Segmentación y estratificación de riesgo en pacientes diagnosticados con la enfermedad de Alzheimer a partir de algoritmos de Machine Learning

Paula Andrea Zuluaga Gómez

Trabajo de grado para optar al título de Bioingeniera otorgado por UdeA

Asesor
John Fredy Ochoa Gómez, Doctor (PhD) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Bioingeniería
Medellín
2022

Cita	Zuluaga Gómez [1]
Referencia Estilo IEEE (2020)	[1] P. A. Zuluaga Gómez, “Segmentación y estratificación de riesgo en pacientes diagnosticados con la enfermedad de Alzheimer a partir de algoritmos de Machine Learning”, Práctica empresarial, Bioingeniería, Universidad de Antioquia, Medellín, 2022.







Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: John Fredy Ochoa Gómez.

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

A mis padres que, con apoyo incondicional, amor y confianza, permitieron que logre culminar mi carrera profesional. A mis hermanas, Jéssika y Catherine, por su cariño, motivación y también apoyo incondicional durante todo este proceso.

Agradecimientos

Quiero agradecer a la Universidad de Antioquia por permitirme crecer y formar como profesional. A mi asesor John Fredy Ochoa Gómez por su ayuda, dedicación y orientación en la realización de este proyecto. A Claudia Aponte por su tiempo y sobre todo por compartir sus conocimientos técnicos sobre el área. A mis profesores del programa de bioingeniería por la formación intelectual y personal. También agradezco a la empresa SURA IM por brindarme la oportunidad de realizar mis prácticas académicas en el área de cumplimiento.

TABLA DE CONTENIDO

RESUMEN 12
ABSTRACT 13
I. INTRODUCCIÓN 14
II. OBJETIVOS 16
A. Objetivo general 16
B. Objetivos específicos 16
III. MARCO TEÓRICO 17
1. Datos utilizados 17
2. Herramientas tecnológicas 17
2.1. Python 17
2.2. Google Colab 17
3. Conceptos médicos 18
3.1. Enfermedad de Alzheimer (EA) 18
3.1.1. Enfermedad de Alzheimer autosómica dominante (ADAD) 18
3.1.1.1. Mutación E280A 19
3.2. Evaluación neuropsicológica en individuos con Enfermedad de Alzheimer 19
3.2.1. Minimental State Examination (MMSE) 19
3.2.2. Fluidez verbal semántica 19
3.2.3. Denominación 19
3.2.4. Memoria de una lista de palabras (MLP) 20
3.2.5. Copia y evocación de figuras Geométricas (Praxias) 20
3.2.6. Trail Making Test (TMT) 20
3.2.7. Figura de Rey-Osterrieth 21

3.2.8. Test de clasificación de cartas de Wisconsin (WCST) 21
3.2.9. Memory Impairment Screen (MIS) 21
3.2.10. Dígitos y símbolos- WAIS 21
3.2.11. Memory capacity test (MCT) 22
3.2.12. Escala de trastornos de memoria 22
3.2.13. Escala De Depresión Geriátrica de Yesavage 22
3.2.14. Functional Assessment Staging de Reisberg (FAST) 22
3.2.15. Escala de Deterioro Global de Reisberg (EDG) 23
3.2.16. Índice de Barthel (IB) 23
3.2.17. Índice de Katz 23
3.2.18. Escala de Lawton & Brody 23
3.2.19. Cancelación visual de la A 23
3.2.20. Raven 23
3.2.21. Test de Aritmética. 24
4. Minería de datos 24
4.1. Machine Learning 24
4.1.1. Aprendizaje no supervisado 24
4.1.1.1. Modelo de clusterización K-means 24
4.1.1.2. Modelo de clusterización K-medoids 25
4.1.2. Coeficiente de Silhouette 25
4.2. Metodología CRISP-DM 26
IV. METODOLOGÍA 28
Análisis y segmentación de la población de estudio sin criterios de personal experto en el área 28
Preprocesamiento de las bases de datos independientes 29

Aplicación de la metodología CRISP-DM 29
1. Comprensión del problema 29
1.1. Problemática del estudio 30
1.2. Objetivos del proyecto 30
1.3. Evaluación de la situación 30
2. Comprensión de los datos 30
2.1. Recolección de los datos 30
2.2. Descripción de los datos 31
2.3. Exploración de los datos 31
3. Preparación de los datos 32
3.1. Selección de los datos 32
3.2. Limpieza de datos 32
3.3. Transformación de los datos 32
4. Modelado 33
4.1. Selección de la técnica de modelado 33
5. Evaluación del modelo 33
6. Implementación del modelo 33
Análisis y segmentación de la población de estudio con criterios de personal experto en el área 33
Preprocesamiento de las bases de datos independientes 33
Aplicación de la metodología CRISP-DM 34
1. Preparación de los datos 34
1.1. Limpieza de los datos 34
2. Modelado 35
2.1. Selección de la técnica de modelado 35

V. RESULTADOS Y ANÁLISIS 36
Análisis y segmentación de la población de estudio sin criterios de personal experto en el área 36
Análisis exploratorio de datos 36
Exploración de variables cuantitativas 36
Exploración de variables cualitativas 40
Exploración de variables cuantitativas y cualitativas 41
Definición de los segmentos 42
Validación interna 42
Validación externa 43
Análisis y segmentación de la población de estudio con criterios de personal experto en el área 46
Análisis exploratorio de datos 46
Exploración de variables cuantitativas 46
Exploración de variables cualitativas 54
Exploración de variables cuantitativas y cualitativas 58
Definición de los segmentos 59
Validación interna 59
Validación externa 60
Estratificación de riesgo 66
VI. CONCLUSIONES 70
REFERENCIAS 72

LISTA DE TABLAS

Tabla 1. Resumen estadístico de las variables cuantitativas – Metodología 1. 37
Tabla 2. Atributos cuantitativos del clúster 0 y 1 – Metodología 1. 44
Tabla 3. Resumen estadístico de las variables cuantitativas – Metodología 2. 47
Tabla 4. Correlaciones entre los atributos cuantitativos – Metodología 2. 52
Tabla 5. Atributos cuantitativos del clúster 0 y 1 – Metodología 2. 61

LISTA DE FIGURAS

Fig 1. Esquemático de la metodología CRISP-DM 28
Fig 2. a) Diagrama de cajas sin aumento para cada uno de los atributos, b) Diagrama de cajas con aumento para cada uno de los atributos 38
Fig 3. Matriz de correlación de variables cuantitativas 39
Fig 4. Histograma de la ocupación de los pacientes. 40
Fig 5. Relaciones relevantes entre algunas variables cualitativas. 41
Fig 6. Relaciones relevantes entre las variables cuantitativas y cualitativas. 42
Fig 7. Representación del método Silhouette – Metodología 1. 43
Fig 8. Distribución del clúster 0 y 1 – Metodología 1. 44
Fig 9. Pacientes del clúster 0 y 1 divididos por su estrato según EEPM 45
Fig 10. Proporción de los pacientes pertenecientes al clúster 0 y 1 con respecto a la variable target. 46
Fig 11. Diagrama de cajas para los atributos cuantitativos Edad, Escolaridad, IMC, Presión arterial sistólica y Presión arterial diastólica. 48
Fig 12. Diagrama de cajas para los atributos cuantitativos Depresión/Disforia, Ansiedad_y, Irritabilidad/Liabilidad emocional, Sueño/Trastornos del sueño. 50
Fig 13. Diagrama de cajas para los atributos cuantitativos MLP - Intrusiones Ensayo 1 y MLP – Intrusiones Evocación. 50
Fig 14. Diagrama de cajas para los atributos cuantitativos TMT - tiempo y Figura de rey % de retención. 51
Fig 15. Distribución de las ubicaciones de dónde provinieron los pacientes que participaron en el estudio. 54
Fig 16. Relaciones relevantes de los atributos 'Ubicacion nac.' y 'Zona'. 55
Fig 17. Atributo 'Zona' frente a las tres etapas de la enfermedad de Alzheimer. 56
Fig 18. Atributo 'Alteradas actividades instrumentales en la vida cotidiana' frente a las tres etapas de la enfermedad de Alzheimer. 56
Fig 19. Atributo 'Status' frente a las tres etapas de la enfermedad de Alzheimer. 57
Fig 20. Atributo 'EDG' frente a las tres etapas de la enfermedad de Alzheimer. 57
Fig 21. Atributo 'Sexo' frente a las tres etapas de la enfermedad de Alzheimer. 58

Fig 22. Atributo 'Zona' frente a las tres etapas de la enfermedad de Alzheimer. 58
Fig 23. Atributo 'Escolaridad' frente a las tres etapas de la enfermedad de Alzheimer. 59
Fig 24. Representación del método Silhouette – Metodología 2. 60
Fig 25. Distribución del clúster 0 y 1 – Metodología 2. 61
Fig 26. Edad promedio de inicio en la etapa sin impacto, con impacto y en la demencia de la enfermedad para los pacientes pertenecientes de los clústeres 0 y 1. 63
Fig 27. Pacientes del clúster 0 y 1 divididos por su ubicación proveniente. 64
Fig 28. Pacientes del clúster 0 y 1 divididos por su zona proveniente, rural o urbana. 64
Fig 29. Pacientes del clúster 0 y 1 divididos por Status, Alteraciones y EDG. 65
Fig 30. Clasificación de los pacientes por clúster en las tres etapas de la enfermedad. 65
Fig 31. Clasificación de los pacientes del clúster 0 y del clúster 1 en pacientes portadores y no de la mutación E280A. 67
Fig 32. Matriz de confusión para el clúster 0 y 1. 68

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

EA	Enfermedad de Alzheimer
ADAD	Enfermedad de Alzheimer Autosómica Dominante
PSEN1	Presenilin-1
PSEN2	Presenilin-2
APP	Proteína precursora de amiloide
E280A	Mutación del gen de presenilin-1 en el codón 280
DCL	Deterioro cognitivo leve
ML	Machine Learning

RESUMEN

Este proyecto tiene como finalidad desarrollar un modelo de segmentación en función de niveles de riesgo en pacientes con alto riesgo de desarrollar la enfermedad de Alzheimer, específicamente los pacientes portadores de la mutación ‘E280A’, debido a la gran importancia que existe en la actualidad la detección temprana de la enfermedad ya que permite realizar ensayos clínicos e intervenciones eficaces cuando el daño neurológico es relativamente leve en el paciente. Para el cumplimiento del estudio se hace uso de una de las metodologías más utilizadas en la minería de datos conocida como ‘CRISP-DM’. A partir de esta, se realiza el respectivo procesamiento de las bases de datos y la implementación de los modelos que permiten segmentar a la población de estudio. Finalmente, se implementa una métrica de desempeño conocida como ‘Coeficiente de Silhouette’ para evaluar la calidad del agrupamiento obtenido mediante los algoritmos de clustering.

Los resultados obtenidos por los dos modelos de segmentación diseñados dividieron a la población de estudio en dos segmentos. El segundo modelo, el cual obtuvo el mejor puntaje en la métrica de desempeño, logró diferenciar a los dos segmentos en función de niveles de riesgo. Sin embargo, en los dos segmentos definidos existen pacientes que son portadores de la mutación ‘E280A’.

Palabras clave — **Enfermedad de Alzheimer, mutación E280A, Machine Learning, metodología CRISP-DM, clustering.**

ABSTRACT

This project aims to develop a segmentation model based on risk levels in patients at high risk of developing Alzheimer's disease, specifically patients carrying the 'E280A' mutation, due to the great importance of early detection of the disease because it allows clinical trials and effective interventions when the neurological damage is relatively mild in the patient. To carry out the study, use is made of one of the most widely used methodologies in data mining known as 'CRISP-DM'. From this, the respective processing of the databases and the implementation of the models that allow segmenting the study population is performed. Finally, a performance metric known as 'Silhouette Coefficient' is implemented to evaluate the quality of the grouping obtained by the clustering algorithms.

The results obtained by the two segmentation models designed divided the study population into two segments. The second model, which obtained the best score in the performance metric, managed to differentiate the two segments based on risk levels. However, in the two defined segments there are patients who are carriers of the 'E280A' mutation.

Keywords — **Alzheimer's disease, E280A mutation, Machine Learning, CRISP-DM methodology, clustering.**

I. INTRODUCCIÓN

La enfermedad de Alzheimer (EA) es la forma más común de deterioro cognitivo incapacitante en las personas mayores y tiene un impacto social devastador. Los elementos postulados de la cascada patógena incluyen la acumulación de péptidos amiloide β (Aβ); agregación y fosforilación de tau; neuro inflamación; disfunción sináptica; y pérdida neuronal. La enfermedad de Alzheimer autosómica dominante (ADAD) representa del 1% al 2% de todos los casos de EA. Las mutaciones de los genes presenilin1 (PSEN1), presenilin2 (PSEN2) y proteína precursora de amiloide (APP) se heredan como rasgos autosómicos dominantes completamente penetrantes que generalmente provocan síntomas de EA a la edad de 65 años. Aunque existen diferencias genéticas y biológicas entre la ADAD y la EA esporádica, las consecuencias bioquímicas son similares y la acumulación de Aβ en el cerebro desempeña un papel temprano.

En 1987 se describió una familia colombiana con ADAD de inicio temprano con una mutación del gen de presenilin-1 (PSEN1) en el codón 280 (E280A). La presentación clínica más frecuente en esta familia es la pérdida gradual de la memoria, seguida de cambios en el comportamiento y deterioro del lenguaje. Basados en los resultados, se dedujo que el perfil cognitivo de la EA PSEN1 E280A no difiere sustancialmente del de la EA esporádica. Por esta razón, la ADAD representa una población crucial para identificar estrategias de prevención que podrían modificar el curso de la enfermedad en personas sin deterioro cognitivo (EA "preclínica"), pero con riesgo inminente de desarrollar síntomas debido a la enfermedad de Alzheimer (EA) [1].

Actualmente, la detección de la EA en etapas tempranas se ha convertido en uno de los principales focos de investigación en enfermedades neurodegenerativas. Su detección y diagnóstico temprano permite realizar ensayos clínicos e intervenciones eficaces cuando el daño neurológico es aun relativamente leve [2]. Los algoritmos de aprendizaje automático ofrecen un método alternativo para la detección que puede proporcionar un proceso automatizado y valioso conocimiento para el diagnóstico temprano.

En el presente estudio, el objetivo es desarrollar un modelo de Machine Learning que, mediante el uso de información sociodemográfica, evaluaciones médicas neurológicas y

evaluaciones médicas neuropsicológicas neurodegenerativas de la población de estudio sea capaz de segmentar a los pacientes, con alto riesgo de desarrollar Alzheimer por una mutación, en dos niveles de riesgo: pacientes que son portadores y pacientes que no son portadores de la mutación E280A en PS1 en etapas tempranas.

Para la clasificación se desarrollaron dos modelos basados en Machine con la diferencia de que el segundo modelo fue diseñado bajo criterios de un profesional médico. Ambos modelos dividieron a la población de estudio en dos segmentos. No obstante, el segundo modelo desarrollado obtuvo mejores resultados que el primero con un score de 0.145.

Dentro de las conclusiones, se tiene que no se verificaron las dos hipótesis, tanto el clúster 0 como el clúster 1 incluyen pacientes que son portadores de la mutación E280A en una proporción similar.

II. OBJETIVOS

A. Objetivo general

Desarrollar un modelo de segmentación en función de niveles de riesgo para la optimización del servicio de salud en pacientes diagnosticados con la enfermedad de Alzheimer poniendo en práctica la metodología y las técnicas desarrolladas en la compañía SURA Investment Management para la segmentación de sus clientes.

B. Objetivos específicos

- Establecer una metodología de minería de datos para realizar un proceso eficiente.
- Realizar un análisis exploratorio de datos con las variables que caracterizan los pacientes diagnosticados con la enfermedad de Alzheimer.
- Identificar las principales variables de segmentación con cuáles se realizará la segmentación de los pacientes.
- Construir una base de datos consolidada que permita la correcta aplicación de los algoritmos de Machine Learning.
- Validar el algoritmo de aprendizaje no supervisado que permita segmentar de manera adecuada los pacientes diagnosticados con la enfermedad de Alzheimer.

III. MARCO TEÓRICO

1. Datos utilizados

Los datos utilizados para la realización de este proyecto fueron adquiridos por medio del grupo de Neurociencias de la Universidad de Antioquia. En las bases de datos se registran variables sociodemográficas de los pacientes, así como sus resultados frente a evaluaciones médicas neurológicas y neuropsicológicas neurodegenerativas. La cantidad de registros llega a más de 6000 pacientes.

2. Herramientas tecnológicas

2.1. Python

Python es un lenguaje de programación versátil multiplataforma y multiparadigma que se destaca por su código legible y limpio. La licencia de código abierto permite su utilización en distintos contextos. Su objetivo es la automatización de procesos y se utiliza en muchas industrias y campos científicos como ciencia de datos, para extraer información valiosa de los datos, y el aprendizaje automático basados en algoritmos de Machine Learning para enseñar a las computadoras a aprender automáticamente de los datos y a efectuar predicciones precisas [3].

2.2.Google Colab

Colab es un servicio Cloud, basado en los Notebooks de Jupyter, que permite el uso gratuito de las GPUs y TPUs de Google, con librerías como: Scikit-learn, PyTorch, TensorFlow, Keras y OpenCV. Todo ello con bajo Python 2.7 y 3.6. Es una herramienta ideal para el desarrollo de aplicaciones de Machine Learning y Deep Learning, sin tener que invertir en recursos hardware o del Cloud.

Con Colab se pueden crear notebooks o importar los que han sido creados, además de compartirlos y exportarlos cuando sea necesario. Además, se puede trabajar con información o bases de datos contenidas en Google Drive, unidad de almacenamiento local, GitHub e incluso en otros sistemas de almacenamiento Cloud, como el S3 de Amazon [4].

