de hipoglucemia en pacientes con DMT1. Entre posibles líneas futuras se encuentran:

- Añadir al análisis otras variables como pueden ser las relacionadas con factores demográficos y socioeconómicos para evaluar su influencia en el desarrollo de la hipoglucemia.
- Involucrar a pacientes procedentes de diferentes zonas geográficas para evaluar la influencia de los factores culturales en la aparición de los episodios de hipoglucemia.

- Añadir cuestionarios más detallados y orientados al estilo de vida de los pacientes, que evalúen aspectos como la calidad del sueño, la actividad física o el nivel de estrés. Con esto se podría determinar si existen relaciones significativas entre estos aspectos y la aparición de los episodios de hipoglucemia.
- Explorar otras técnicas de ML para identificar patrones y relaciones entre variables que no hayan sido detectadas con las herramientas utilizadas.

Bibliografía

[1] Sandra Páez Ramos. Relación lógica entre el cuidado y las enfermedades crónicas. *Revista de Enfermería Ene*, 13(4), 2019.

[2] Rocio Robledo Martínez and Fabio Alberto Escobar Díaz. Las enfermedades crónicas no transmisibles en colombia. *Boletín del Observatorio en Salud*, 3(4), 2010.

[3] Enrique Ardila. Las enfermedades crónicas. *Revista del Instituto Nacional de Salud Biomédica*, 38(1):5–6, 2018.

[4] Elizabeth Rodríguez, Rusty Molina, and Cruz Rodríguez. Definición, clasificación y diagnóstico de la diabetes mellitus. *Revista Venezolana de Endocrinología y Metabolismo*, 10(1):7–12, 2012.

[5] Alejandro Almaguer Herrera, Pedro Enrique Miguel Soca, Carlos Reynaldo Será, Antonio Luis Mariño Soler, and René Carlos Oliveros Guerra. Actualización sobre diabetes mellitus. *Correo Científico Médico*, 16(2), 2012.

[6] Manuel Emiliano Licea Puig and Teresa Marqarita González Calero. Estrategias para la prevención de la diabetes mellitus tipo 1. *Revista Cubana de Salud Pública*, 39:733–751, 2013.

[7] Néboa Zozaya, Renata Villoro, Álvaro Hidalgo, Juan Oliva, Marta Rubio, Sonia Garcia-Perez, et al. Estudios de coste de la diabetes tipo 2: una revisión de la literatura. 2015.

[8] Edelmiro Luis Menéndez Torre, Jessica Ares Blanco, Santiago Conde Barreiro, Gemma Rojo Martínez, Elías Delgado Alvarez, et al. Prevalencia de diabetes mellitus en 2016 en españa según la base de datos clínicos de atención primaria (bdcap). *Endocrinología, Diabetes y Nutrición*, 68(2):109–115, 2021.

[9] Statista. Estadísticas de la diabetes, 2023. Available at https://es.statista.com/estadisticas/702527/

gasto-sanitario-en-personas-con-diabetes-a-nivel-mundia/#:~:text=En%20el%20a%C3%B1o%202021%2C%20se,960.000%20millones%20de%20d%C3%B3lares%20estadounidenses.

[10] Roberto Milton Di LorenziBruzzone, Lorena Bruno, Marcelo Pandolfi, Gerardo Javiel, and Mabel Goñi. Hipoglucemia en pacientes diabéticos. *Revista Uruguaya de Medicina Interna*, 2(3):51–60, 2017.

[11] Pedro J Pinés Corrales, Cristina Arias Lozano, Cortes Jiménez Martínez, Luz M López Jiménez, Alejandro E Sirvent Segovia, Lourdes García Blasco, and Francisco Botella Romero. Prevalencia de hipoglucemia grave en una cohorte de pacientes con diabetes tipo 1. *Endocrinología, Diabetes y Nutrición*, 68(1):47–52, 2021.

