UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 2 - Análisis de Datos

Integrantes: Matías Figueroa Contreras

Curso: Análisis de Datos

Profesor: Max Chacón Pacheco

Ayudante: Marcelo Álvarez

12 de Mayo de 2024

Tabla de contenidos

1. Introducción

El síndrome clínicamente aislado (SCA) es un primer episodio de síntomas neurológicos que puede o no evolucionar hacia la esclerosis múltiple (EM), una enfermedad crónica del sistema nervioso central que provoca una serie de síntomas físicos y mentales («Esclerosis múltiple - Síntomas y causas», s.f.). La capacidad para distinguir entre los pacientes de SCA que eventualmente desarrollarán EM de aquellos que no lo harán es crucial, dada la diversidad de síntomas iniciales y lo imprevisible de la enfermedad. Estudiar cómo evolucionan los pacientes con SCA a lo largo del tiempo es fundamental para comprender los factores de riesgo y las características clínicas que pueden influir en su progresión hacia la EM, ya que esto permite implementar intervenciones más tempranas en aquellos pacientes con alto riesgo de desarrollar la enfermedad («Síndrome Clínicamente Aislado (CIS)», s.f.). Para este estudio, en este laboratorio se trabaja con el algoritmo de clustering K-Prototypes, que permite agrupar eficazmente datos que incluyen tanto variables numéricas como categóricas Huang, 1998. Esto proporciona una ventaja al analizar los datos a trabajar «Conversion Predictors of CIS to Multiple Sclerosis», 2023, ya que permite capturar la diversidad de características demográficas, síntomas iniciales y resultados médicos que pueden influir en la progresión del SCA a esclerosis múltiple (EM).

1.1. Objetivos

- Realizar limpieza, imputación y normalización de datos para asegurar la integridad y coherencia de los mismos.
- Evaluar y determinar el numero de clústers a trabajar.
- Identificar patrones distintivos en los datos que puedan predecir la progresión de pacientes con SCA a EMCD.
- Obtener conocimiento que ayude a comprender mejor la relación entre el SCA y la EMCD, así como caracterizar las trayectorias de progresión de la enfermedad.

2. Marco Teórico

2.1. Clustering

El clustering o agrupamiento es una técnica de aprendizaje no supervisado que busca agrupar objetos en subconjuntos llamados clústers, donde los objetos dentro de cada grupo son más similares entre sí en comparación con los de otros grupos. Estos grupos se forman según la similitud de características, usando métricas como la distancia euclidiana. Entre los algoritmos más comunes de clustering están K-means, que divide los datos en K grupos con centroides cercanos; clustering jerárquico, que forma jerarquías de clusters mediante métodos aglomerativos o divisivos; y DBSCAN, que identifica conglomerados basados en densidades y aisla puntos atípicos Institute, 2023.

2.2. Centroide

En el contexto del clustering, el centroide es un punto representativo de un clúster. En los clústers basados en datos numéricos, como en K-Means o K-Prototypes, el centroide es típicamente la media de todas las observaciones en el clúster («Centroid», s.f.).

2.3. K-Prototypes

K-Prototypes es una variante del algoritmo K-Means que se utiliza para agrupar conjuntos de datos que contienen tanto variables numéricas como categóricas. En lugar de calcular la distancia euclidiana entre observaciones, K-Prototypes utiliza una combinación de la distancia euclidiana para variables numéricas y una medida de disimilitud apropiada para variables categóricas (Huang, 1998).

2.4. Disimilitud Euclidiana

La disimilitud euclidiana es una medida de la distancia entre dos puntos en un espacio euclidiano. En el contexto del clustering, se utiliza para calcular la distancia entre observaciones que contienen variables numéricas («Euclidean distance», s.f.).

2.5. Disimilitud por Coincidencia

La disimilitud por coincidencia, también conocida como coincidencia ocupada, es una medida de disimilitud utilizada para variables categóricas. Cuenta el número de atributos coincidentes entre dos observaciones categóricas y se utiliza para calcular la distancia entre ellas («Overlap distance», s.f.).

2.6. Método del Codo

El método del codo es una técnica utilizada para determinar el número óptimo de clústers en un conjunto de datos. Se basa en trazar el valor de la función de costo del clustering en función del número de clústers y observar el punto donde se observa una disminución significativamente menor en la función de costo (Mobility, 2019).