3. Conceptos médicos

3.1. Enfermedad de Alzheimer (EA)

La enfermedad de Alzheimer (EA) es la enfermedad neurodegenerativa más prevalente de inicio en la edad adulta, caracterizada por un deterioro progresivo de la cognición y la memoria [5]. Está descrita por 3 estadios: la etapa preclínica, en el que inicia la acumulación atípica de β-amiloide sin manifestaciones clínicas; el deterioro cognitivo leve (DCL), que puede ser de tipo amnésico y no amnésico; y, por último, la etapa de demencia, en la que se identifican alteración de la memoria y falta de independencia del individuo para realizar tareas de la vida diaria. Las características patológicas más importantes de la EA abarcan la acumulación excesiva de placas de β-amiloide extracelular y la presencia de ovillos neurofibrilares de proteína tau hiperfosforilada intracelulares, que da lugar a una configuración atípica que se extiende por toda la corteza cerebral [6].

Se ha establecido que esta enfermedad comienza con la muerte neuronal en la corteza entorrinal, una región que está conectada con el hipocampo, que juega un papel importante en el aprendizaje y está involucrada en la transformación de recuerdos a corto plazo en recuerdos a largo plazo. La atrofia de estas áreas cerebrales explica los síntomas de olvido observados en las primeras etapas de la enfermedad, pero también están presentes otras alteraciones cognitivas, como cambios en la atención y la capacidad de resolución de problemas [5].

*3.1.1. Enfermedad de Alzheimer autosómica dominante (ADA*D)

La enfermedad de Alzheimer autosómica dominante (ADAD); también conocida como EA familiar o EA de herencia dominante, es una forma hereditaria de la enfermedad del Alzheimer causada por mutaciones de un solo gen en los genes APP, PSEN1 o PSEN2. Se estima que menos del 1% de todos los casos de Alzheimer del mundo está causado por mutaciones genéticas. Suele tener un inicio muy precoz y los síntomas aparecen frecuentemente antes de los 60 años, pudiendo incluso llegar a manifestarse en la década de los 30 años. Si una persona tiene una de estas mutaciones es muy probable que desarrolle Alzheimer y existe un 50% de probabilidad de transmitirla a sus hijos [7].

3.1.1.1. Mutación E280A

Se han identificado familias con la mutación de PSEN1 en el codón 280 (E280A) que viven principalmente en Antioquia, Colombia. La presentación clínica más frecuente en esta familia es la pérdida gradual de la memoria, seguida de cambios en el comportamiento y deterioro del lenguaje. El perfil cognitivo de la EA PSEN1 E280A no difiere sustancialmente del de la EA esporádica. La edad de inicio del depósito de Aβ fibrilar fue de 28 años, con un patrón de depósito similar al observado en la EA esporádica. La mediana de edad de aparición fue de 44 años para deterioro cognitivo leve (DCL) y de 49 años para demencia [1].

3.2. Evaluación neuropsicológica en individuos con Enfermedad de Alzheimer

Las evaluaciones neuropsicológicas tienen como objetivo determinar el estado de salud de la persona que padece probablemente la Enfermedad de Alzheimer. Las escalas de evaluación incluyen ítems de atención, memoria, percepción visual, praxias, lenguaje, cálculo, función ejecutiva y nivel de independencia para actividades cotidianas. Con estas evaluaciones se puede determinar de manera breve, rápida y general, si la persona tiene alteraciones anormales en sus funciones cognitivas a partir de un puntaje total.

3.2.1. Minimental State Examination (MMSE)

Instrumento que sirve para detectar y evaluar la progresión del trastorno cognitivo asociado a enfermedades neurodegenerativas. Explora 5 áreas cognitivas: orientación, fijación, concentración y cálculo, memoria y lenguaje. Su puntuación total máxima es 30 [8].

3.2.2. Fluidez verbal semántica

Los tests de fluidez verbal semántica (FVS) miden el número de elementos de una misma categoría (animales, frutas, herramientas, etc.) que un sujeto puede evocar en un tiempo determinado [9].

3.2.3. Denominación

El test consiste en presentar una serie de láminas con figuras impresas las cuales deben ser nombradas por los sujetos. La denominación está sujeta a 15 figuras diferentes agrupadas en grupos de cinco según grado de familiaridad con la palabra [9].

3.2.4. Memoria de una lista de palabras (MLP)

El test MLP mide el aprendizaje de un listado de 10 palabras que se presentan con distinto orden en tres ensayos consecutivos. Mide el volumen de memoria, que involucra tanto retención a corto plazo como capacidades de aprendizaje. Este test está dividido en tres secciones [9]:

- **Recuerdo inmediato de una lista de palabras:** Consiste en el recuerdo inmediato de una lista de 10 palabras independientes presentadas por escrito en tres ensayos consecutivos. En el primer ensayo, se pide al sujeto que las lea en voz alta y cuando haya acabado de leerlas debe decir de memoria todas las que recuerde. Se realizan otros dos ensayos consecutivos siguiendo el mismo procedimiento. Se registra el número total de palabras recordadas en cada ensayo, siendo la puntuación total la suma de las palabras recordadas en los tres ensayos.
- **Recuerdo diferido de una lista de palabras:** Mide el recuerdo libre de la lista de las 10 palabras, transcurrido un tiempo de entre 5 y 10 minutos tras la interferencia de tarea distractora para evitar mecanismos de refuerzo de la memoria. Se registra el número total de palabras correctamente recordadas en este intervalo.
- **Reconocimiento de lista de palabras:** Valora la capacidad del sujeto para reconocer las 10 palabras de la lista mostradas previamente junto a otras 10 palabras distractoras. Las palabras van presentándose una a una y el sujeto debe indicar si pertenecen o no a la lista previa, contestando "sí" cuando se trate de las palabras objetivo y "no" cuando se trate de palabras distractoras.

3.2.5. Copia y evocación de figuras Geométricas (Praxias)

El test evalúa habilidades construccionales y memoria no verbal. El sujeto debe dibujar con lápiz y papel cuatro figuras geométricas (círculo, rombo, rectángulo y cubo). Posterior al recuerdo diferido, se pide al participante que evoque las figuras geométricas previamente escritas. El puntaje máximo es 11 [9].

3.2.6. Trail Making Test (TMT)

El test Making Trail part A se utiliza para evaluar la velocidad de procesamiento, atención, escaneo visual y secuenciación. El sujeto debe unir, mediante en una línea consecutiva (sin levantar el lápiz), los números del 1 al 25 lo más rápido posible. El puntaje máximo es 25 [9].

3.2.7. Figura de Rey-Osterrieth

La prueba consiste en pedirle al sujeto que realice dos tipos de tareas:

1. Debe copiar, con el modelo a la vista, el dibujo que el instructor le presente (Copia).
2. Al transcurrir un tiempo, se le solicita que reproduzca el dibujo sin tener el modelo a la vista (Evocación).

Este test mide organización perceptual y memoria visual. El puntaje máximo es 36 [9].

3.2.8. Test de clasificación de cartas de Wisconsin (WCST)

La prueba de clasificación de tarjetas de Wisconsin (WCST) se utiliza para evaluar las funciones ejecutivas de nivel superior o el cambio. El WCST clásico consta de 64 cartas; las cartas están compuestas por la combinación de tres clases de atributos: la forma (triángulo, estrella, cruz y círculo), el color (rojo, azul, verde y amarillo) y el número (uno, dos, tres o cuatro elementos). La tarea consiste en repartir las cartas siguiendo un criterio específico, por ejemplo, el color. Cuando el sujeto realiza diez respuestas correctas consecutivas, consigue una categoría, y a partir de la última se cambia el criterio de clasificación sin previa advertencia. Si continúa clasificando las cartas con el criterio de la categoría anterior, va puntuando en errores perseverativos [10].

3.2.9. Memory Impairment Screen (MIS)

Es una prueba breve de trastornos de memoria que utiliza técnicas de aprendizaje controlado y recuerdo selectivamente facilitado para optimizar los procesos de codificación. Consta de cuatro palabras que deben ser leídas e identificadas con la categoría a la que corresponda. Se prosigue con una tarea de interferencia. Finalmente, el sujeto debe recordar las cuatro palabras aprendidas previamente en cualquier orden (recuerdo libre). En caso de no hacerlo, se emplean las claves semánticas (categorías) que se emplearon en el aprendizaje, con el objetivo de facilitar el recuerdo de los ítems que no se lograron evocar libremente [8]. El puntaje máximo es 8 [9].

3.2.10. Dígitos y símbolos- WAIS

El test de WAIS mide aspectos de velocidad visuomotora. En la parte superior de la hoja se presenta un modelo en el que, a cada número, debajo del mismo, le corresponde un símbolo

distintivo. El sujeto debe completar en unas casillas vacías el número equivalente al símbolo presentado [9].

3.2.11. Memory capacity test (MCT)

Es la versión extensa del MIS. Emplea técnicas de aprendizaje controlado y recuerdo selectivamente facilitado para optimizar los procesos de codificación. Consta de dos listas con 16 palabras cada una, las cuales se presentan en cuatro láminas de cuatro palabras cada una. Cada palabra hace parte de una categoría específica, la cual el sujeto debe asociar [8]. El puntaje máximo es 32 [9].

3.2.12. Escala de trastornos de memoria

La prueba permite cuantificar el funcionamiento de la memoria cotidiana del paciente. La escala se realiza al sujeto y a un familiar. Consta de 15 preguntas cada uno con 4 opciones de respuesta así: nunca (0) rara vez (1) a veces (2) y casi siempre (3). El sujeto debe establecer con qué frecuencia ocurre cada una de las 15 preguntas formuladas. 17 es el punto de corte con máxima puntuación de 45. La encuesta contestada por el paciente se denomina QP y QF la contestada por su familiar [8].

3.2.13. Escala De Depresión Geriátrica de Yesavage

Escala auto aplicable diseñada para adultos mayores, compuesta por 15 ítems (10 síntomas positivos y 5 negativos). Según la calificación de la prueba, una puntuación entre 0 y 5 puntos, indica normalidad, entre 6 y 9 puntos, indica depresión leve y 10 o más puntos indican depresión establecida [11].

3.2.14. Functional Assessment Staging de Reisberg (FAST)

Escala que permite valorar los niveles de funcionalidad del paciente en la vida diaria. Sirve para evaluar la duración de la enfermedad en cada uno de los estados y para determinar el orden de pérdida funcional de los pacientes hasta llegar a la demencia. La puntuación de la escala corresponde al ítem en el cual se encuentra el paciente: 1, 2, 3, 4, etc. hasta 16 [8].

3.2.15. Escala de Deterioro Global de Reisberg (EDG)

Escala que permite valorar los niveles de funcionalidad del paciente en la vida diaria. Va desde el nivel 1 (normal, no demente) hasta el nivel 7 (demencia tardía) [8].

3.2.16. Índice de Barthel (IB)

Es un índice utilizado para medir el nivel de independencia del sujeto en actividades de la vida diaria. La IB evalúa 10 áreas de la vida diaria que son: comer, trasladarse entre la silla y la cama, aseo personal, uso del baño, bañarse, desplazarse, subir y bajar escaleras, vestirse y desvestirse, control del intestino y control de orina. La prueba se puntúa sobre 100, donde 0 indica total dependencia en las actividades de la vida diaria, en tanto que 100 denota un sujeto autónomo, capaz de realizar sus actividades de manera independiente [11].

3.2.17. Índice de Katz

Índice que permite el desarrollo de las actividades básicas de la vida diaria del sujeto [12].

3.2.18. Escala de Lawton & Brody

Escala que permite evaluar el desarrollo de las actividades cotidianas instrumentales complejos del sujeto [12].

3.2.19. Cancelación visual de la A

Es una prueba que sirve para evaluar procesos atencionales, especialmente la atención sostenida. Consiste en una serie de letras (más de 60) que suelen distribuirse de manera aleatoria en una hoja en blanco, de forma que el sujeto tiene que tachar en el menor tiempo posible todas las letras A que encuentre. Se puntúan las respuestas correctas, los errores y el tiempo. (no tachar la letra indica un error de omisión) [13].

3.2.20. Raven

Es una tarea sensible a daño frontal y puede reflejar también funcionamiento ejecutivo en cuanto a resolución de problemas. La tarea incluye 12 láminas. La instrucción es completar diseños incompletos; el sujeto seleccionará de seis opciones posibles, aquella que complete mejor el diseño

modelo. Las láminas aumentan su grado de dificultad conforme se avanza en la tarea. El puntaje máximo es 12 [9].

3.2.21. Test de Aritmética.

En el test, al sujeto evaluado se le presenta una serie de problemas aritméticos en palabras para que los resuelva mentalmente, sin el uso de lápiz y papel, dentro de un tiempo límite cada uno de ellos. (15 a 60 segundos) [14].

4. Minería de datos

4.1. Machine Learning

Es una derivación de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones para realizar tareas específicas y con mínima intervención humana. Además, son algoritmos capaces de mejorar con el tiempo. Existen dos tipos principales de Machine Learning: aprendizaje supervisado y aprendizaje no supervisado. Este proyecto se centrará en aprendizaje no supervisado [15].

4.1.1. Aprendizaje no supervisado

Es un tipo de aprendizaje donde no se usan valores verdaderos o etiquetas, los sistemas tienen como finalidad la comprensión y abstracción de patrones de información de manera directa. Este modelo de problema se conoce como clustering. Su análisis por agrupamiento pretende realizar la clasificación en clústeres para que las observaciones en cada grupo sean similares entre sí bajo ciertas características. Teóricamente, los puntos de datos que están en el mismo grupo deben tener características similares, mientras que los puntos de datos en diferentes grupos deben tener características diferentes. Dentro de los algoritmos de clasificación no supervisados se tiene al K-means y al K-medoids, ambos algoritmos de partición [16].

4.1.1.1. Modelo de clusterización K-means

K-means es un algoritmo de agrupamiento no supervisado el cual calcula los centroides para cada grupo y luego calcula las distancias entre el vector de datos actual y cada uno de los centroides, asignando el vector al grupo de datos cuyo centroide sea el más cercano a este. El agrupamiento

con K-means es llevado a cabo en términos de distancia, es decir, divide el conjunto de datos de estudio en un número determinado de clústeres, acorde a una función de distancia establecida [17].

Los pasos que realiza el algoritmo K-means se basan en un proceso iterativo como el siguiente:

1. Seleccionar los (k) puntos como centroides iniciales.
2. Formar clústeres (k) asignando cada punto a su centroide más cercano.
3. Recalcular el centroide de cada clúster.
4. Repetir.
5. Hasta que el criterio de convergencia sea conocido.

No obstante, el algoritmo realiza una iteración que está acompañada de una amplia variedad de medidas de distancia para calcular los centroides más cercanos. La elección de estas medidas puede afectar de forma significativa la asignación del centroide y la calidad de la solución final.

4.1.1.2. Modelo de clusterización K-medoids

K-medoids es un algoritmo flexible a los valores atípicos y es más robusto en comparación con K-means, pero la complejidad computacional de K-medoids es más alta y por lo tanto no es adecuado para grandes conjuntos de datos [17]. Los algoritmos basados en el método de K-medoids tienen el propósito de dividir un conjunto de observaciones en grupos, teniendo como principal diferencia con el algoritmo K-means la utilización de los datos que forman parte del conjunto de datos a ser analizados como representantes de las agrupaciones, que son denominados "medoids". Cada observación restante es agrupada con el medoid más cercano. Estos algoritmos tienen características más robustas ante el ruido que se puede dar en las observaciones, siendo uno de los más representativos el algoritmo Partición Alrededor de Medoids (PAM) [16].

4.1.2. Coeficiente de Silhouette

El coeficiente de Silueta es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering. El análisis de silueta se puede utilizar para estudiar la distancia de separación entre los grupos resultantes. Muestra una medida de qué tan cerca está cada punto en un grupo de puntos en los grupos vecinos y, por lo tanto, proporciona una forma de evaluar parámetros como el número de grupos visualmente. Esta medida tiene un rango de [-1, 1]. Valores

cercanos a +1 indican que la muestra está lejos de los conglomerados vecinos. Un valor de 0 indica que la muestra se encuentra en el límite de decisión entre dos conglomerados vecinos o muy cerca del mismo, y los valores negativos indican que es posible que esas muestras se hayan asignado al conglomerado equivocado [18].

4.2. Metodología CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) integra todas las tareas necesarias en los proyectos de minería de datos o ciencia de datos, desde la fase de comprensión del problema hasta la puesta en producción de sistemas automatizados analíticos o predictivos. Además, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas [19].

CRISP-DM conceptualiza el proceso de minería de datos en 6 fases [20]:

- **Comprensión del negocio:** La primera fase comprende en entender de la manera más completa el problema que se desea resolver, esto permite recolectar los datos correctos e interpretar correctamente los resultados.
- **Comprensión de los datos:** La segunda fase comprende la recolección inicial de datos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis.
- **Preparación de datos:** En la tercera fase se seleccionan los datos que se transforman de acuerdo con los resultados de la etapa anterior a fin de utilizarlos en la etapa de modelado. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.
- **Modelado:** En la cuarta etapa se obtiene propiamente el modelo de minería de datos.
- **Evaluación del modelo:** En la quinta fase se determina la calidad del modelo con base en el análisis de ciertas métricas estadísticas del mismo, comparando los resultados con resultados previos, o bien, analizando los resultados con apoyo de expertos en el dominio del problema.

- **Implementación del modelo:** En la sexta y última fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio.

IV. METODOLOGÍA

La metodología CRISP-DM de minería de datos orientada a la segmentación de pacientes diagnosticados con la enfermedad de Alzheimer consiste en la aplicación ordenada y sistemática de las siguientes etapas como se presenta en la Fig 1:

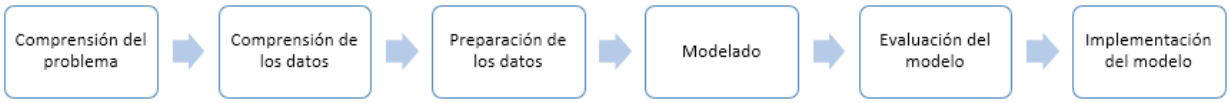

Fig 1. Esquemático de la metodología CRISP-DM

La aplicación de esta metodología se ve reflejada en dos sesiones principales:

1. Análisis y segmentación de la población de estudio sin criterios de personal experto en el área.
2. Análisis y segmentación de la población de estudio con criterios de personal experto en el área.

A continuación, se describe la metodología llevada a cabo para cada sesión mencionada.