[12] Javier Mora Pineda. Modelos predictivos en salud basados en aprendizaje de maquina (machine learning). *Revista Médica Clínica Las Condes*, 33(6):583–590, 2022.

[13] Gloria López. Diabetes mellitus: clasificación, fisiopatología y diagnóstico. *Medwave*, 9(12), 2009.

[14] María P Russo, María F Grande-Ratti, Mariana A Burgos, Anahí A Molaro, and María B Bonella. Prevalencia de diabetes, características epidemiológicas y complicaciones vasculares. *Archivos de Cardiología de México*, 93(1):30–36, 2023.

[15] World Health Organization et al. Classification of diabetes mellitus. 2019.

[16] OMS. Datos de la diabetes, 2023. Available at https://www.who.int/es/news-room/fact-sheets/detail/diabetes.

[17] Eduardo Cabrera Rode, Pedro A Perich Amador, and Manuel E Licea Puig. Diabetes autoinmune latente del adulto o diabetes tipo 1 de lenta progresión: definición, patogenia, clínica, diagnóstico y tratamiento. *Revista Cubana de Endocrinología*, 13(1):0–0, 2002.

[18] Yuri Arnold Domínguez, Manuel E Licea Puig, and José Hernández Rodríguez. Algunos apuntes sobre la epidemiología de la diabetes mellitus tipo 1. *Revista Cubana de Salud Pública*, 44:e1127, 2018.

[19] Sociedad Española de Diabetes. Estrategia DM tipo I, 2023. Available at https://www.revistadiabetes.org/tratamiento/diabetes-tipo-1/estrategias-de-prediccion-y-prevencion-en-la-diabetes-mellitus-tipo-1/.

[20] FDA. Medicamento DM tipo I, 2023. Available at https://www.fda.gov/news-events/press-announcements/la-fda-aprueba-el-primer-medicamento-que-puede-retrasar-la-aparicion-de-la-diab~:text=Hoy%2C%20la%20Administraci%C3%B3n%20de%20Alimentos,tipo%201%20en%20estadio%202.

[21] Diana Elizabeth Pérez Guiracocha. Revisión bibliográfica del temaïnmunogenética de la diabetes mellitus tipo 1". 2020.

[22] Philip E Cryer. *Hypoglycemia: pathophysiology, diagnosis, and treatment*. Oxford University Press, USA, 1997.

[23] James B Field. Hypoglycemia: definition, clinical presentations, classification, and laboratory tests. *Endocrinology and Metabolism Clinics of North America*, 18(1):27–43, 1989.

[24] C Colas. Les hypoglycémies, un sujet de préoccupation majeur pour les personnes atteintes de diabète et leur entourage: résultats français de dawn2™. *Médecine des Maladies Métaboliques*, 7:S30–S33, 2013.

[25] David S Oyer. The science of hypoglycemia in patients with diabetes. *Current Diabetes Reviews*, 9(3):195–208, 2013.

[26] JJ Alfaro Martínez, I Mora Escudero, I Huguet Moreno, and C Gonzalvo Díaz. Hipoglucemia. *Medicine-Programa de Formación Médica Continuada Acreditado*, 11(18):1089–1095, 2012.

[27] PHILIP E Cryer. Glucose counterregulation: prevention and correction of hypoglycemia in humans. *American Journal of Physiology-Endocrinology and Metabolism*, 264(2):E149–E155, 1993.

[28] Pamela Apablaza, Néstor Soto, and Ethel Codner. De la bomba de insulina y el monitoreo continuo de glucosa al páncreas artificial. *Revista Médica de Chile*, 145(5):630–640, 2017.

[29] Jorge Bondía. Páncreas artificial. *Revista Esp Endocrinol Pediatr*, 11(1):8–13, 2020.

[30] Marianella Álvarez Vega, Laura María Quirós Mora, Mónica Valeria Cortés Badilla, et al. Inteligencia artificial y aprendizaje automático en medicina. *Revista médica sinergia*, 5(8):e557–e557, 2020.