2.7. Método de la Silueta

El método de la silueta es una medida de la cohesión y separación de los clústers en un conjunto de datos. Proporciona una medida de qué tan similar es una observación a su propio clúster en comparación con otros clústers (Mobility, 2019).

2.8. Índice de Davies-Bouldin

El índice de Davies-Bouldin es una medida de la calidad de los clústers en un conjunto de datos. Se calcula como la media de la similitud entre cada clúster y su clúster más similar, dividido por la distancia entre los centroides de los clústers (Davies y Bouldin, 1979).

3. Desarrollo

3.1. Pre-procesamiento

Antes de proceder a la generación de clústers, es esencial realizar un pre-procesamiento de los datos. Durante esta fase, se llevaron a cabo las operaciones detalladas a continuación:

3.1.1. Limpieza de datos

La limpieza de datos comenzó con la identificación de valores faltantes en el conjunto de datos. Encontrando que las variables *Schooling*, *Initial_Symptom*, *Initial_EDSS* y *Final_EDSS* presentan ausencias de datos, siendo mas notable la cantidad de datos faltantes para las últimas dos variables. Basándose en estas observaciones, se efectuaron las siguientes acciones:

(A) **Imputación de valores**: Para las variables *Schooling* e *Initial_Symptom*, se optó por utilizar estrategias de imputación basadas en la naturaleza de cada variable, con el objetivo de mantener la integridad del conjunto de datos evitando la pérdida de información valiosa contenida en las filas de los datos faltantes. Además, es poco probable que la imputación de datos introduzca un sesgo significativo, ya que la cantidad de datos faltantes es baja en comparación con el total de datos. Esta decisión se encuentra respaldada por la literatura (Dong y Peng, 2013), que sugiere lo dicho anteriormente. A continuación se menciona las técnicas usadas para cada variable:

- *Schooling*: Se imputa el dato faltante usando la mediana de la variable, dado que es una variable numérica entera y proporciona una medida central menos susceptible a valores extremos.
- *Initial_Symptom*: Se imputa el dato faltante usando la moda de la variable, dado que es una variable categórica y esta medida representa la categoría más frecuente en la muestra.

(B) **Eliminación de columnas**: Se decidió excluir las columnas *Initial_EDSS* y *Final_EDSS* del análisis, dado que el grupo 2 (no EMCD) tenia una ausencia completa

de datos para estas variables. Lo anterior debido a que estas representan una escala de discapacidad específicamente aplicada a pacientes con esclerosis múltiple clínicamente definida (EMCD) (Trust, 2022), y su ausencia en el grupo no EMCD se debe a que estos individuos no desarrollaron EMCD y por lo tanto no se le realizaron estos estudios, para monitorear el desarrollo de la enfermedad. Incluir estos datos mediante imputación artificial o utilizarlos de forma incompleta podría introducir un sesgo significativo en el análisis. La omisión de estas variables se justifica porque no se pueden estimar las medidas de discapacidad en pacientes que no han progresado a EMCD. Así, para preservar la integridad y precisión del análisis de clustering, que busca identificar patrones predictivos de progresión de SCA a EM, es adecuado excluir estas variables y centrarse en otras características clínicas y demográficas comunes a ambos grupos.

3.1.2. Normalización de datos

Se aplicó una normalización a las variables numéricas *Age* y *Schooling*, con el fin de asegurar que estas tengan la misma escala y no se vean afectadas por la magnitud de los valores al momento de realizar el clustering. Este enfoque convierte los valores a un rango común, facilitando la comparación y análisis posteriores.

3.2. Obtención de Clústers

Para la obtención de clústers, se ha seleccionado el algoritmo de K-Prototypes. Esta decisión se fundamenta en las características del algoritmo, que lo hacen idóneo para manejar conjuntos de datos que incluyen tanto variables categóricas como numéricas. El algoritmo de K-Prototypes es una extensión del algoritmo K-Means, diseñado específicamente para abordar esta diversidad de datos (Huang, 1998). La elección de K-Prototypes se justifica por su capacidad para asignar clústers de manera efectiva en el conjunto de datos, que presenta una combinación de variables numéricas y categóricas. Esta metodología permite una mejor segmentación comparada con otros algoritmos que manejan un solo tipo de variable, asegurando una mayor precisión y relevancia en los resultados obtenidos. Por tanto, el uso de K-Prototypes permitirá extraer y agrupar características más profundas y representativas del conjunto de datos analizado, facilitando la toma de decisiones basada en un entendimiento

más completo e integrado de las variables involucradas.