Análisis y segmentación de la población de estudio sin criterios de personal experto en el área

La información utilizada para la realización del proyecto fue adquirida por medio de grupo de Neurociencias de la Universidad de Antioquia. No obstante, dicha información se encontraba distribuida en cuatro bases de datos diferentes:

1. Datos generales del paciente.
2. Evaluaciones médicas neurológicas.
3. Evaluaciones neuropsicológicas neurodegenerativas.
4. Genotipificaciones.

Previa a la aplicación de la metodología CRISP-DM, fue necesario realizar un preprocesamiento de las cuatro bases de datos con la finalidad de obtener una sola base de datos que abarcara toda la información necesaria para el proyecto.

Preprocesamiento de las bases de datos independientes

Para el preprocesamiento de las cuatro bases de datos se consideraron los siguientes pasos:

a) De la base de datos ‘Genotipificaciones’ se seleccionaron todos los registros de pacientes que presentaron un valor diferente de nulo con respecto al atributo ‘E280A’ y se eliminaron los que no cumplieron con la condición.
b) Se filtraron los registros de pacientes seleccionados en la base de datos ‘Datos generales del paciente’ por medio del atributo ‘Blind Code’.
c) Para las hojas de datos ‘Examen médico’,’ Funciones mentales superiores’,’ Escala CDR’, ‘Dx. neurol y psiq’, ‘Escala Cummings’, contenidas en la base de datos ‘Evaluaciones médicas neurológicas’, se creó un nuevo atributo nombrado ‘ID’ el cual fue generado por la concatenación de los atributos ‘No. Evaluación’ y ‘Blind Code’.
d) De la base de datos ‘Evaluaciones neuropsicológicas neurodegenerativas’ se creó, de igual manera, el atributo nombrado ‘ID’ generado por la concatenación de los atributos ‘No. Evaluación’ y ‘Blind Code’.
e) Desde la hoja de datos ‘Examen médico’ se seleccionaron los registros de pacientes estándar para identificar los ‘ID’ coincidentes de estos registros en las hojas datos ‘Examen médico’,’ Funciones mentales superiores’,’ Escala CDR’, ‘Dx. neurol y psiq’, ‘Escala Cummings’ y en la base de datos ‘Evaluaciones neuropsicológicas neurodegenerativas’.
f) De los registros que fueron coincidentes, se eliminaron los valores duplicados.
g) Se creó la base de datos global que contiene todos los atributos necesarios con los registros de pacientes seleccionados uniendo las bases de datos independientes con sus respectivas hojas después de haber realizado los filtros mencionados y eliminando el atributo de comparación ‘E280A’ de la base de datos ‘Genotipificaciones’.

Aplicación de la metodología CRISP-DM

1. *Comprensión del problema*

La primera fase consistió en entender la problemática, puntualizar los objetivos del proyecto de minería de datos y evaluar la situación actual.

1.1.Problemática del estudio

La enfermedad de Alzheimer es una enfermedad neurodegenerativa cerebral progresiva e irreversible la cual puede ser causada por la mutación E280A en el gen de presenilina 1. Esta mutación se encuentra asociada al grupo familiar más grande del mundo con la enfermedad [21]. La presencia de este gen mutante determina por completo que el paciente padezca la dolencia. En la actualidad no hay tratamientos efectivos para curar dicha enfermedad.

1.2.Objetivos del proyecto

El objetivo principal de este trabajo consistió en obtener un modelo de segmentación en función de niveles de riesgo en pacientes diagnosticados con la enfermedad de Alzheimer por medio de una base de datos que recopila información acerca de atributos que caracterizan a los mismos pacientes.

1.3.Evaluación de la situación

Se contó con cuatro bases de datos con información detallada de los pacientes y de evaluaciones médicas neurológicas y neuropsicológicas degenerativas que se les realizaron; y con un entorno de programación donde se realizó el procesamiento de estas bases de datos.

2. Comprensión de los datos

En la segunda fase se realizó la recolección, descripción y exploración de los datos por medio de métodos estadísticos básicos.

2.1.Recolección de los datos

Las bases de datos utilizadas para la realización de este proyecto fueron adquiridas por medio del grupo de Neurociencias de la Universidad de Antioquia. Específicamente se tuvo acceso a más de 6000 pacientes donde algunos de ellos son personas portadoras de la mutación E280A en el gen de la presenilina 1 y el grupo restante son personas no afectadas. La mayoría de estas personas presentan más de una evaluación médica realizada a lo largo del estudio.

2.2.Descripción de los datos

Las cuatro bases de datos se encontraban almacenadas en libros de Microsoft Excel. Cada base de datos contiene información personal del paciente evaluado:

- *Datos generales del paciente:* La base de datos recolecta información acerca de características de la población de estudio como edad, sexo, escolaridad, estado civil, ocupación, entre otros. Su volumetría está dada por 6295 registros de pacientes y 49 atributos que los describen.
- *Evaluaciones médicas neurológicas:* El libro de Excel comprende más de dos bases de datos relacionadas con exámenes médicos y diferentes exámenes neurológicos que le realizan a los pacientes como son funciones mentales superiores, escala CDR y escala Cummings. Su volumetría está dada por 6 bases de datos con un promedio de 7000 registros de pacientes y un total aproximado de 100 atributos.
- *Evaluaciones médicas neuropsicológicas neurodegenerativas:* De igual manera que las evaluaciones médicas neurológicas, estas evaluaciones comprenden exámenes neurológicos de evaluaciones realizadas a los pacientes los cuales pueden sufrir de una enfermedad neurodegenerativa. Su volumetría está dada por 10686 registros de pacientes y 162 atributos que los describen.
- *Genotipificaciones:* La base de datos solo contiene un atributo de interés relacionado con la mutación E280A el cual nos da información de los pacientes que se ven afectados y los que no se ven afectados por esta mutación. Su volumetría está dada por 10686 registros de pacientes y 162 atributos que los describen.

2.3.Exploración de los datos

Una vez descrito los datos, se procedió a explorarlos. Esto implica la aplicación de métodos estadísticos para observar la propiedad de los datos y como estos se encuentran distribuidos:

- *Medidas de tendencia central:* Los atributos cuantitativos de los pacientes fueron representados en tablas que resumen en un solo valor un conjunto grande de valores.
- *Diagramas de cajas:* Los diagramas de caja identificaron la distribución y la tendencia central de los atributos cuantitativos de los pacientes.

- *Diagramas de barra:* Los diagramas de barras dividieron los datos utilizados en sectores ilustrando la proporción numérica de un atributo cuantitativo que describe a los pacientes frente a otro.
- *Gráficos de líneas:* Los gráficos de línea compararon los atributos cualitativos contra los atributos cuantitativos de los pacientes.
- *Matriz de correlación:* La matriz de correlación midió el grado de linealidad entre los atributos de estudio de tipo cuantitativo.

3. Preparación de los datos

En esta tercera fase de la metodología se seleccionaron los datos que se iban a transformar a fin de utilizarlos en la etapa de modelado.

3.1.Selección de los datos

Se identificaron y se eliminaron los atributos que tenían solo un valor constante en todos sus registros y los atributos redundantes que no aportan información adicional al proyecto.

3.2.Limpieza de datos

Posterior a la selección de los datos, se identificaron gran cantidad valores faltantes o nulos en registros de los pacientes evaluados. De acuerdo con estos resultados se realizó lo siguiente:

1. Se eliminaron los atributos que comprendieran más del 40% de registros de pacientes nulos.
2. Se eliminaron el 100% de los registros de pacientes que comprendieran uno o más atributos con algún valor nulo.

3.3.Transformación de los datos

Se transformó el formato y la estructura de los datos sin modificar su significado para facilitar la etapa de modelado como se describe:

1. Se normalizaron los datos con la finalidad de reducir todos los valores de los atributos de estudio a un pequeño rango específico por medio de la técnica '*Min Max'*.

2. Se redujo la dimensionalidad de la base de datos mediante PCA para eliminar características excesivas y que no son necesarias en el modelo.

4. Modelado

Una vez se tenían los datos preparados, se obtuvo el modelo de minería de datos

4.1.Selección de la técnica de modelado

Se seleccionó 'K-medoids' como el modelo más apropiado conforme a la base de datos de estudio obtenida dado que no contiene valores nulos.

5. Evaluación del modelo

En la quinta etapa se determinó la calidad y validez del modelo en dos formas:

1. *Validación interna:* Se utilizó la métrica de desempeño para clustering 'Silhouette'.
2. *Validación externa:* Se compararon los resultados obtenidos con un conjunto de datos del que se conoce el clúster al que pertenece cada observación.

6. Implementación del modelo

En este proyecto no se incluyó la etapa final de la metodología dado que en la etapa anterior no se cumplieron los criterios de éxito para el proyecto.

Análisis y segmentación de la población de estudio con criterios de personal experto en el área

La información utilizada en este análisis es la misma que se utilizó para el análisis sin criterios de personal experto en el área.

Preprocesamiento de las bases de datos independientes

Para el procesamiento de las cuatro bases de datos se consideraron los siguientes pasos:

a) De la base de datos ‘Genotipificaciones’ se seleccionaron todos los registros de pacientes que presentaron un valor diferente de nulo con respecto al atributo ‘E280A’ y se eliminaron los que no cumplen con la condición.
b) Se seleccionaron los atributos definidos por el personal experto en el área en las bases de datos ‘Información general del paciente’, ‘Evaluaciones médicas neurológicas’ y ‘Evaluaciones neuropsicológicas neurodegenerativas’.
c) Se seleccionaron los registros de los pacientes que coincidieron en todas las bases de datos.
d) Se realizó una revisión de cada atributo con la finalidad de identificar los valores nulos. Según el criterio del personal experto en el área algunos registros de pacientes con valores nulos fueron creados, modificados o eliminados.
e) Se eliminaron los atributos que en todos sus registros (o al menos el 90%) tuvieron el mismo valor constante.
f) Se eliminaron los atributos que correspondían a operaciones totales en las bases de datos respectivas a las evaluaciones realizadas a los pacientes.
g) Se creó la base de datos global que contiene todos los atributos necesarios con los registros de pacientes seleccionados eliminando el atributo de comparación ‘E280A’ de la base de datos ‘Genotipificaciones’.

Aplicación de la metodología CRISP-DM

La aplicación de la metodología en este análisis se realizó de manera análoga al análisis sin criterios de personal experto en el área. No obstante, a continuación, se presentan solo las etapas donde hubo diferenciación:

1. *Preparación de los datos*

1.1.Limpieza de los datos

No hubo necesidad de aplicar un filtro para eliminar atributos y registros con valores nulos debido a que con el preprocesamiento realizado previamente se descartaron todos los valores vacíos o nulos. No obstante, se identificaron valores atípicos en los diagramas de cajas por parte del

personal experto en el área. Con estos valores atípicos se realizó una imputación de valores, es decir, se reescribieron los valores atípicos por valores considerados normales según los criterios del personal experto.

2. Modelado

2.1.Selección de la técnica de modelado

Se seleccionó ‘K-means’ como el modelo más apropiado conforme a la base de datos de estudio obtenida dado que no contiene ni valores nulos ni valores atípicos.

V. RESULTADOS Y ANÁLISIS

Así mismo como se describió anteriormente en la metodología, los resultados y análisis del proyecto también se dividen en dos secciones:

1. Análisis y segmentación de la población de estudio sin criterios de personal experto en el área.
2. Análisis y segmentación de la población de estudio con criterios de personal experto en el área.

A continuación, se describe detalladamente los resultados obtenidos para cada una de las dos secciones. Sin embargo, como el análisis de la segunda sección obtuvo mejores resultados con respecto a la segmentación, fue con esta sección con la cual se realizó la estratificación de riesgo de los pacientes.

Análisis y segmentación de la población de estudio sin criterios de personal experto en el área

En esta sección se presentan los resultados obtenidos del análisis exploratorio y de la validación interna y externa de la definición de los segmentos de la primera metodología.

Análisis exploratorio de datos

La base de datos, posterior al preprocesamiento, cuenta con 44 atributos que describen 1114 pacientes. De estos 44 atributos, 19 son cuantitativos y 25 cualitativos.

Exploración de variables cuantitativas

En la Tabla 1 se presenta un resumen estadístico de las variables cuantitativas:

Tabla 1. Resumen estadístico de las variables cuantitativas – Metodología 1.

	count	mean	std	min	25%	50%	75%	max
Presión arterial sistólica	1114	115,1023	14,56134	10	106	112	120	190
Presión arterial diastólica	1114	73,50987	10,14559	50	70	70	80	110
Perímetro cefálico (cms.)	1114	54,98106	1,764156	49	54	55	56	65
Peso (kgs.)	1114	67,01795	23,8539	35	57	65	74	703
Talla (cms.)	1114	157,4966	32,63644	0	156	163	170	191
Minimental - Orientación y tiempo	1114	4,798025	0,639769	0	5	5	5	5
Minimental - Orientación y lugar	1114	4,785458	0,562152	0	5	5	5	5
Minimental - Memoria de fijación	1114	2,991023	0,111794	1	3	3	3	3
Minimental - Atención y cálculo	1114	4,614004	1,014726	0	5	5	5	5
Minimental - Memoria de evocación	1114	2,501795	0,828782	0	2	3	3	3
Minimental - Lenguaje	1114	8,883303	0,449448	4	9	9	9	9
Fast	1114	1,32316	0,70147	1	1	1	2	11
Barthel	1114	49,86535	2,061455	0	50	50	50	50
KATZ	1114	0,048474	0,509904	0	0	0	0	6
Lawton y Brody	1114	7,914722	0,672191	0	8	8	8	8
Edad	1114	40,75583	12,83875	21	30	39	49	81
Escolaridad	1114	9,33842	4,424815	0	5	11	12	25
No. años repetidos	1114	0,517953	0,928795	0	0	0	1	7

La base de datos está conformada por un total de 1114 registros de pacientes en todos sus atributos lo que indica que, efectivamente, no existen valores faltantes. Además, los atributos que presentaron mayor desviación estándar, previa a la imputación de valores, fueron talla, peso, presión arterial sistólica, edad y presión arterial diastólica. Por lo general, estas grandes desviaciones estándar de los datos se vieron alteradas por valores atípicos en cada uno de estos atributos. Por ejemplo, hubo valores incongruentes tanto en talla como en peso dado que no existe persona con una talla de 0 metros o una persona con un peso superior de 700 Kg.

Para verificar los percentiles y la mediana, en la Fig 2 se realizó un diagrama de cajas por cada uno de los atributos:

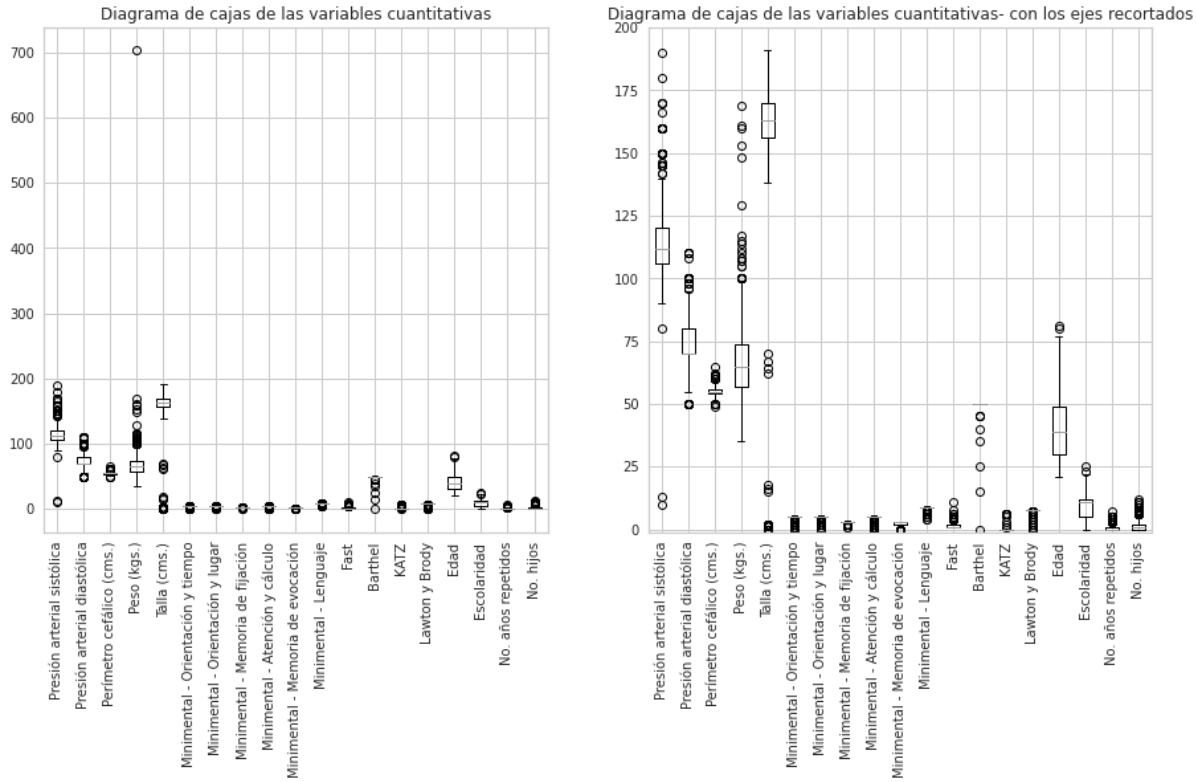

Fig 2. a) Diagrama de cajas sin aumento para cada uno de los atributos, b) Diagrama de cajas con aumento para cada uno de los atributos

En la gráfica a) se identifica que el atributo relacionado con el peso presentó un valor atípico demasiado alto en comparación con los demás diagramas de cajas lo que afecta la visualización de la distribución de los otros atributos. Con respecto a la gráfica b), la cual tiene aumento, se evidencia que los atributos relacionados con la presión arterial sistólica, el peso, la talla, la edad y el número de hijos de los pacientes prácticamente presentaron una distribución simétrica, es decir, la mayoría de sus datos estuvieron igualmente distribuidos tanto por debajo como por encima de la mediana. Por el contrario, los atributos relacionados con la presión arterial diastólica y la escolaridad de los pacientes presentaron una distribución asimétrica. La distribución de los datos de la presión arterial diastólica no presentó una gran dispersión por debajo de la mediana, es decir, al menos el 50% de los pacientes tuvo una presión arterial diastólica muy similar y alrededor de 70 mmHg. En cambio, la distribución de la escolaridad de los pacientes presentó una alta dispersión por debajo y poca por encima de la mediana, lo cual quiere decir que al menos el 50% de los pacientes presentaron una escolaridad similar que estuvo alrededor del onceavo año.