[31] Humberto Chaviano Arteaga. Técnicas de aprendizaje supervisado y no supervisado para el aprendizaje automatizado de computadoras. In *Memorias del primer Congreso Internacional de Ciencias Pedagógicas: Por una educación integral, participativa e incluyente*, pages 549–564. Instituto Superior Tecnológico Bolivariano, 2015.

[32] APD Redacción. ¿ cuáles son los tipos de algoritmos del machine learning, 2019.

[33] Joseph R Dettori and Daniel C Norvell. The anatomy of data. *Global spine journal*, 8(3):311–313, 2018.

[34] SGOPAL Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[35] Claudia Hernández, Jorge Enrique Rodríguez Rodríguez, et al. Preprocesamiento de datos estructurados. *Revista vínculos*, 4(2):27–48, 2008.

[36] Leandre R Fabrigar and Duane T Wegener. *Exploratory factor analysis*. Oxford University Press, 2011.

[37] Daire Hooper. Exploratory factor analysis. 2012.

[38] Parul M Jain and VK Shandliya. A survey paper on comparative study between principal component analysis (pca) and exploratory factor analysis (efa). *International Journal of Managment, IT and Engineering*, 3(6):415–424, 2013.

[39] Jean Dean Brown. Principal components analysis and exploratory factor analysis &ndash definitions, differences, and choices. *Statistics*, 13(1):26–30, 2009.

[40] Carolina Méndez Martínez and Martín Alonso Rondón Sepúlveda. Introducción al análisis factorial exploratorio. *Revista Colombiana de Psiquiatría*, 41(1):197–207, 2012.

[41] Kelvin Pizarro Romero and Omar Martínez Mora. Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral kmo y esfericidad de bartlett para determinar factores principales. *Journal of Science and Research*, 5(CININGEC):903–924, 2020.

[42] Emilio Cabello and Jesús L Chirinos. Validación y aplicabilidad de encuestas servqual modificadas para medir la satisfacción de usuarios externos en servicios de salud. *Revista Médica Herediana*, 23(2):88–95, 2012.

[43] Susana Lloret-Segura et al. El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3):1151–1169, 2014.

[44] Andrés Fernández Aráuz. Aplicación del análisis factorial confirmatorio a un modelo de medición del rendimiento académico en lectura. *Revista de Ciencias Económicas*, 33(2):39–65, 2015.

[45] J Brown. Choosing the right number of components or factors in pca and efa. *JALT Testing & Evaluation SIG Newsletter*, 13(2), 2009.

[46] Cadeyrn J Gaskin and Brenda Happell. On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, 51(3):511–521, 2014.

[47] María Luisa Garmendia. Análisis factorial: una aplicación en el cuestionario de salud general de goldberg, versión de 12 preguntas. *Revista Chilena de Salud Pública*, 11(2):57–65, 2007.

[48] Murat Yıldırım and Abdurrahim Güler. Factor analysis of the covid-19 perceived risk scale: A preliminary study. *Death Studies*, 46(5):1065–1072, 2022.

[49] Seiichi Omura, Kazuaki Shimizu, Motoi Kuwahara, Miyuki Morikawa-Urase, Susumu Kusunoki, and Ikuo Tsunoda. Exploratory factor analysis determines latent factors in guillain–barré syndrome. *Scientific Reports*, 12(1):21837, 2022.

[50] Hayfa Almutary, Clint Douglas, and Ann Bonner. Multidimensional symptom clusters: an exploratory factor analysis in advanced chronic kidney disease. *Journal of Advanced Nursing*, 72(10):2389–2400, 2016.

[51] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, and Tutut Herawan. Big data clustering: a review. In *Computational Science and Its Applications–ICCSA 2014: 14th International Conference, June 30–July 3, 2014*, pages 707–720, Guimarães, Portugal, 2014. Springer.

[52] Enrique Ruspini. A new approach to clustering. *Information and Control*, 15(1):22–32, 1969.