3.2.1. Métricas de disimilitud

El algoritmo de K-Prototypes emplea dos métricas de disimilitud distintas para tratar de manera efectiva los tipos de datos mixtos presentes en el conjunto de datos, de los cuales se escogieron las dos métricas estándar que utiliza el algoritmo:

- **Disimilitud Euclidiana para valores Numéricos:** La disimilitud euclidiana es una de las métricas más comunes para medir la distancia entre puntos en un espacio numérico. Se define como la raíz cuadrada de la suma de las diferencias cuadradas entre los puntos correspondientes de dos vectores. Esta métrica es especialmente útil en datos numéricos porque refleja la magnitud de la diferencia entre los puntos, ofreciendo una representación precisa de su distancia relativa en un espacio multidimensional (Harmouch, 2023). Su uso en K-Prototypes asegura que los clústers reflejen agrupaciones naturales en los datos numéricos, basadas en la proximidad euclidiana.

- **Disimilitud por Coincidencia para valores Categóricos:** La disimilitud por coincidencia ocupada, es efectiva para datos categóricos porque cuenta el número de coincidencias entre dos vectores. Esta métrica simplemente suma una unidad por cada discrepancia entre categorías correspondientes de los vectores (Huang, 1998). La aplicación de esta métrica es crucial para los datos categóricos dentro de K-Prototypes, ya que no existe un orden o una magnitud de diferencia que pueda ser medido de forma numérica. Por lo tanto, la disimilitud por coincidencia ocupada proporciona un método adecuado para evaluar qué tan similares o diferentes son dos objetos categóricos.

3.2.2. Método de inicialización

Como método de inicialización del algoritmo se selecciono el método *Cao*. La elección de este se basa en su capacidad para proporcionar una inicialización de centroides más efectiva y robusta, especialmente en conjuntos de datos con características mixtas (numéricas y categóricas). Este método fue desarrollado por Cao et al. y está diseñado específicamente para optimizar la calidad de los clústers iniciales en el contexto de datos mixtos (Cao et al.,

2009). El método Cao evalúa tanto la frecuencia como la densidad de los datos en el espacio de características, determinando los centroides iniciales basándose en un criterio que maximiza la densidad de datos alrededor de un centroide propuesto (Cao et al., 2009). Por lo tanto, la elección del método de *Cao* está justificada por su eficacia en tratar con la complejidad y diversidad de datos mixtos, ofreciendo una base sólida y confiable para el análisis de clústers. Esto garantiza que el proceso de clustering no sólo sea eficiente sino también robusto frente a las variaciones en los datos de entrada.

3.2.3. Selección de numero de Clústers

Para establecer el número adecuado de clústers, se emplearon los métodos del codo y de la silueta, evaluando posibles configuraciones que varían entre 2 y 9 clústers. Los resultados de estos análisis están presentes en las imágenes 1 y 2, donde se observa lo siguiente:

- **Método del Codo:** La gráfica correspondiente muestra una disminución en la tasa de descenso del costo al incrementar el número de clústers. Notablemente, se observa una desaceleración pronunciada alrededor de los 4 clústers, sugiriendo que esta podría ser una cantidad óptima de clústers para nuestro conjunto de datos.
- **Método de la Silueta:** Este revela una tendencia general de disminución en el puntaje de silueta con el aumento de clústers. Principalmente mostrando la disminución más pronunciada después del 4to clúster, indicando que 4 clústers también podrían proporcionar una segmentación eficaz de los datos.

Basado en las observaciones de ambos métodos, la selección de 4 clústers es la opción más consistente. Esta cantidad no solo coincide con el çodo.ºbservado en el gráfico del Método del Codo, sino que también se alinea con un pico en el puntaje de silueta. Esto sugiere un equilibrio óptimo entre una buena partición de los datos y mantener el costo relativamente bajo. Por lo tanto, se concluye que 4 clústers es el número más adecuado para proceder con el análisis.