En la Fig 3 se presenta la correlación de estos atributos cuantitativos posterior a la imputación de valores para eliminar los valores atípicos. Esta matriz recogió las similitudes entendidas como el grado de cercanía entre los elementos analizados donde los valores más altos indican una mayor cercanía; mientras que, si los casos se encuentran separados, los valores tienden a tener un valor bajo. Como se puede apreciar, la diagonal contiene el valor de 1 al relacionarse cada atributo consigo mismo.

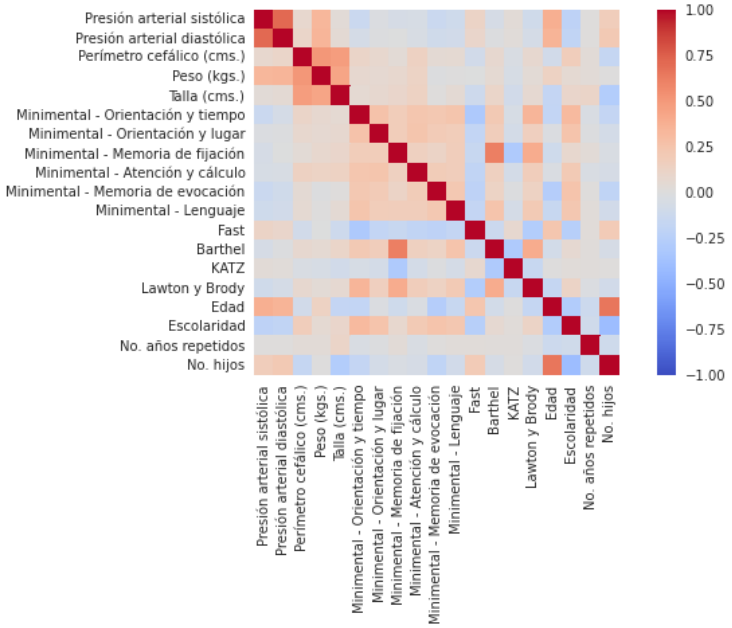

Fig 3. Matriz de correlación de variables cuantitativas

El análisis de la Fig 6 muestra que existen atributos que presentan una correlación positiva alta como también una correlación negativa alta. Dentro de las correlaciones positivas se identifican las siguientes relaciones:

- La presión arterial diastólica y la presión arterial sistólica presentan una de las correlaciones más fuertes. Como su coeficiente es positivo, ambas variables tienden a aumentar a la vez, es decir, el valor más alto de la presión arterial diastólica está asociado con el valor más alto de la presión arterial sistólica.

- Minimental - Memoria de fijación y Barthel también presentan una correlación relativamente fuerte. Como su coeficiente es positivo, ambas variables tienden a aumentar a la vez.
- La edad y el número de hijos de los pacientes también se relacionaron. Según la matriz de correlación, a mayor edad, mayor es el número de hijos.

Para las correlaciones negativas se obtienen las siguientes relaciones:

- La escolaridad y el número de hijos presentan una de las correlaciones más fuertes. Como su coeficiente es negativo, el valor más alto de la escolaridad está asociado con el valor más bajo del número de hijos o viceversa.
- Minimental - Orientación y tiempo también presentan una correlación relativamente fuerte. Como su coeficiente es negativo, mientras una aumenta, la otra tiende a disminuir.

Exploración de variables cualitativas

En la Fig 4 se presenta la proporción de pacientes por ocupación. Del análisis de las 174 ocupaciones, se obtuvo una agrupación de 21 ocupaciones.



Fig 4. Histograma de la ocupación de los pacientes.

La categoría de ‘Otras_ocupaciones’ presenta la mayor cantidad de pacientes registrados. Sin embargo, esta variable fue el resultado de unir las 153 ocupaciones restantes que no presentaban

un número considerable de pacientes. Por ende, la mayoría de las pacientes que se incluyeron en el estudio son ama de casa o estudiante.

En la Fig 5 se presentan las relaciones del atributo ‘Queja de memoria (QP/QF)’ contra los atributos más relevantes, entre estos, el sexo y las alteraciones presentadas en las actividades instrumentales en la vida cotidiana de los pacientes.

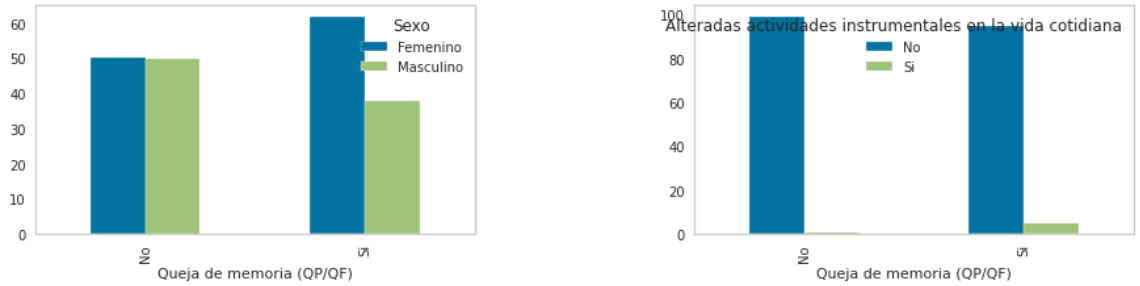

Fig 5. Relaciones relevantes entre algunas variables cualitativas.

No existe diferencias significativas entre hombres y mujeres que no presentan queja de memoria. Sin embargo, si es mayor la proporción de mujeres que de hombres que sufren queja de memoria. En cuanto a las alteraciones en las actividades instrumentales en la vida cotidiana, solo los pacientes que presentan queja de memoria también presentan alteraciones en dichas actividades.

Exploración de variables cuantitativas y cualitativas

En la Fig 6 se presentan siete gráficos de línea donde se muestran las relaciones más relevantes entre las variables cuantitativas y cualitativas:

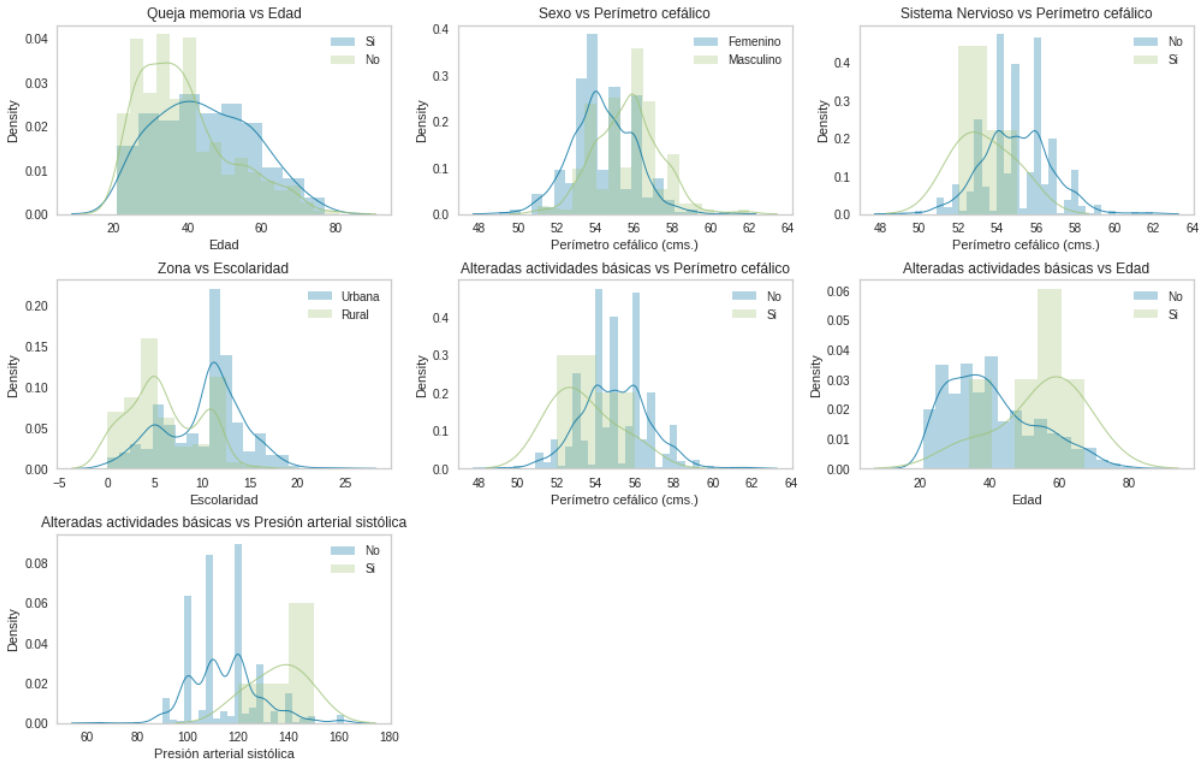

Fig 6. Relaciones relevantes entre las variables cuantitativas y cualitativas.

Se observa que, por cada gráfico de línea, existe una característica relevante que se cumple para la mayoría de los pacientes. En primer lugar, se identifica que los pacientes de mayor edad son los que presentan queja de memoria y alteraciones en las actividades básicas en la vida cotidiana. En segundo lugar, se evidencia que los pacientes con bajo perímetro cefálico, por lo general, pertenecen al sexo femenino y cuentan con problemas en el sistema nervioso como con alteraciones en las actividades básicas en la vida cotidiana. En tercer lugar, los pacientes con baja escolaridad son provenientes de zona rural. En cuarto y último lugar, los pacientes con presión arterial sistólica alta presentan alteraciones en las actividades básicas en la vida cotidiana.

Definición de los segmentos

Validación interna

El análisis de silueta o método de la silueta permitió medir la calidad del agrupamiento o clustering, es decir, qué tan compactos fueron los clústeres en su interior (homogeneidad) y qué

tan separados o alejados se encontraron entre sí (heterogeneidad) mediante el algoritmo K-medoids implementado. Este algoritmo fue seleccionado para esta metodología dado que la base de datos conformada, después de la imputación, sigue conteniendo valores atípicos y, según la teoría, este tipo de algoritmo no se ve afectado por valores atípicos como es el algoritmo K-means.

En la Fig 7, se expone el método de la silueta:

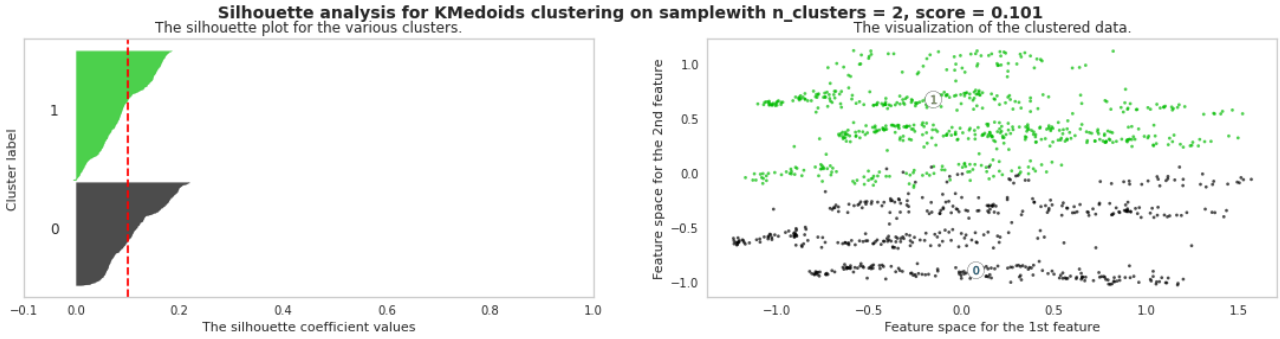

Fig 7. Representación del método Silhouette – Metodología 1.

Se contempla que el score de la métrica de desempeño fue muy bajo, tan solo fue de 0.101. Esto indica que la técnica implementada para la de eliminación no es lo suficiente efectiva para la diferenciación, pues al eliminar las filas y columnas que tienen el mayor porcentaje de valores nulos, se pudo perder información muy valiosa, es decir, atributos que son fundamentales para descubrir pacientes diagnosticados con la enfermedad de Alzheimer. No obstante, los dos clústeres obtenidos no presentaron valores negativos, lo que indica que los datos están clasificados correctamente dado que están más cerca de su propio clúster que al clúster del vecino. Además, ambos clústeres obtuvieron, en promedio, un score y un tamaño muy similar, pues el número de pacientes clasificados en ambos clústeres es aproximadamente equivalente.

Validación externa

La población de estudio se dividió en dos segmentos como se mostró anteriormente. Para la validación externa de estos dos segmentos se hizo uso de un conjunto de datos del cual se tenía conocimiento y el cual contenía el clúster al que pertenece cada observación.

En la Fig 8, se expone la distribución de los clústeres que fueron definidos:

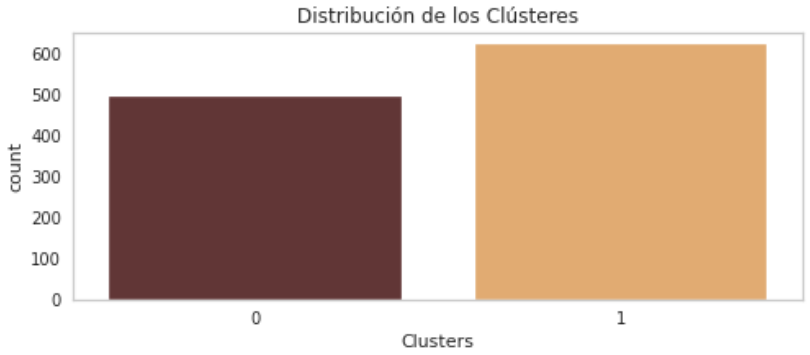

Fig 8. Distribución del clúster 0 y 1 – Metodología 1.

El clúster 0 está conformado por 494 pacientes siendo el 62.77% de la población de estudio total. El clúster 1 está conformado por 620 pacientes siendo el 78.78% de la población de estudio total.

En cuanto a los atributos cuantitativos, en la Tabla 2, se presentan las características más relevantes de los dos segmentos:

Tabla 2. Atributos cuantitativos del clúster 0 y 1 – Metodología 1.

	Clústeres	
	0	**1**
Presión arterial sistólica	114,98	115,25
Presión arterial diastólica	72,98	73,93
Perímetro cefálico (cms.)	54,87	55,06
Peso (kgs.)	66,01	66,62
Talla (cms.)	159,78	163,58
Minimental - Orientación y tiempo	4,83	4,77
Minimental - Orientación y lugar	4,81	4,77
Minimental - Memoria de fijación	2,99	2,99
Minimental - Atención y cálculo	4,64	4,6
Minimental - Memoria de evocación	2,51	2,5
Minimental - Lenguaje	8,9	8,87
Fast	1,34	1,31
Barthel	49,76	49,95
KATZ	0,05	0,05
Lawton y Brody	7,9	7,92
Edad	41,69	40,01
Escolaridad	10,13	8,71
No. años repetidos	0,4	0,6
No. hijos	1,55	1,5

Aunque existe una diferencia entre segmentos, la diferencia es mínima, lo que indica que no existe realmente una diferencia significativa entre los dos segmentos en cuanto a los atributos cuantitativos los cuales hacen referencia a las pruebas neurológicas realizadas por los pacientes. Por ende, los pacientes pertenecientes al clúster 0 como los pertenecientes al clúster 1, presentan un desempeño similar en las pruebas neurológicas desarrolladas.

En cuanto a los atributos cualitativos, en la Fig 9, se presenta una de las características más relevantes que logró diferenciar los dos segmentos:

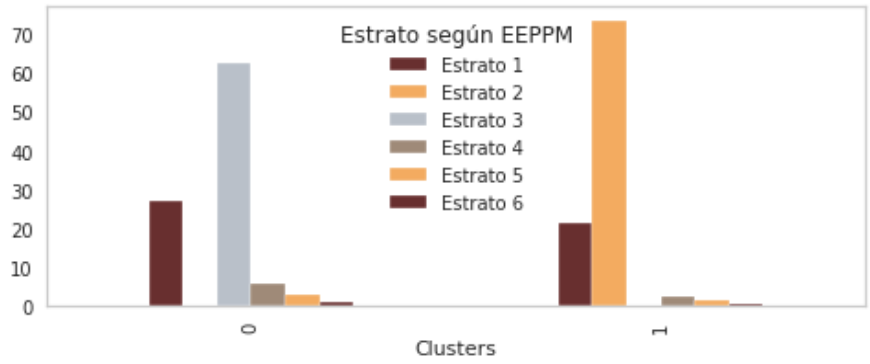

Fig 9. Pacientes del clúster 0 y 1 divididos por su estrato según EEPM

Se evidencia que la mayoría de los pacientes del clúster 0 pertenecen al estrato 1 y 3, mientras que los pacientes del clúster 1 al estrato 1 y 2. Los pacientes que pertenecen a los estratos 4, 5 y 6 quedaron divididos entre los dos segmentos.

En la Fig 10 se presenta la proporción de los pacientes pertenecientes al clúster 0 y 1 con respecto a la variable objetivo, es decir, el conjunto de datos del cual se tenía conocimiento y el cual contenía el clúster al que pertenecía cada observación. Esta variable objetivo, nombrada ‘E280A’, hace referencia a los pacientes que son portadores de este tipo de mutación. En este caso, los pacientes con el estado ‘A’ de esta variable objetivo indicaban que no eran portadores de la mutación y los pacientes con el estado ‘B’, que si eran portadores:

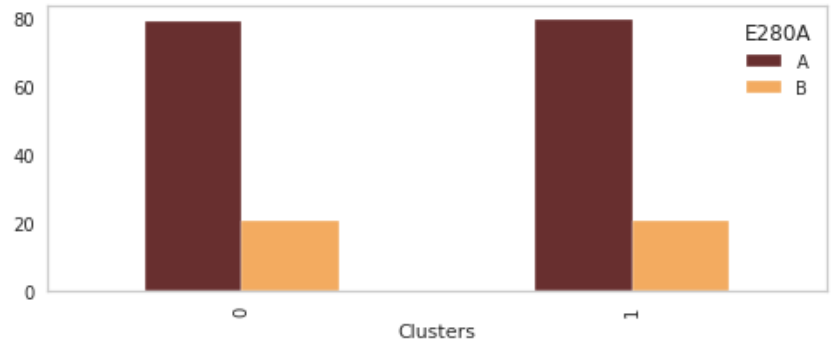

Fig 10. Proporción de los pacientes pertenecientes al clúster 0 y 1 con respecto a la variable target.