[53] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[54] Gladys M Casas Cardoso, Gladys Cardoso Romero, Vivian Guerra Morales, and Luis Felipe Herrera Jiménez. Técnicas de detección de clusters aplicadas a la investigación psicológica. *Revista Cubana de Psicología*, 19(1), 2002.

[55] Yunjae Jung, Haesun Park, Ding-Zhu Du, and Barry L Drake. A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, 25(1):91–111, 2003.

[56] Duy-Tai Dinh, Tsutomu Fujinami, and Van-Nam Huynh. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *Knowledge and Systems Sciences: 20th International Symposium*, pages 1–17, Da Nang, Vietnam, December 2019. Springer.

[57] Congming Shi, Bingtao Wei, Shoulin Wei, Wen Wang, Hai Liu, and Jialei Liu. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, 2021(1):1–16, 2021.

[58] Xu Wang and Yusheng Xu. An improved index for clustering validation based on silhouette index and calinski-harabasz index. In *IOP Conference Series: Materials Science and Engineering*, volume 569, page 052024. IOP Publishing, 2019.

[59] Rabassa Gutierrez Monica. Sld263-implementacion del indice de dunn para la evaluacion de la tendencia al agrupamiento de conjuntos de datos quimioinformaticos. In *VIII Congreso Internacional de Informática en la Salud. II Congreso Moodle Salud*, 2010.

[60] Yudhistira Arie Wijaya, Dedy Achmad Kurniady, Eddy Setyanto, Wahdan Sanur Tarihoran, Dadan Rusmana, and Robbi Rahim. Davies bouldin index algorithm for optimizing clustering case studies mapping school facilities. *TEM J*, 10(3):1099–1103, 2021.

[61] Rui Veloso, Filipe Portela, Manuel Filipe Santos, Alvaro Silva, Fernando Rua, António Abelha, and José Machado. A clustering approach for predicting readmissions in intensive medicine. *Procedia Technology*, 16:1307–1316, 2014.

[62] Gustavo Lorca, José Arzola, and Osvaldo Pereira. Segmentación de imágenes médicas digitales mediante técnicas de clustering. *Revista Aporte Santiaguino*, 3:2, 2010.

[63] Ruth S Weinstock, Stephanie N DuBose, Richard M Bergenstal, Naomi S Chaytor, Christina Peterson, Beth A Olson, Medha N Munshi, Alysa JS Perrin, Kellee M Miller, Roy W Beck, et al. Risk factors associated with severe hypoglycemia in older adults with type 1 diabetes. *Diabetes Care*, 39(4):603–610, 2016.

[64] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396–403, 2015.

[65] Jerome A Yesavage. Geriatric depression scale. *Psychopharmacol Bull*, 24(4):709–711, 1988.

[66] Linda A Gonder-Frederick, Karen M Schmidt, Karen A Vajda, Megan L Greear, Harsimran Singh, Jaclyn A Shepard, and Daniel J Cox. Psychometric properties of the hypoglycemia fear survey-ii for adults with type 1 diabetes. *Diabetes Care*, 34(4):801–806, 2011.

[67] Iciar Martín-Timón and Francisco Javier del Cañizo-Gómez. Mechanisms of hypoglycemia unawareness and implications in diabetic patients. *World Journal of Diabetes*, 6(7):912, 2015.

[68] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. Montreal cognitive assessment. *The American Journal of Geriatric Psychiatry*, 2003.

[69] Trinidad Hoyl, Eduardo Valenzuela, and Pedro Paulo Marín. Depresión en el adulto mayor: evaluación preliminar de la efectividad, como instrumento de tamizaje, de la versión de 5 ítems de la escala de depresión geriátrica. *Revista Médica de Chile*, 128(11):1199–1204, 2000.

[70] Alexandre Henrique. Algoritmo EFA, 2023. Available at https://www.kaggle.com/code/alexandrehsd/binary-multiclass-classification-factor-analysis/notebook.