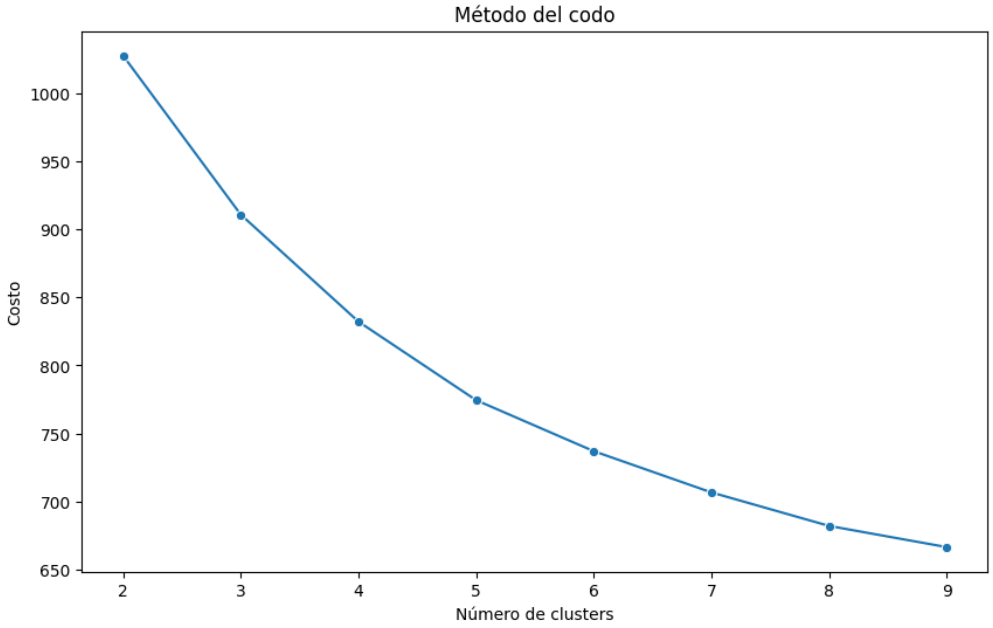


Figura 1: Gráfico método del codo

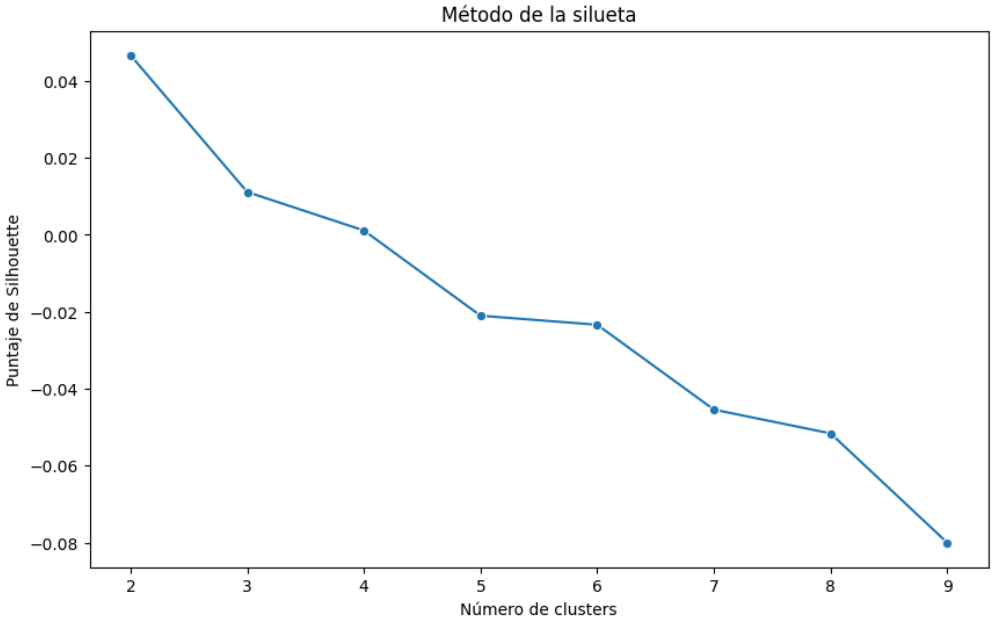


Figura 2: Gráfico método de la silueta

3.2.4. Evaluación de calidad de los Clústers

En el análisis de clústers realizado, se aplicaron dos criterios clave para evaluar la calidad de los clústers formados mediante el algoritmo K-Prototypes. A continuación, se presentan los criterios utilizados junto con los resultados obtenidos:

Puntaje de Silueta: Este criterio evalúa la cohesión interna y la separación entre clústers y fue observado previamente en la sección de selección de numero de clústers. Un valor cercano a +1 indica que los clústers están bien separados y que los elementos dentro de un clúster son similares entre sí, mientras que un valor cercano a 0 sugiere que los clústers se superponen («sklearn.metrics.silhouette_score», 2024). En nuestro caso, el puntaje de silueta obtenido fue de 0.001114201681195495, lo cual indica que los clústers formados tienen una baja distinción entre ellos.

Índice Davies-Bouldin: Este índice mide la calidad de los clústers basándose en la relación entre la dispersión interna de los clústers y su separación mutua. Un valor bajo en este índice indica clústers que están bien separados y que tienen una baja variación interna («sklearn.metrics.davies_bouldin_score», 2024). El índice Davies-Bouldin obtenido fue de 3.938712282767227, indicando una calidad de clústers subóptima.

Los resultados de estos criterios sugieren que, aunque el algoritmo de K-Prototypes ha sido capaz de formar clústers, la calidad de estos clústers podría mejorarse. La superposición significativa y la falta de separación clara entre los clústers indicada por el bajo puntaje de silueta y el alto índice Davies-Bouldin podría ser un área para explorar mejoras en el proceso de clustering. Dado lo anterior se exploraron otros algoritmos, incluyendo la transformación de los datos numéricos para utilizar el algoritmo K-Modes, sin encontrar mejoras significativas en los índices mencionados.

3.3. Análisis de Resultados

3.3.1. Distribución de los Clústers respecto a EMCD

Partiendo con el análisis de decide observar como se separaron los clústers con respecto a la variable *group*, la cual menciona si el individuo con SCA desarrollo o no EMCD. Según el gráfico observado en la imagen 3 se tiene lo siguiente para cada clúster:

- **Clúster 0:** Tiene una proporción significativa de casos sin EMCD (20.5 %) comparado con su contraparte con EMCD (12.1 %).
- **Clúster 1:** Muestra una baja prevalencia de casos sin EMCD (1.8 %) comparado con los casos con EMCD (19.0 %).
- **Clúster 2:** Exhibe una cantidad moderada de muestras sin EMCD (6.6 %) en relación a los con EMCD (10.3 %).
- **Clúster 3:** Presenta una alta proporción de casos sin EMCD (25.3 %) en comparación con los casos con EMCD (4.4 %).