Efectivamente, tanto el clúster 0 como el clúster 1 incluyen pacientes que son portadores de la mutación E280A en una proporción similar. El clúster 0 tiene 391 pacientes que no son portadores de la mutación y 103 que si son portadores. El clúster 1 tiene 492 pacientes que no son portadores de la mutación y 128 que si son portadores. Por ende, se deduce que no existe diferenciación entre los dos segmentos obtenidos y, por ende, no tiene sentido definir estratificación de riesgo de los pacientes.

Análisis y segmentación de la población de estudio con criterios de personal experto en el área

De la misma manera que la primera metodología, en esta sección se describen los resultados obtenidos del análisis exploratorio y de la validación interna y externa de la segunda metodología, a diferencia de que en esta sección si se realiza la definición de la estratificación de riesgo de los pacientes.

Análisis exploratorio de datos

La base de datos, posterior al preprocesamiento según los criterios del personal experto, cuenta con 66 atributos que describen 787 pacientes. De estos 66 atributos, 54 son cuantitativos y 12 cualitativos.

Exploración de variables cuantitativas

En la Tabla 3 se presenta un resumen estadístico de las variables cuantitativas:

Tabla 3. Resumen estadístico de las variables cuantitativas – Metodología 2.

	count	mean	std	min	25%	50%	75%
Edad	787	41,63151	12,42838	20	32	39	50
Escolaridad	787	8,39263	5,797218	0	5	8	11
No. años repetidos	787	0,547649	1,041727	0	0	0	1
No. hijos	787	1,759848	1,916717	0	0	1	2
Edad inicio QM - sin impacto en la vida	787	3,936468	11,63705	0	0	0	0
Edad inicio QM - con impacto	787	1,758577	8,660998	0	0	0	0
Edad inicio demencia	787	0,951715	6,637675	0	0	0	0
Presión arterial sistólica	787	117,7916	13,78967	80	110	120	120
Presión arterial diastólica	787	75,0521	9,97126	38	70	75	80
Depresión	787	0,151207	0,490373	0	0	0	0
Ansiedad_x	787	0,203304	0,556561	0	0	0	0
Depresión/Disforia	787	0,514612	1,560723	0	0	0	0
Ansiedad_y	787	0,60737	1,598217	0	0	0	0
Irritabilidad/Liabilidad emocional	787	0,345616	1,333648	0	0	0	0
Sueño/Trastornos del sueño	787	0,468869	1,437998	0	0	0	0
Escolaridad en la evaluación	787	8,149936	4,216819	0	5	8	11
Minimental - Orientación y tiempo	787	4,80432	0,501734	1	5	5	5
Minimental - Orientación y lugar	787	4,810673	0,502905	2	5	5	5
Minimental - Atención y cálculo	787	4,668361	0,894173	0	5	5	5
Minimental - Memoria de evocación	787	2,513342	0,815413	0	2	3	3
Fluidez 0 - 15	787	8,083863	2,203532	3	6	8	10
Fluidez 16 - 30	787	4,13723	1,918564	0	3	4	5
Fluidez 31 - 45	787	3,062262	1,702397	0	2	3	4
Fluidez 46 - 60	787	2,447268	1,716764	0	1	2	3
Denominación - Media	787	3,870394	1,176952	1	3	4	5
Denominación - Baja	787	3,246506	1,10756	0	3	3	4
MLP - Ensayo 1	787	3,547649	1,483996	0	3	3	5
MLP - Intrusiones Ensayo 1	787	0,580686	1,060286	0	0	0	1
MLP - Ensayo 2	787	6,015248	1,593489	1	5	6	7
MLP - Intrusiones Ensayo 2	787	0,212198	0,549766	0	0	0	0
MLP - Ensayo 3	787	7,283355	1,56448	3	6	7	8
MLP - Intrusiones Ensayo 3	787	0,147395	0,471722	0	0	0	0
MLP - Evocación palabras	787	6,182973	1,831882	0	5	6	8
MLP - Intrusiones Evocación	787	0,321474	0,806831	0	0	0	0
MLP - Reconocimiento palabras SI	787	9,579416	0,892281	3	9,5	10	10
Praxias - Círculo evocación	787	1,794155	0,588971	0	2	2	2
Praxias - Rombo ejecución	787	2,7446	0,556782	0	3	3	3
Praxias - Rombo evocación	787	2,456163	0,980236	0	2	3	3
Praxias - Rectángulos evocación	787	1,724269	0,677864	0	2	2	2
Praxias - Cubo ejecución	787	3,002541	1,228373	0	2	3	4
Praxias - Cubo evocación	787	2,311309	1,560783	0	1	3	4
TMT - tiempo	787	72,02922	52,52481	0	43,5	57	82,5
Figura de rey Copia	787	27,23698	5,975252	2	24	28	32
Figura de rey Evocación	787	15,35006	7,063625	0	9,5	15,5	21
Figura de rey % de retención	787	55,46411	21,3442	0	39,64461	56,09756	70,84972

Wisconsin aciertos	787	20,58196	8,485841	0	14	20	27
Wisconsin errores	787	27,39263	8,518777	7	21	28	34
Wisonsin categorias	787	2,626429	1,380605	0	2	2	3
Wisconsin Rtas. perseverativas	787	19,76112	8,628691	2	14	19	24
Winsconsin Rtas. nivel conceptual	787	17,8399	8,765668	0	12	17	24
Trastornos memoria QP	787	12,5324	8,594329	0	6	11	18
Yesavage	787	3,111817	3,370309	0	1	2	4
Fast	787	1,264295	0,520602	1	1	1	1

La base de datos está conformada por un total de 787 registros de pacientes en todos sus atributos lo que indica que, efectivamente, no existen valores faltantes. Además, los atributos que presentaron mayor desviación estándar, previa a la imputación de valores, fueron TMT-tiempo, IMC, Figura de rey % de retención, presión arterial sistólica, Edad y Edad inicio QM - sin impacto en la vida. Por lo general, estas grandes desviaciones estándar de los datos se vieron alteradas por valores atípicos en cada uno de estos atributos. Por ejemplo, hubo valores incongruentes tanto en IMC como en Figura de rey % de retención dado que no existe persona con un IMC de 283 kg/m^2 o una persona con un porcentaje de retención superior a 100%. Sin embargo, estos valores atípicos fueron modificados según los criterios del personal experto en el área.

Para verificar los percentiles y la mediana, en las Fig 11, Fig 12, Fig 13 y Fig 14 se realizaron diagramas de cajas de los atributos. Cada diagrama de cajas obtuvo un comportamiento diferente, por esta razón, a continuación, se describe el comportamiento de cada atributo:

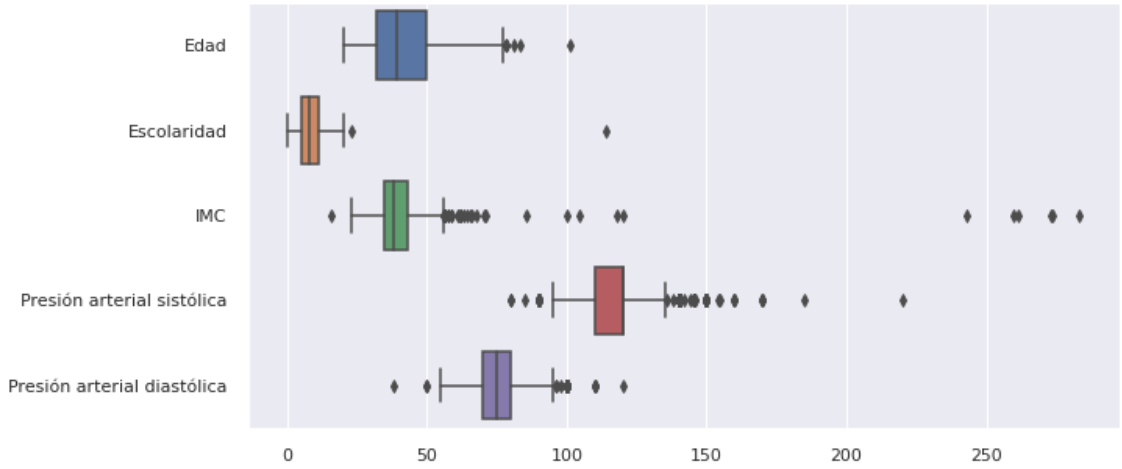

Fig 11. Diagrama de cajas para los atributos cuantitativos Edad, Escolaridad, IMC, Presión arterial sistólica y Presión arterial diastólica.

Edad: Hasta el segundo cuartil (50% de los datos), la distribución de los datos fue homogénea, las edades de los pacientes evaluados comenzaron desde la edad mínima de 20 hasta 39 años. Del segundo cuartil en adelante, la distribución de los datos no fue tan homogénea, las edades de los pacientes evaluados comenzaron desde la edad mínima de 39 hasta 101 años. No obstante, siendo 101 años un valor atípico por el diagrama de cajas, en la realidad no se pudo asegurar que fuera un valor atípico dado que era muy probable que en la población de estudio existiera un paciente con esta edad.

Escolaridad: La distribución de los datos a lo largo de toda la caja fue homogénea. Sin embargo, existía al menos un valor atípico. Uno de estos valores atípicos, el cual estuvo muy alejado de los valores restantes, fue 114. Realmente no había certeza que fuera atípico ya que era probable que en la sociedad existieran personas con este nivel de escolaridad y esta persona hiciera parte de la población de estudio del proyecto.

IMC: Hasta el segundo cuartil (50% de los datos), la distribución de los datos fue homogénea, el IMC de los pacientes evaluados empezaba desde 15.58 hasta 37.97 Kg/m2. Del segundo cuartil en adelante, la distribución de los datos no era tan homogénea, empezaba desde 37.97 hasta 283.3 Kg/m2. Este último valor no era lógico, por ende, hacía parte de los valores atípicos. Además, también contaba con un solo valor atípico de 15.58 Kg/m2. Más adelante hubo la necesidad de realizarles un tratamiento a estos valores atípicos según criterios del personal experto.

Presión arterial sistólica: Hasta el segundo cuartil (50% de los datos), la distribución de los datos fue homogénea, la presión sistólica de los pacientes evaluados empezaba desde 80 hasta 120 mmHg. Del segundo al tercer cuartil (25% de los datos), se mantuvo constante con un valor de 120 mmHg, lo cual fue un valor lógico ya que era la media de la población en general y fue un valor cercano de la media de todos los pacientes evaluados en el estudio. No obstante, hubo valores atípicos por debajo y por encima del bigote menor y mayor respectivamente. Uno de ellos fue el valor mínimo de 80 y mayor de 220 mmHg. Más adelante hubo la necesidad de realizarles un tratamiento a algunos valores atípicos según criterios del personal experto.

Presión arterial diastólica: La distribución de los datos a lo largo de toda la caja fue homogénea. Sin embargo, hubo valores atípicos por debajo del bigote menor y por encima del bigote mayor.

Uno de los valores atípicos por debajo del bigote menor fue de 38 y por encima de 120 mmHg. Fue necesario realizar un tratamiento de datos con estos valores atípicos posteriormente según criterios del personal experto.

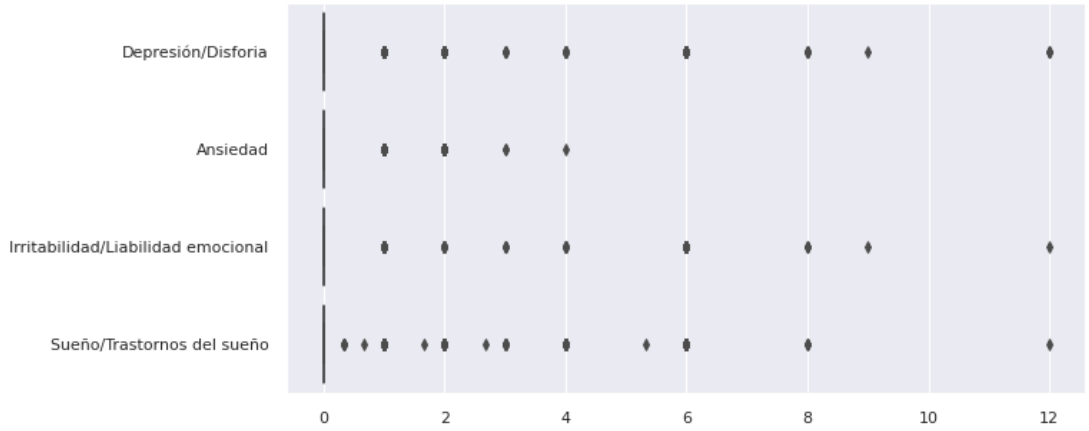

Fig 12. Diagrama de cajas para los atributos cuantitativos Depresión/Disforia, Ansiedad_y, Irritabilidad/Liabilidad emocional, Sueño/Trastornos del sueño.

Todos los atributos involucrados tuvieron desde el bigote menor hasta el mayor un valor constante de cero y solo presentaron valores atípicos por encima del bigote mayor. No obstante, no se pudieron considerar realmente valores atípicos porque estuvieron dentro de los límites establecidos de valores que pueden tomar las evaluaciones médicas estándar.

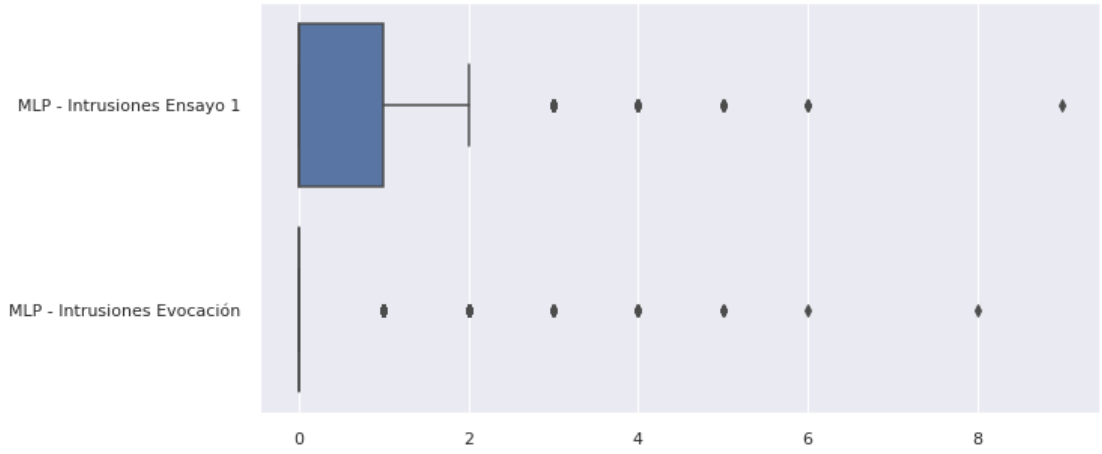

Fig 13. Diagrama de cajas para los atributos cuantitativos MLP - Intrusiones Ensayo 1 y MLP – Intrusiones Evocación.

MLP – intrusiones Ensayo 1: Su comportamiento presentó un valor constante de cero desde el valor mínimo hasta el segundo cuartil (50 % de los datos), del segundo al tercer cuartil, presentó valores entre 0 y 1, del tercer cuartil al bigote mayor presentó valores entre 1 y 2 y del bigote mayor en adelante ambos presentó solo valores atípicos por encima. No obstante, estos valores atípicos en la realidad no son considerados valores atípicos dado que se encuentran dentro de los límites establecidos de valores que pueden tomar las evaluaciones médicas estándar.

MLP- Intrusiones Evocacion: Su comportamiento presentó un valor constante de cero desde el valor mínimo hasta el tercer cuartil y presentó solo valores atípicos por encima del bigote mayor. No obstante, estos valores atípicos en la realidad no son valores atípicos dado que se encuentran dentro de los límites establecidos de valores que pueden tomar las evaluaciones médicas estándar.

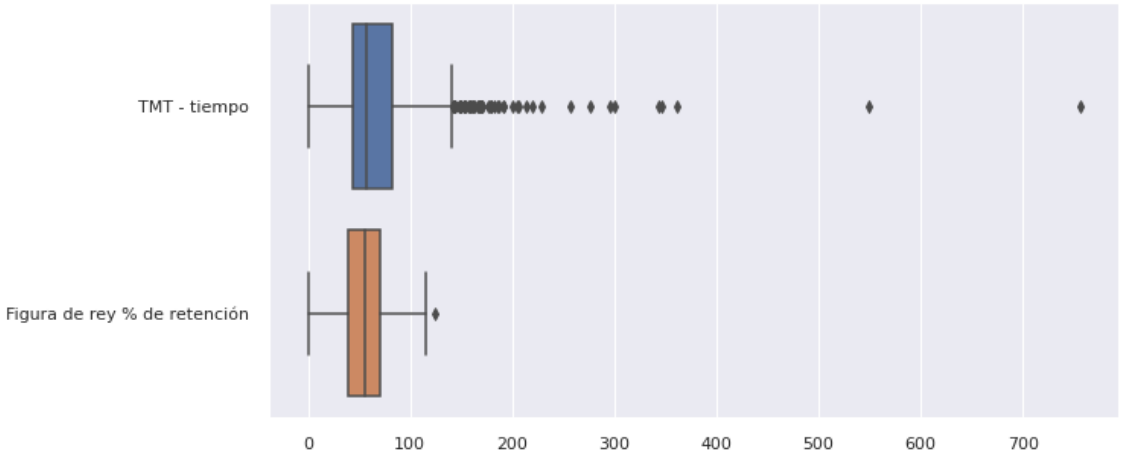

Fig 14. Diagrama de cajas para los atributos cuantitativos TMT - tiempo y Figura de rey % de retención.

TMT – tiempo: Hasta el segundo cuartil (50% de los datos), la distribución de los datos fue homogénea, el tiempo comenzaba desde 0 hasta 57 s. Del segundo cuartil en adelante, la distribución de los datos no fue tan homogénea, el tiempo comenzaba desde 57 hasta 756 s. Uno de los valores atípicos por encima del bigote mayor fue de 756 s, lo cual no se pudo considerar como valor atípico dado que se le pudo dificultar realizar la prueba al paciente.

Figura de rey % de retención: La distribución de los datos a lo largo de toda la caja fue homogénea. Sin embargo, hubo al menos un valor atípico. Uno de estos valores atípicos, por

encima del bigote mayor, fue 124.4. Realmente no se tuvo certeza que fuera un valor atípico.

En la Tabla 4 se presenta, después de la imputación según criterios del personal experto, la correlación de algunos de estos atributos cuantitativos que presentan una correlación relativamente alta. Además, se analiza la significancia de la correlación, pues si la significancia es mayor a 0.05, la correlación entre las variables analizadas no existe.

Tabla 4. Correlaciones entre los atributos cuantitativos – Metodología 2.

X	Y	r	p-unc
Edad	**No. hijos**	0,620833	4,72E-85
Escolaridad	**TMT - tiempo**	-0,60409	1,91E-79
Escolaridad	**Figura de rey Copia**	0,512896	5,12E-54
Edad inicio QM - con impacto	**Edad inicio demencia**	0,656083	4,84E-98
Presión arterial sistólica	**Presión arterial diastólica**	0,737989	3,1E-136
Depresión	**Ansiedad**	0,610262	1,79E-81
Depresión	**Depresión/Disforia**	0,722433	6,4E-128
Denominación - Media	**Denominación - Baja**	0,500887	3,26E-51
MLP - Ensayo 1	**MLP - Ensayo 2**	0,633546	1,53E-89
MLP - Ensayo 1	**MLP - Ensayo 3**	0,538678	1,99E-60
MLP - Ensayo 1	**MLP - Evocación palabras**	0,54416	7,29E-62
MLP - Ensayo 2	**MLP - Ensayo 3**	0,68846	1,3E-111
MLP - Ensayo 2	**MLP - Evocación palabras**	0,680586	3,7E-108
MLP - Ensayo 3	**MLP - Evocación palabras**	0,733658	7,4E-134
MLP - Intrusiones Ensayo 3	**MLP - Intrusiones Evocación**	0,533621	3,98E-59
Praxias - Cubo ejecución	**Praxias - Cubo evocación**	0,643661	2,87E-93
Praxias - Cubo evocación	**Figura de rey Evocación**	0,534764	2,03E-59
Figura de rey Copia	**Figura de rey Evocación**	0,636393	1,41E-90
Figura de rey Evocación	**Figura de rey % de retención**	0,869446	1,3E-242
Wisconsin aciertos	**Wisconsin errores**	-0,99916	0
Wisconsin aciertos	**Wisonsin categorias**	0,804497	8,8E-180
Wisconsin aciertos	**Wisconsin Rtas. perseverativas**	-0,87149	4E-245
Wisconsin aciertos	**Winsconsin Rtas. nivel conceptual**	0,876588	1,5E-251
Wisconsin errores	**Wisonsin categorias**	-0,80504	3,4E-180
Wisconsin errores	**Wisconsin Rtas. perseverativas**	0,870836	2,6E-244
Wisconsin errores	**Winsconsin Rtas. nivel conceptual**	-0,87603	7,6E-251
Wisonsin categorias	**Wisconsin Rtas. perseverativas**	-0,68816	1,8E-111
Wisonsin categorias	**Winsconsin Rtas. nivel conceptual**	0,850023	9,6E-221
Wisconsin Rtas. perseverativas	**Winsconsin Rtas. nivel conceptual**	-0,7328	2,1E-133
Trastornos memoria QP	**Fast**	0,609239	3,91E-81

Se identifica, en primer lugar, que la correlación entre las variables cuantitativas es significativa ya que $p<0.05$. En segundo lugar, se evidencia que muchas de las correlaciones obtenidas son relacionadas con un atributo perteneciente a la misma prueba médica realizada a los

pacientes. Por esta razón, solo se analizan las correlaciones relativamente altas que no tuvieron relación con algún atributo perteneciente a la misma prueba:

Edad vs No. hijos: La correlación presenta un puntaje positivo y relativamente alto, mayor a 0.5, es decir, los pacientes de mayor edad son los pacientes con mayor número de hijos.

Escolaridad vs TMT – tiempo: La correlación presenta un puntaje negativo y relativamente alto, menor a -0.5, es decir, los pacientes que tiene estudios más altos obtienen menores tiempos en la realización de Trail Making Test (TMT).

Escolaridad vs Figura de rey Copia: La correlación presenta un puntaje positivo y relativamente alto, mayor a 0.5, es decir, los pacientes que tienen estudios más altos también tienen mayor puntaje en la realización de la prueba Figura de rey.

Depresión vs Ansiedad: La correlación presenta un puntaje positivo y relativamente alto, mayor a 0.5, es decir, los pacientes que tienen puntajes más altos en el atributo Depresión también obtienen puntajes altos en el atributo Ansiedad.

Praxias - Cubo evocación vs Figura de rey Evocación: La correlación presenta un puntaje positivo y relativamente alto, mayor a 0.5, es decir, los pacientes que tienen resultados altos en la prueba de praxias, específicamente en evocación, también obtienen mayor puntaje en la realización de la prueba Figura de rey Evocación.

Trastornos memoria QP vs Fast: La correlación presenta un puntaje positivo y relativamente alto, mayor a 0.5, es decir, los pacientes que obtienen mayores puntajes en la prueba de trastornos de memoria, contestada por el paciente, también obtienen mayores puntajes en la prueba Fast dado que el nivel de funcionalidad del paciente en la vida diaria se va volviendo menor.

Exploración de variables cualitativas

En la Fig 15, se presentan las primeras 9 ubicaciones (departamento, ciudad) de donde provienen la mayor cantidad de pacientes que participaron en el estudio. La décima ubicación nombrada 'Otro' hace referencia a las ubicaciones restantes.

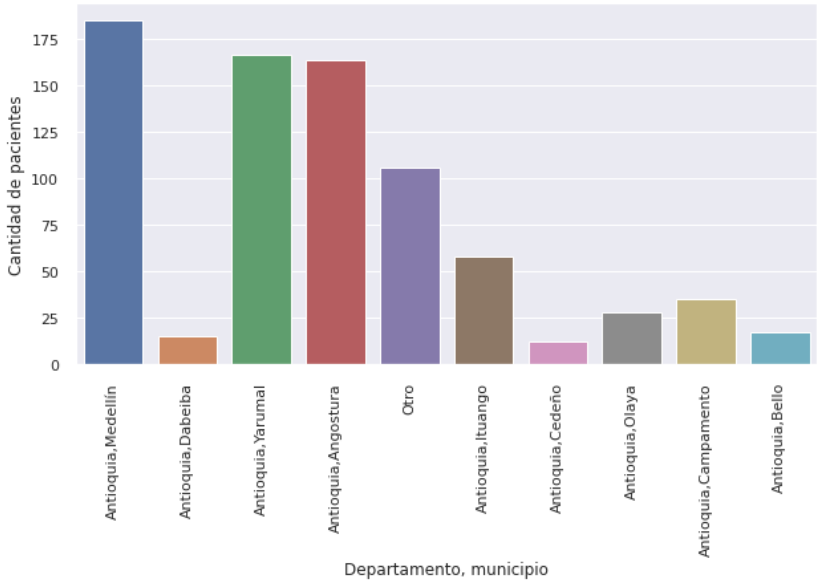

Fig 15. Distribución de las ubicaciones de dónde provinieron los pacientes que participaron en el estudio.

Antioquia, Medellín fue la ubicación con mayor número de pacientes seguido de Antioquia, Yarumal y Antioquia, Angostura. Evidentemente, todos los municipios seleccionados anteriormente pertenecen al departamento de Antioquia. Es decir, la mayoría de los pacientes son pertenecientes al departamento de Antioquia.

En la Fig 16 se presenta la interpretación de los atributos que tienen relación con el atributo 'ubicación' y 'zona':

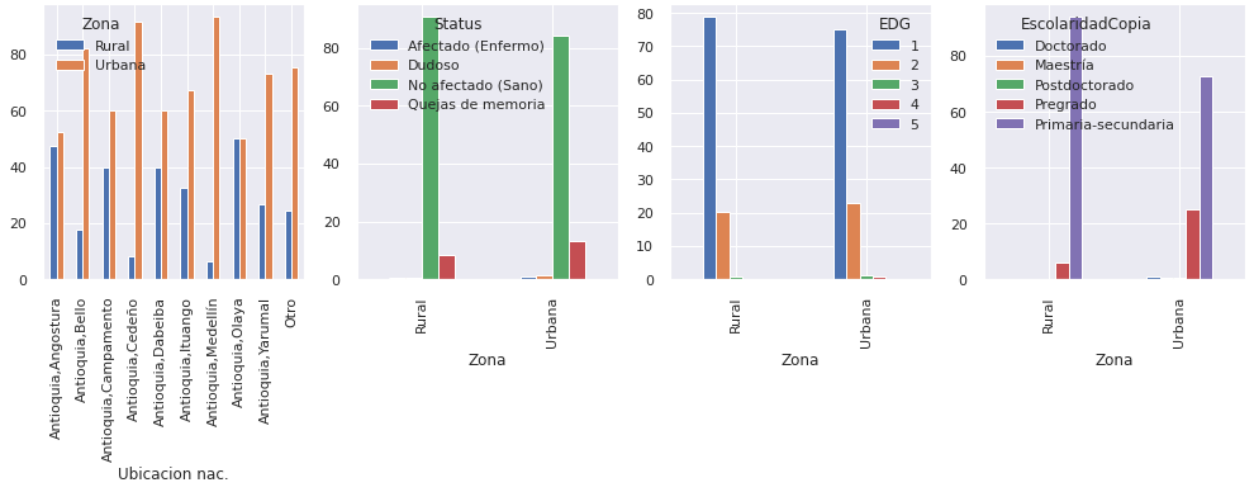

Fig 16. Relaciones relevantes de los atributos 'Ubicacion nac.' y 'Zona'.

Medellín, Bello y Cedeño son los municipios con mayor cantidad de pacientes provenientes de zona urbana. Medellín es el municipio con mayor población en la zona urbana dada que es una de las ciudades principales de Colombia. Asimismo, Bello es un municipio que pertenece al área metropolitana del Valle de Aburrá de donde la mayoría de los pacientes son provenientes. No obstante, Cedeño, siendo uno de los municipios con menor número de pacientes nacidos allí, pertenece a uno de los 7 corregimientos de Antioquia, Yarumal.
En cuanto al estatus, los pacientes que son provenientes de la zona urbana son los únicos que obtienen estados de ‘Dudoso’ y ‘Afectado’. Así mismo sucede con la prueba EDG, pues los pacientes que son provenientes de zona urbana son los únicos que obtienen puntajes superiores a 2. Los resultados con el atributo ‘Escolaridad’ eran de esperarse, los pacientes que son provenientes de zonas urbanas también obtienen una mayor escolaridad, es decir, por encima del pregrado.

En las Fig 17, Fig 18, Fig 19 y Fig 20, se presentan los atributos ‘Zona’, ‘Alteradas actividades instrumentales en la vida cotidiana’, ‘Status’ y ‘EDG’ frente a las tres etapas de la enfermedad de Alzheimer:

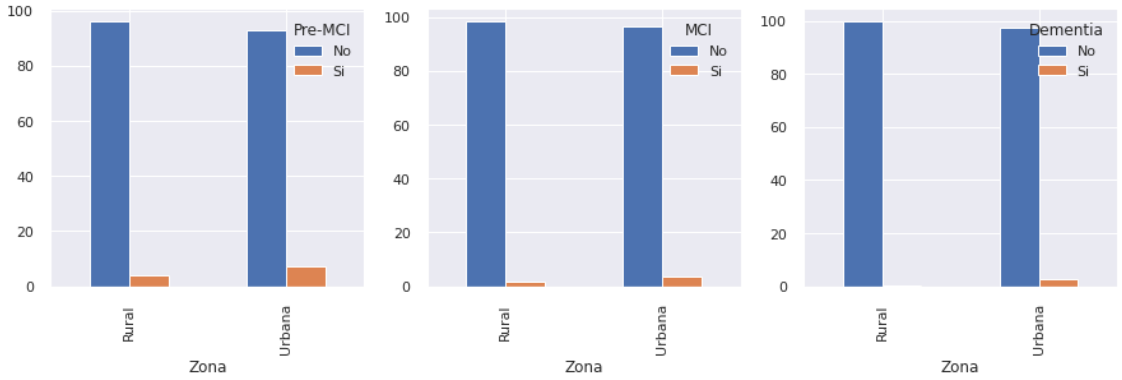

Fig 17. Atributo 'Zona' frente a las tres etapas de la enfermedad de Alzheimer.

Como primer punto se tiene que la cantidad de pacientes, tanto en la zona rural y urbana, va disminuyendo a medida que aumenta la gravedad de la etapa de la enfermedad de Alzheimer. Además, se evidencia que los únicos pacientes que desarrollan demencia son los pacientes que provienen de la zona urbana.

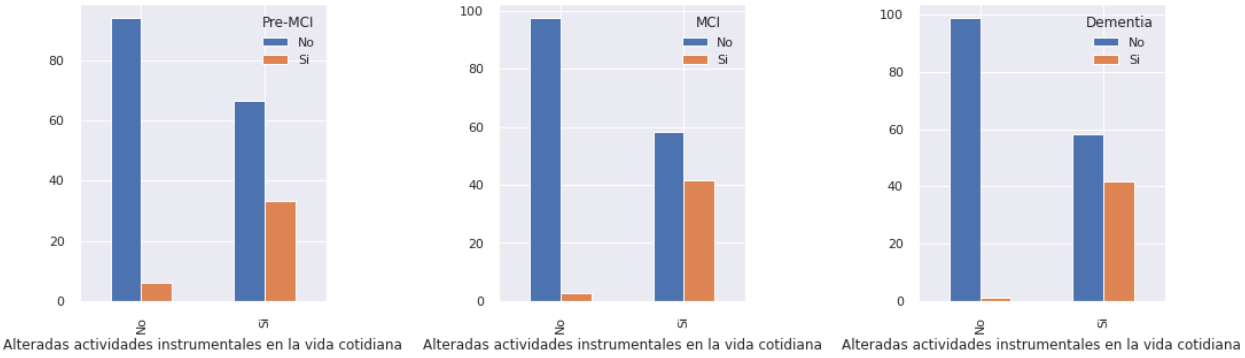

Fig 18. Atributo 'Alteradas actividades instrumentales en la vida cotidiana' frente a las tres etapas de la enfermedad de Alzheimer.

En segundo lugar, se evidencia que la cantidad de pacientes que no presentan alteraciones va disminuyendo, así como va aumentando la cantidad de los que si presentaban alteraciones a medida que aumentaba la gravedad de la etapa. Además, se identifica, en la última etapa, que prácticamente todos los pacientes que sufren de demencia presentan alteraciones en las actividades instrumentales en la vida cotidiana.

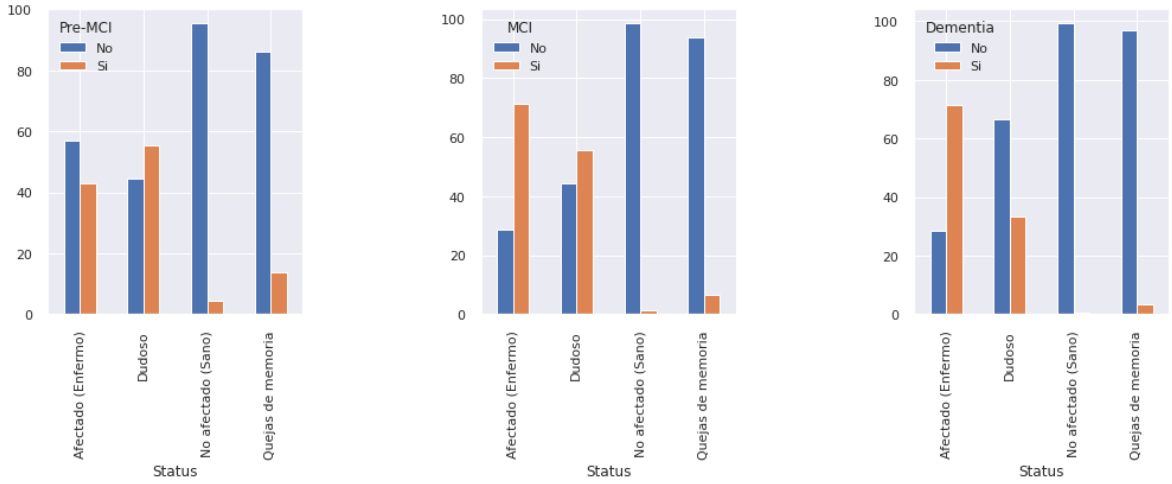

Fig 19. Atributo 'Status' frente a las tres etapas de la enfermedad de Alzheimer.

También se observa que, a medida que la etapa de la enfermedad es más grave, el número de pacientes con quejas de memoria y no afectados(sano) disminuye, pero el número de afectados aumenta considerablemente.

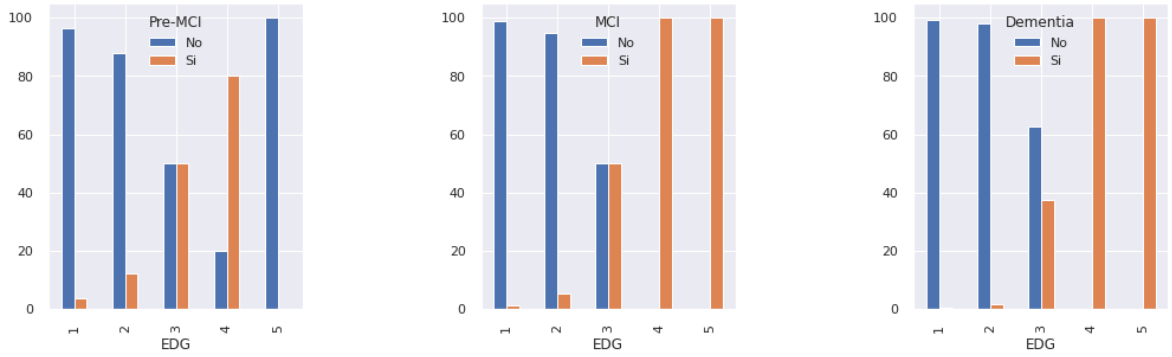

Fig 20. Atributo 'EDG' frente a las tres etapas de la enfermedad de Alzheimer.