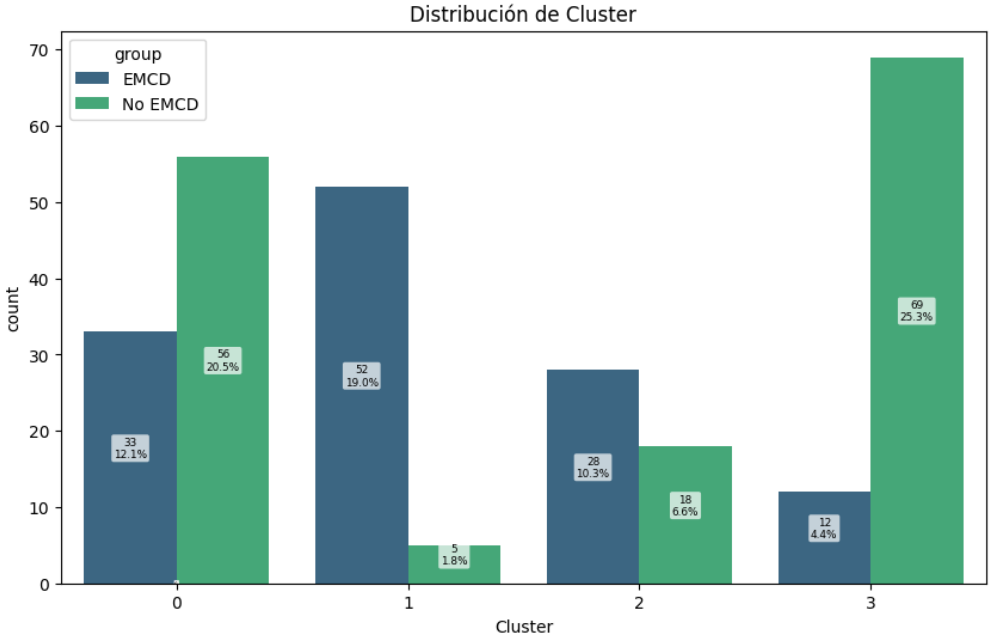


Figura 3: Distribución de Clúster con group

3.4. Análisis de variables numéricas

Para las variables numéricas *Age* y *Schooling* se generaron las tablas resumen 1 y 2 sobre distintos estadísticos para cada clúster, y los graficos sobre la distribución de estos 4a y 4b, a partir de esto se realizo el siguiente análisis para cada clúster:

- **Clúster 0:** Esta compuesto principalmente por individuos con una edad media de 28.42 años y una escolaridad media de 19.02 años, muestra una concentración de jóvenes adultos con educación superior (rango escolar de 15-25 años). Este presenta una mayor cantidad de casos sin EMCD, aunque no tan grande como en otros grupos. El nivel educativo más alto podría tener correlación con un mejor manejo de la salud y acceso a recursos médicos, lo que podría ayudar a mitigar el riesgo de desarrollar EMCD. Según la literatura, el alto nivel de educación podría asociarse con mejores hábitos de salud y mayor conciencia sobre el manejo de condiciones crónicas (UNESCO, 2019). Por otro lado, la edad de los individuos la cual es relativamente baja sugiere que la concentración de jóvenes adultos puede representar un mejor perfilamiento en relación al desarrollo de SCA a EMCD, lo cual no coincide con lo establecido por la literatura («Esclerosis múltiple - Síntomas y causas», s.f.), la cual sugiere que esta se diagnostica generalmente entre 20 a 40 años, esto se podría deber a que este análisis es preliminar solo con las variables numéricas.

- **Clúster 1:** muestra una media de edad similar al Clúster 0 (27.93 años), pero con una escolaridad significativamente más baja (media de 15.02 años). este clúster muestra una alta prevalencia de EMCD. En relación con la escolaridad, se puede señalar que coincide con lo establecido en el clúster 0 sobre esta medida, ya que a diferencia de el clúster 0 este presenta rangos y una media más baja. Por otro lado, con respecto a la edad, este si se coincide con lo establecido por la literatura («Esclerosis múltiple - Síntomas y causas», s.f.) a diferencia del Clúster 0.

- **Clúster 2:** Con una edad promedio de 51.91 años y una escolaridad variada (media de 14.74 años), el Clúster 2 abarca un rango de edad más amplio, extendiéndose hasta el caso particular de 77 años. Este clúster no tiene una gran diferencia entre los grupos que desarrollaron o no EMCD, aun asi existe una leve inclinación sobre el desarrollo de EMCD, lo que no se coincide con la literatura, ya que como se explico anteriormente la EMCD se diagnostica generalmente entre 20 a 40 años.

- **Clúster 3:** A pesar de tener la más baja escolaridad media (11.31 años) y una edad media de 34.44 años, este clúster muestra la menor cantidad de desarrollo de EMCD. Este

hallazgo puede indicar que, aunque la educación es baja y el rango de edad esta entre la edad común de diagnostico de EMCD («Esclerosis múltiple - Síntomas y causas», s.f.), otros factores podrían estar afectando la composición de este clúster, de todas maneras se estudiara en la siguientes secciones, ya que según la literatura lo que indica que la baja escolaridad puede aumentar el riesgo de enfermedades debido a factores como el menor conocimiento de las prácticas de salud preventiva (UNESCO, 2019).