De igual manera, cuando el puntaje de EDG es más alto, el paciente se encuentra en una etapa más avanzada de la enfermedad de Alzheimer, por ejemplo, demencia.

Exploración de variables cuantitativas y cualitativas

Esta inspección de variable también se realizó con base a las tres etapas de la enfermedad de Alzheimer. En las Fig 21, Fig 22 y Fig 23 se presentan los atributos ‘Sexo’, ‘Zona’ y ‘Escolaridad’ frente a las tres etapas de la enfermedad de Alzheimer:

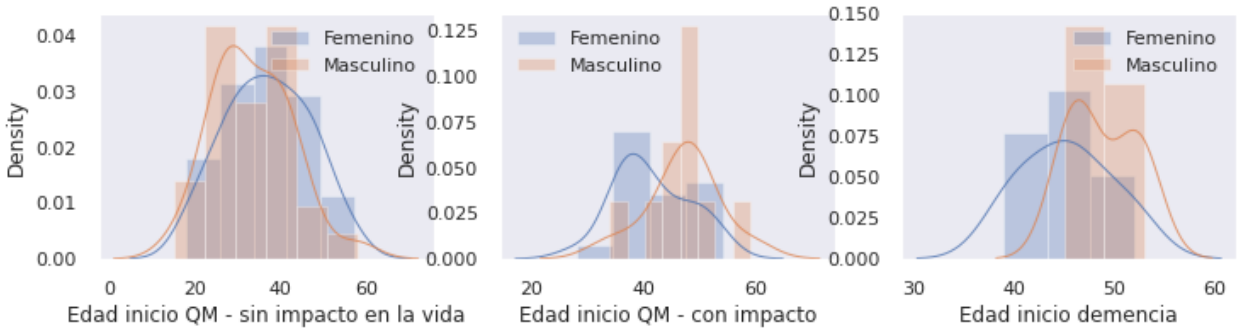

Fig 21. Atributo 'Sexo' frente a las tres etapas de la enfermedad de Alzheimer.

No existe mayor diferencia entre las mujeres y los hombres en la etapa sin impacto en la vida. Sin embargo, si hay una diferencia entre hombres y mujeres que pertenecen a las etapas con impacto y con demencia. Por lo general, las mujeres que se encuentran presentes en estas etapas cuentan con una menor edad que la de los hombres. En la etapa con impacto, la edad media de las mujeres es de aproximadamente 38 años, mientras que la de los hombres es de 48 años. En la etapa con demencia, las mujeres cuentan con una edad aproximada de 45 años y los hombres con una edad aproximada de 50 años. Esto indica que las mujeres presentan síntomas de la enfermedad más rápido que los hombres.

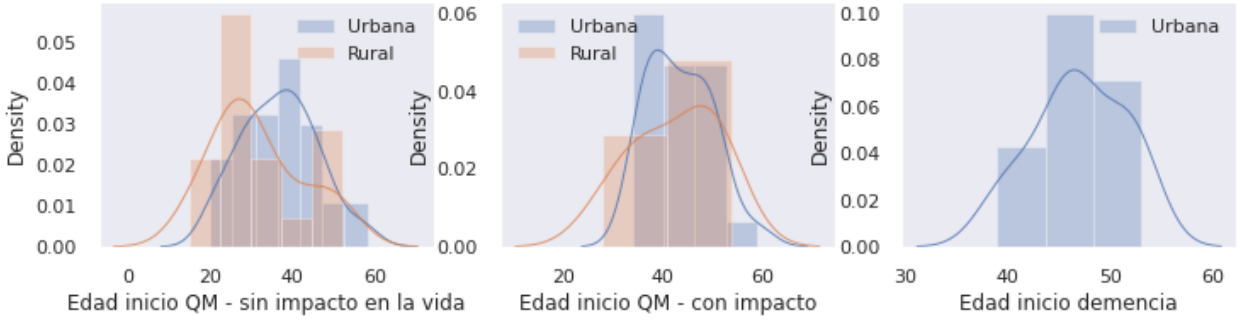

Fig 22. Atributo 'Zona' frente a las tres etapas de la enfermedad de Alzheimer.

En la etapa sin impacto en la vida, por lo general los pacientes de menor edad son los pertenecientes a la zona rural y los de mayor edad a los de la zona urbana. Sin embargo, en las dos etapas siguientes, en las cuales hay presencia de síntomas, no ocurre de la misma manera. En la etapa de inicio con impacto, la media de la edad de los pacientes que viven en la zona urbana es menor que los que viven en la zona rural, lo cual indica que los pacientes que viven en la zona urbana pueden presentar síntomas más rápido que los que viven en la zona rural. En la etapa de inicio de demencia, no existen pacientes pertenecientes a la zona rural. Todos los pacientes con demencia son pertenecientes a la zona urbana con una edad promedio de 47 años.

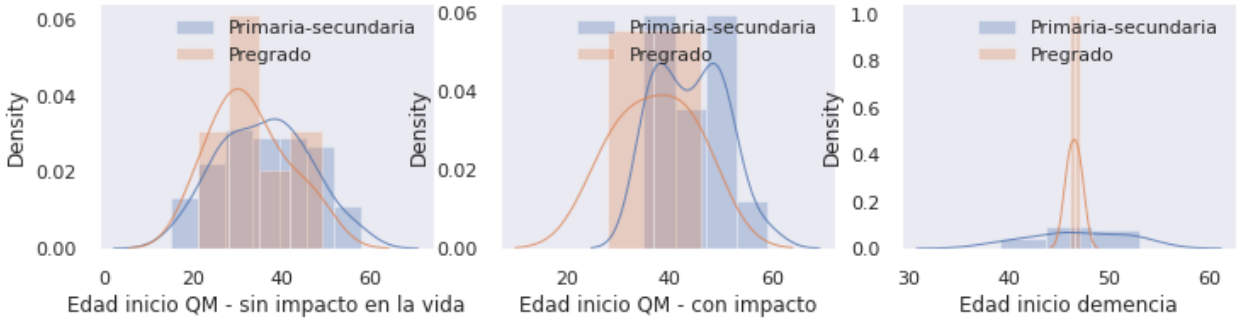

Fig 23. Atributo 'Escolaridad' frente a las tres etapas de la enfermedad de Alzheimer.

El mayor número de años de educación actúa como un modificador de la enfermedad que adelanta la aparición precoz de la enfermedad. Se evidencia que los pacientes con escolaridad alta son los que entran a la etapa con impacto de la enfermedad en edades inferiores a los pacientes con escolaridad baja. Además, en la última etapa que hace referencia a la demencia, prácticamente todos los pacientes tienen escolaridad alta, es decir, una escolaridad superior al bachillerato.

Definición de los segmentos

Validación interna

La validación interna de esta metodología también se evaluó con el análisis de la silueta. Sin embargo, los resultados obtenidos están basados en el algoritmo K-means el cuál fue seleccionado dado que la base de datos conformada no contiene valores atípicos, es decir, estos valores atípicos

fueron imputados por valores que se consideran normales según criterios del personal experto.

En la Fig 24, se expone el método de la silueta:

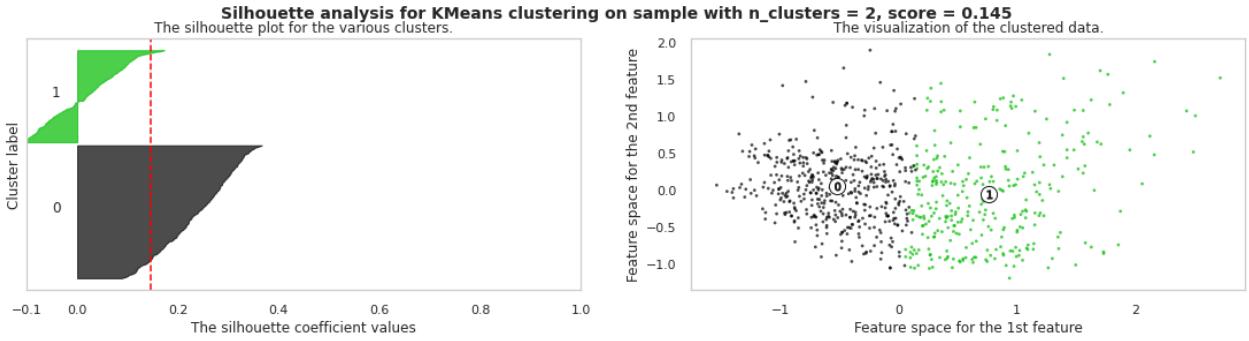

Fig 24. Representación del método Silhouette – Metodología 2.

Se contempla que el score de la métrica de desempeño también fue muy bajo, tan solo fue de 0.145. No obstante, el clúster 0 presenta un score mucho más alto, aproximadamente de 0.4, presenta una geometría definida y no presenta valores negativos. El clúster 1, por el contrario, no presenta una geometría definida y si presenta valores negativos, lo que indica que existen datos que están incorrectamente clasificados, pues están más cerca al clúster 0 que a su propio clúster. También se identifica que el clúster 0 es mucho más grande, esto indica que existen más pacientes clasificados en el clúster 0 que en el clúster 1.

Validación externa

La población de estudio, basados en la metodología 2, también se dividió en dos segmentos como se mostró anteriormente. Para la validación externa de estos dos segmentos también se hizo uso del conjunto de datos del cual se tenía conocimiento y el cual contenía el clúster al que pertenece cada observación.

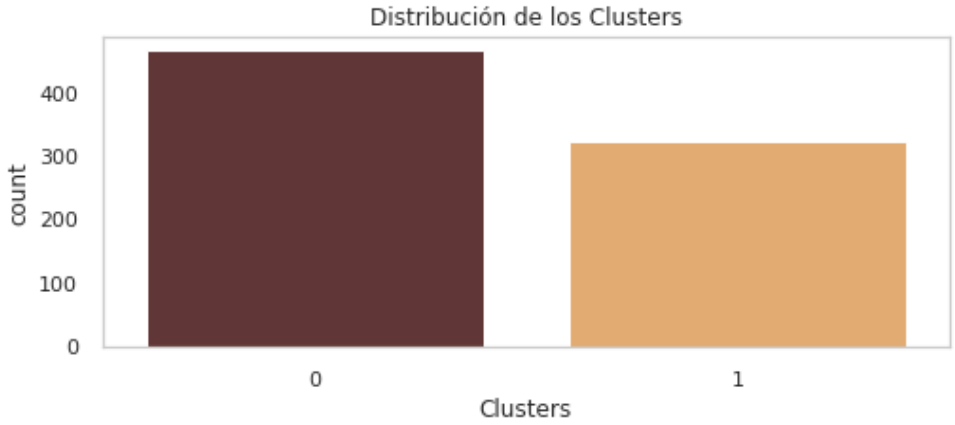

Fig 25. Distribución del clúster 0 y 1 – Metodología 2.

El clúster 0 está conformado por 465 pacientes siendo el 59.08% de la población de estudio total. El clúster 1 está conformado por 322 pacientes siendo el 40.91% de la población de estudio total.

En cuanto a los atributos cuantitativos, en la Tabla 5, se presentan las características más relevantes de los dos segmentos:

Tabla 5. Atributos cuantitativos del clúster 0 y 1 – Metodología 2.

	Clústeres	
	0	**1**
Edad	38,02	46,84
Escolaridad	10,6	4,92
No. años repetidos	0,53	0,57
No. hijos	1,18	2,59
Edad inicio QM - sin impacto en la vida	2,95	5,36
Edad inicio QM - con impacto	0,68	3,31
Edad inicio demencia	0,4	1,75
Presión arterial sistólica	115	121,83
Presión arterial diastólica	73,43	77,43
Depresión	0,11	0,21
Ansiedad	0,17	0,26
Depresión/Disforia	0,43	0,64
Irritabilidad/Liabilidad emocional	0,31	0,4
Sueño/Trastornos del sueño	0,36	0,62
Minimental - Orientación y tiempo	4,89	4,67
Minimental - Orientación y lugar	4,93	4,64
Minimental - Atención y cálculo	4,86	4,4
Minimental - Memoria de evocación	2,79	2,11
Fluidez 0 - 15	8,83	7,01
Fluidez 16 - 30	4,58	3,49
Fluidez 31 - 45	3,38	2,61

Fluidez 46 - 60	2,7	2,08
Denominación - Media	4,44	3,05
Denominación - Baja	3,64	2,68
MLP - Ensayo 1	4,08	2,78
MLP - Intrusiones Ensayo 1	0,4	0,84
MLP - Ensayo 2	6,69	5,03
MLP - Intrusiones Ensayo 2	0,12	0,34
MLP - Ensayo 3	7,94	6,34
MLP - Intrusiones Ensayo 3	0,08	0,25
MLP - Evocación palabras	7,01	4,99
MLP - Intrusiones Evocación	0,17	0,53
MLP - Reconocimiento palabras SI	9,82	9,23
Praxias - Círculo evocación	1,93	1,59
Praxias - Rombo ejecución	2,85	2,59
Praxias - Rombo evocación	2,78	1,99
Praxias - Rectángulos evocación	1,9	1,48
Praxias - Cubo ejecución	3,5	2,29
Praxias - Cubo evocación	3,15	1,11
TMT - tiempo	51,09	102,27
Figura de rey Copia	29,92	23,36
Figura de rey Evocación	18,93	10,19
Figura de rey % de retención	63,2	44,3
Wisconsin aciertos	24,11	15,49
Wisconsin errores	23,85	32,51
Wisonsin categorias	3,18	1,83
Wisconsin Rtas. perseverativas	15,97	25,23
Winsconsin Rtas. nivel conceptual	21,32	12,81
Trastornos memoria QP	10,97	14,79
Yesavage	2,95	3,34
Fast	1,14	1,44
IMC	38,93	39,92

El algoritmo logra identificar las diferencias, en cuanto a los atributos cuantitativos, de los pacientes pertenecientes al clúster 0 como al clúster 1. Por lo general, los pacientes pertenecientes al clúster 1 presentaron resultados menos favorables que los pacientes pertenecientes al clúster 0 en las pruebas médicas realizadas y en los atributos relacionados con la escolaridad y el número de años repetidos. El clúster 1 contiene pacientes, en promedio, con escolaridad menor, mayor número de años repetidos, mayor número de hijos y menor desempeño en las pruebas médicas realizadas que el clúster 0.

No obstante, los atributos relacionados con la edad en la etapa de sin impacto en la vida, con impacto y demencia no muestran valores significativos ya que muchos valores contenidos en

estos atributos eran cero. Por esta razón, se realizó un análisis individual de estos atributos sin contener estos valores iguales a cero. En la Fig 26 se presenta el resultado:

```
Edad inicio QM - sin impacto en la vida
----------------------------------------
Cluster 0: 34.325
Clúster 1: 36.702127659574465

Edad inicio QM - con impacto
----------------------------------------
Cluster 0: 39.625
Clúster 1: 44.458333333333336

Edad inicio demencia
----------------------------------------
Cluster 0: 46.25
Cluster 1: 47.0
```

Fig 26. Edad promedio de inicio en la etapa sin impacto, con impacto y en la demencia de la enfermedad para los pacientes pertenecientes de los clústeres 0 y 1.

En general, se identifica que la edad promedio de inicio de los pacientes pertenecientes al clúster 1 es mayor que la de los pacientes pertenecientes al clúster 0 en las tres etapas de la enfermedad. En la etapa de inicio QM - sin impacto en la vida, la edad promedio de los pacientes pertenecientes al clúster 0 es de 34 años mientras que los pacientes pertenecientes al clúster 1 es de 37 años. En la etapa de inicio QM - con impacto, la edad promedio de los pacientes pertenecientes al clúster 0 es de 39 años mientras que los pacientes pertenecientes al clúster 1 es de 44 años. En la etapa de inicio demencia, la edad promedio de los pacientes pertenecientes al clúster 0 es de 46 años mientras que los pacientes pertenecientes al clúster 1 es de 47 años.

En cuanto a los atributos cualitativos, en las Fig 27 y Fig 28 se presentan las características más relevantes que logró diferenciar los dos segmentos:

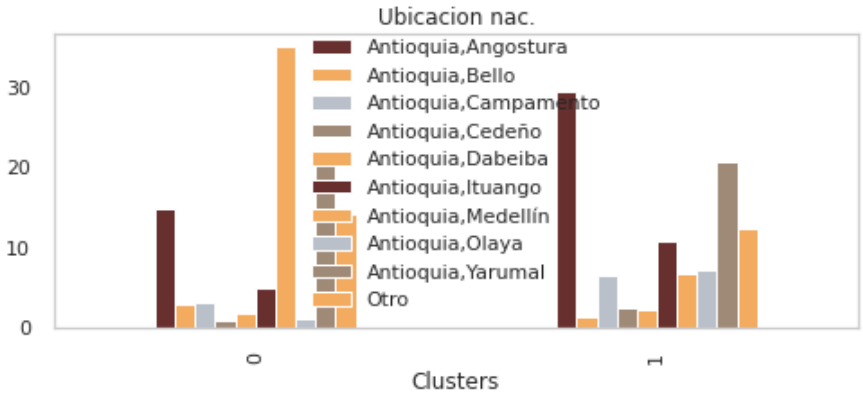

Fig 27. Pacientes del clúster 0 y 1 divididos por su ubicación proveniente.

La mayoría de los pacientes que son pertenecientes al clúster 0 son provenientes de Medellín, Antioquia. Por el contrario, los pacientes pertenecientes al clúster 1 son provenientes, en su mayoría, de Angostura, Antioquia o de Yarumal, Antioquia.

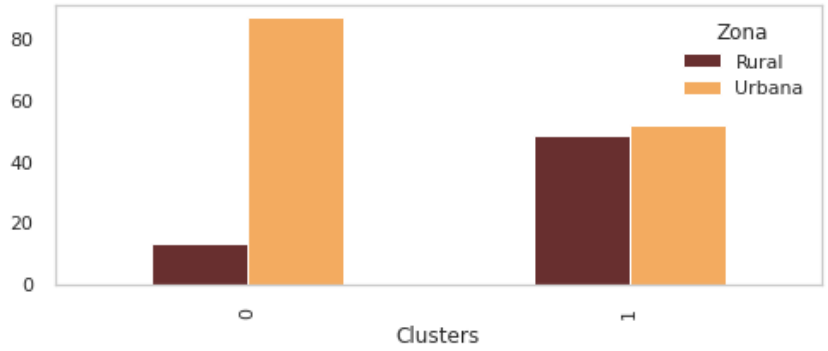

Fig 28. Pacientes del clúster 0 y 1 divididos por su zona proveniente, rural o urbana.

Los pacientes pertenecientes al clúster 1 no se encuentran diferenciados en cuanto a la zona de donde son provenientes. Los pacientes pertenecientes del clúster 0 si se encuentran diferenciados, en este caso, la mayoría de los pacientes del clúster 0 son provenientes de zonas urbanas.

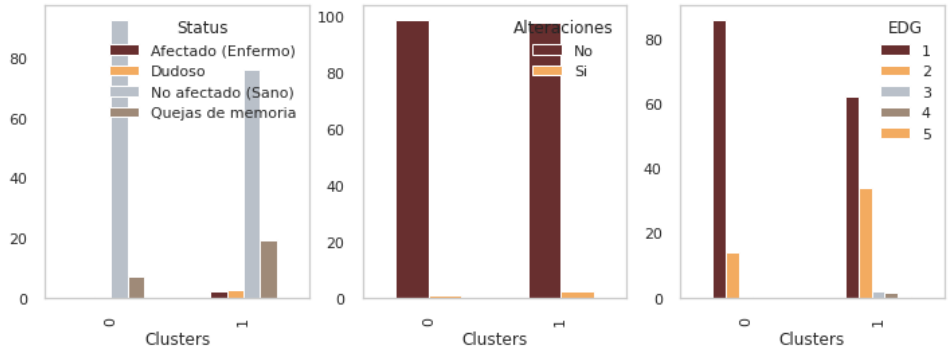

Fig 29. Pacientes del clúster 0 y 1 divididos por Status, Alteraciones y EDG.

En la Fig 29, se evidencia que los pacientes que se encuentran afectados (enfermos) y los que presentan quejas de memoria y alteraciones en la vida cotidiana pertenecen al clúster 1. Además, estos mismos pacientes obtienen mayores puntuaciones en la prueba EDG, por encima de 2. Los pacientes pertenecientes al clúster 0 son personas, en su mayoría, no afectados, no presentan alteraciones en la vida cotidiana y presentan puntuaciones menores en las pruebas EDG, igual o menores a 2.

Además, en la Fig 30 se presenta la cantidad de pacientes pertenecientes al clúster 0 y al clúster 1 con respuesta afirmativa para las tres etapas de la enfermedad:

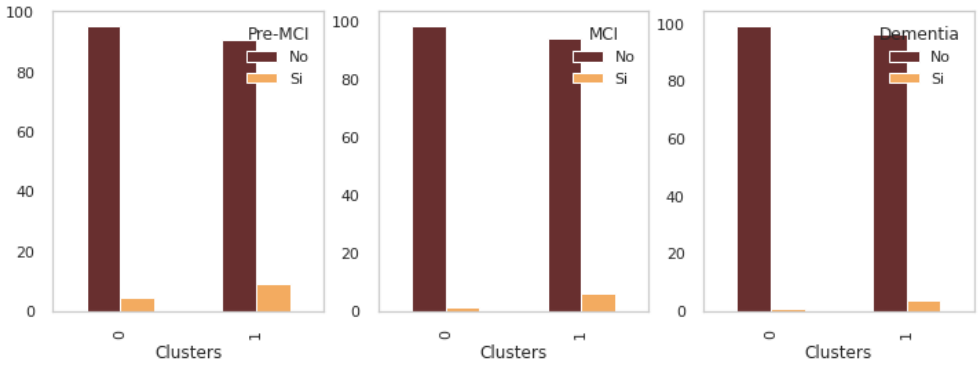

Fig 30. Clasificación de los pacientes por clúster en las tres etapas de la enfermedad.

En la etapa Pre-MCI, las proporciones de la cantidad de pacientes en respuesta afirmativa y negativa para el clúster 0 y clúster 1 son equivalentes. Por el contrario, tanto en la etapa MCI y demencia, la cantidad de pacientes en la respuesta afirmativa disminuye para el clúster 0 y solo

quedan respuestas afirmativas para el clúster 1. Esto indica que la mayoría de los pacientes con respuesta afirmativa en las etapas MCI y demencia pertenecen al clúster 1.

Estratificación de riesgo

Con base a los resultados obtenidos para la definición de los segmentos, cada clúster se define de la siguiente manera:

Población de riesgo elevado: Esta población pertenecen al clúster 1 y cumplen, en gran proporción, estas características:

- Resultados menos favorables en las pruebas neurológicas y neuropsicológicas neurodegenerativas realizadas.
- Resultados menos favorables en el EDG (puntuaciones altas).
- Status afectado (Enfermo) y presenta quejas de memoria.
- Alteraciones de las actividades instrumentales en la vida cotidiana.
- La edad de inicio de la enfermedad en promedio es de 36 años.

Este tipo de pacientes deben recibir atención médica lo más pronto posible debido a que sus características son muy graves con la finalidad de prevenir o retrasar los síntomas de la enfermedad de Alzheimer.

Pacientes de riesgo bajo: Esta población pertenecen al clúster 0 y cumplen, en gran proporción, estas características:

- Resultados más favorables en las pruebas neurológicas y neuropsicológicas neurodegenerativas realizadas.
- Resultados más favorables en el EDG (puntuaciones bajas).
- Status No afectado (Sano).
- La edad de inicio de la enfermedad en promedio es de 34 años.
- Proveniente de zona urbana.
- Proveniente de Medellín, Antioquia.

Con este tipo de pacientes no es necesario prestar atención médica tan rápido como los pacientes pertenecientes al clúster 1 debido a que sus características presentan un nivel de riesgo menor.

Comparando las definiciones realizadas con la variable objetivo, los resultados esperados estaban basados en las siguientes suposiciones:

1. Los pacientes que son portadores de la mutación E280A son las personas que presentan un mayor riesgo y, por ende, deben recibir atención médica lo más pronto. Aquellos pacientes deberían estar clasificados dentro del clúster 1, es decir, en la población de riesgo elevado.
2. Los pacientes que no son portadores de la mutación E280A son las personas que presentan un menor riesgo y, por ende, no es necesario recibir atención médica demasiado rápido. Aquellos pacientes deberían estar clasificados dentro del clúster 0, es decir, en la población de riesgo bajo.

Sin embargo, los resultados obtenidos no coinciden con los resultados esperados como se presenta en la Fig 31:

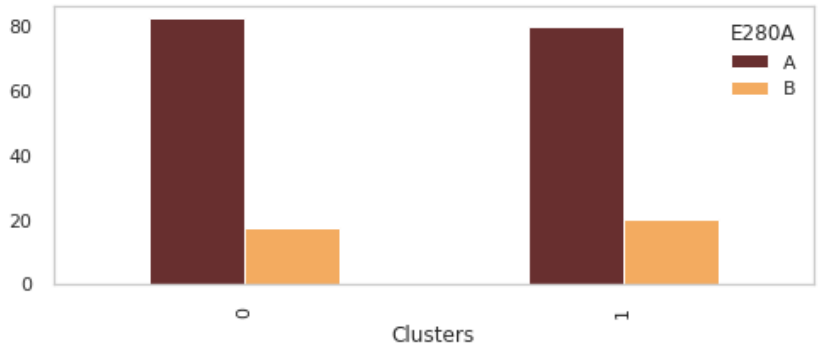

Fig 31. Clasificación de los pacientes del clúster 0 y del clúster 1 en pacientes portadores y no de la mutación E280A.

Tanto el clúster 0 como el clúster 1 incluyen pacientes que son portadores de la mutación E280A en una proporción similar. El clúster 0 tiene 383 pacientes que no son portadores de la

mutación y 82 que si son portadores. El clúster 1 tiene 258 pacientes que no son portadores de la mutación y 64 que si son portadores.

Con la matriz de confusión se observan cuántos pacientes fueron clasificados verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos según la definición anteriormente descrita:

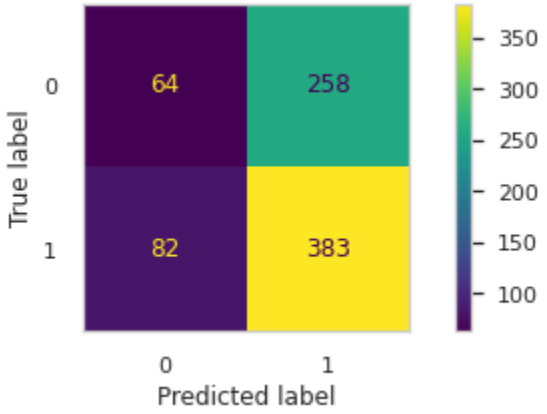

Fig 32. Matriz de confusión para el clúster 0 y 1.

Verdaderos positivos: Existen 64 pacientes que son portadores de la mutación y el algoritmo predijo que el paciente si era portador.

Verdaderos negativos: Existen 383 pacientes que no son portadores de la mutación y el algoritmo predijo que el paciente no era portador.

Falsos positivos: Existen 258 pacientes que no son portadores de la mutación y el algoritmo predijo que el paciente si era portador.

Falsos negativos: Hay 82 pacientes que si son portadores de la mutación y el algoritmo predijo que el paciente no era portador.

Basados en los resultados de la matriz de confusión, se identifica que el algoritmo es muy sensible en la identificación de pacientes que efectivamente no son portadores. Además, es muy

alta la tasa de falsos positivos, lo que indica que quien use el algoritmo para predecir, está prestando atención médica a 269 pacientes que no son portadores lo cual genera gastos dentro del centro médico. Sin embrago, así no sea tan alta la tasa de falsos negativos, es preocupante que 81 pacientes se queden sin atención médica a tiempo cuando realmente sus síntomas son muy graves.

VI. CONCLUSIONES

En este proyecto se planteó el desarrollo de un modelo de segmentación de pacientes con alto riesgo de desarrollar la enfermedad de Alzheimer por una mutación en función de niveles de riesgo basados en algoritmos de clusterización de Machine Learning. Para abordar este objetivo general, se realizaron dos procedimientos diferentes bajo la metodología CRISP-DM de minería de datos. El primer procedimiento estuvo basado en construir una base de datos para la respectiva segmentación sin criterios de personal experto en el área. El segundo procedimiento, a diferencia del primero, tomó en cuenta criterios de personal experto en el área.

En cuanto a los objetivos específicos, con ambos procedimientos se realizó un análisis exploratorio de datos que permitió indagar la base de datos resultante posterior al preprocesamiento e identificar los valores atípicos. Asimismo, después de la imputación de valores atípicos, se obtuvieron dos bases de datos consolidadas, sin registros nulos y con las respectivas variables de segmentación que permitieron desarrollar el algoritmo pertinente para la segmentación de los pacientes. El rendimiento de dichos algoritmos se evaluó mediante una validación interna y externa del modelo diseñado. En la validación externa, ni el primer procedimiento ni el segundo procedimiento lograron segmentar a los pacientes en dos grupos como pacientes portadores de la mutación y pacientes que no son portadores de la mutación E280A. Sin embargo, en la validación interna, solo el segundo procedimiento, bajo criterios de personal experto en el área, cumplió uno de los objetivos requeridos, pues este procedimiento logró segmentar a los pacientes en dos niveles de riesgo: pacientes de riesgo bajo y pacientes de riesgo alto.

Los resultados del proyecto evidenciaron que los pacientes de alto riesgo se caracterizan por obtener resultados menos favorables en las pruebas neurológicas y neuropsicológicas neurodegenerativas y en la prueba EDG. Además, su estado se define como afectado (enfermo), presenta quejas de memoria y alteraciones de las actividades instrumentales en la vida cotidiana. Su edad de inicio de la enfermedad en promedio es de 36 años. Por el contrario, los pacientes de bajo riesgo se caracterizan por obtener resultados más favorables en las pruebas neurológicas y neuropsicológicas neurodegenerativas y en la prueba EDG. Su estado se define como no afectado

(sano) y la edad de inicio de la enfermedad en promedio es de 34 años. Por lo general, este tipo de pacientes de bajo riesgo son provenientes de una zona urbana de Medellín, Antioquia.

Este proyecto es un punto de partida para segmentar a los pacientes en función de niveles de riesgo. Para futuros trabajos se propone continuar explorando la metodología implementada, los modelos y las variables de segmentación que permitan segmentar adecuadamente a las personas con alto riesgo de desarrollar la enfermedad de Alzheimer por la mutación E280A.

REFERENCIAS

[1] P. N. Tariot *et al.*, "The Alzheimer's Prevention Initiative Autosomal-Dominant Alzheimer's Disease Trial: A study of crenezumab versus placebo in preclinical PSEN1 E280A mutation carriers to evaluate efficacy and safety in the treatment of autosomal-dominant Alzheimer's disease, including a placebo-treated noncarrier cohort," *Alzheimer's Dement. Transl. Res. Clin. Interv.*, vol. 4, pp. 150–160, Jan. 2018.

[2] R. G. LanfRanco, P. ManRíquez-navaRRo, L. G. aveLLo, and andRés canaLes-Johnson, "Evaluación de la enfermedad de Alzheimer en etapa temprana: biomarcadores y pruebas neuropsicológicas," *artículo revisión rev Med chile*, vol. 140, pp. 1191–1200, 2012.

[3] Universia, "Para qué sirve Python: qué es y usos," 17-Dec-2020. [Online]. Available: https://www.universia.net/es/actualidad/orientacion-academica/para-que-sirve-phyton-que-es-y-usos-1154393.html. [Accessed: 06-Sep-2022].

[4] datahack, "INTRODUCCIÓN A GOOGLE COLAB PARA DATA SCIENCE - Datahack," 17-Jun-2019. [Online]. Available: https://www.datahack.es/google-colab-para-data-science/. [Accessed: 06-Sep-2022].

[5] A. Barrera-Ocampo, F. Lopera, A. Barrera-Ocampo, and F. Lopera, "Amyloid-beta immunotherapy: the hope for Alzheimer disease?," *Colomb. Med.*, vol. 47, no. 4, pp. 203–212, 2016.

[6] A. Puche, J. F. Ochoa-Gómez, Y. Agudelo, J. Rodas, and C. Tobon, "Alteración de la entropía en la precuña y la corteza cingulada posterior en la enfermedad de Alzheimer: estudio de resonancia magnética funcional en reposo," *TecnoLógicas*, vol. 24, Dec. 2021.

[7] "Roche presenta datos actualizados del estudio de la Iniciativa para la Prevención del Alzheimer que evalúa crenezumab en Enfermedad de Alzheimer autosómica dominante de inicio temprano | Roche España." [Online]. Available: https://www.roche.es/es_es/comunicacion/actualidad/2022/junio/roche-presenta-datos-actualizados-del-estudio-de-la-iniciativa-p.html. [Accessed: 05-Sep-2022].

[8] S. J. Romero -Vanegas, C. M. Valencia-Marín, D. C. Aguirre-Acevedo, H. Buschke, and F. Lopera, "Alteraciones de la memoria episódica verbal en fases preclínica y temprana de la enfermedad de Alzheimer familiar precoz por mutación E280A en PS1," 2010.

[9] A. Zuno-Reyes, "Efecto de la Escolaridad, Edad y Género en el desempeño del protocolo

neuropsicológico CERAD en el Estado de Jalisco," 2016.

[10] S. Czapka and J. Festman, "Wisconsin Card Sorting Test reveals a monitoring advantage but not a switching advantage in multilingual children," *J. Exp. Child Psychol.*, vol. 204, p. 105038, Apr. 2021.

[11] A. M. Sánchez, K. Arévalo, M. Vallecilla, M. C. Quijano, and J. Arabia B, "La memoria audioverbal en adultos mayores con deterioro cognitivo leve y un grupo control," *CES Psicol.*, vol. 7, no. 1, pp. 35–47, 2014.

[12] J. Carvajal-Castrillón, D. C. Aguirre Acevedo, and F. J. Lopera Restrepo, "Perfil clínico y cognitivo de la atrofia cortical posterior y sus diferencias con la enfermedad de Alzheimer esporádica tardía y familiar precoz," *Acta Neurológica Colomb.*, vol. 26, no. 2, pp. 75–86, 2010.

[13] S. Julieta and R. Vanegas, "VALIDACIÓN DEL INSTRUMENTO 'ALZHEIMER`S DISEASE ASSESMENT SCALE' (ADAS) PARA POBLACIONES CON ENFERMEDAD DE ALZHEIMER Y DETERIORO COGNOSCITIVO LEVE EN COLOMBIA.," 2014.

[14] L. I. Santamaría and C. P. Jiménez, "Características neuropsicológicas y neurofisiológicas de la memoria de trabajo del traductor," p. 169, 2013.

[15] "¿Qué es Machine Learning y para qué sirve?" [Online]. Available: https://www.inbest.cloud/comunidad/que-es-machine-learning-y-para-que-sirve. [Accessed: 20-Feb-2022].

[16] H. R. Gonzáles Argote and U. A. Ticona Gonzáles, "Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means," *Rev. Latinoam. Desarro. Económico*, Nov. 2019.

[17] J. Sebastián and R. Gómez, "Evaluación de algoritmos de aprendizaje de máquina no supervisados sobre datos climáticos," 2019.

[18] "Selecting the number of clusters with silhouette analysis on KMeans clustering — scikit-learn 1.1.2 documentation." [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. [Accessed: 06-Sep-2022].

[19] Biogen, "Conceptos básicos de los ensayos clínicos," *Cure SMA*, 2017.

[20] J. Espinosa, "Aplicación de metodología CRISP-DM para segmentación geográfica de una

base de datos pública," *Scielo*, 2020.

[21] A. Villegas, M. M. Castañeda, and L. F. Arias, "Evaluación de la producción de β-amiloide por la mutación E280A en el gen de la presenilina," *Biomédica*, vol. 27, pp. 372–84, 2007.