Tabla 1: Datos de Edad por Cluster

Cluster	median	mean	std	min	max
0	28.0	28.41573	7.086814	16	46
1	28.0	27.929825	6.627485	15	43
2	50.0	51.913043	7.363049	41	77
3	36.0	34.444444	7.607562	17	51

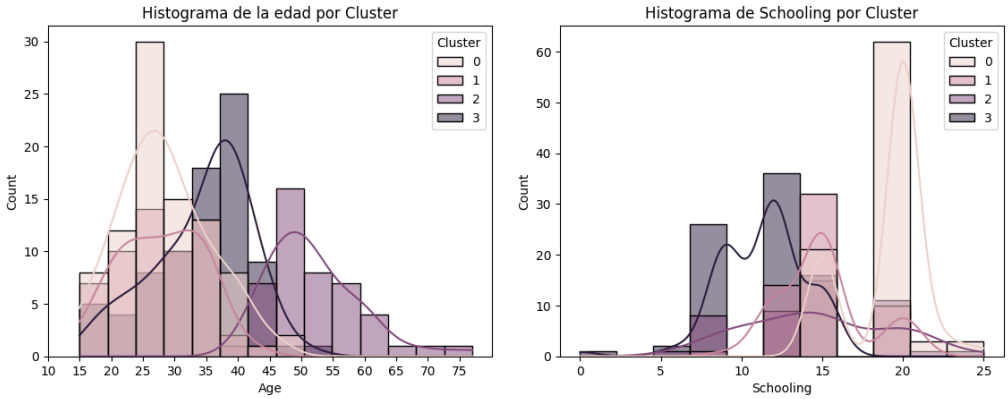
Tabla 2: Datos de Schooling por Cluster

Cluster	median	mean	std	min	max
0	20.0	19.022472	2.388312	15	25
1	15.0	15.017544	2.715637	9	20
2	15.0	14.739130	4.399165	6	25
3	12.0	11.308642	2.639327	0	15

3.5. Análisis de variables categóricas

Para las variables categóricas se generaron las tablas 3, 4 y 5 con el valor más frecuente para cada clúster y variable, además se generaron los gráficos 5, 6, 7, 8 y 9 para analizar las distintas variables:

- **Clúster 0:** El Clúster 0 está compuesto mayoritariamente por mujeres jóvenes, muchas de las cuales no han desarrollado EMCD, con solo un 12,1 % reportando la enfermedad,



(a) Histograma edad por Clúster (b) Histograma Schooling por Clúster

en comparación con un 20,5 % que no la ha desarrollado. Este grupo se caracteriza por tener antecedentes de lactancia materna y exposición a varicela, factores que pueden haber contribuido a un sistema inmunológico más robusto y posiblemente a una menor susceptibilidad a desarrollar EMCD. La prevalencia de síntomas mayoritariamente polisintomáticos sugiere una variedad de manifestaciones clínicas. Los resultados diagnósticos son en su mayoría negativos para bandas oligoclonales, LLSEP, ULSEP, VEP, y BAEP, lo que indica una baja actividad de enfermedad neurológica, en línea con la baja prevalencia de EMCD observada. Además, los resultados de MRI en áreas periventricular, cortical, infratentorial y medular son mayoritariamente negativos, lo que corrobora una afectación neurológica mínima.

- **Clúster 1:** Este clúster se distingue por tener una alta proporción de hombres que han desarrollado EMCD, con un 19 % de los casos confirmados frente a un pequeño 1.8 % que no ha desarrollado la enfermedad. Este grupo muestra una alta prevalencia de síntomas polisintomáticos, que indican una forma más agresiva y progresiva de EMCD, dado el desarrollo de múltiples síntomas. Los participantes en este clúster tienden a tener una menor historia de exposición a varicela en comparación con otros clústers, lo que influye al desarrollo de la EMCD según se menciona en la literatura (Bermudez et al., 2016). A pesar de que muchos fueron amamantados, lo que generalmente contribuye a un mejor sistema inmunológico (Sara Collorone, 2022), este factor no parece haber sido suficiente para contrarrestar el desarrollo de la enfermedad en este grupo. En cuanto a las pruebas

diagnósticas, hay una notable cantidad de resultados positivos en bandas oligoclonales, así como en pruebas de LLSEP y ULSEP, lo que refleja una significativa actividad inflamatoria y daño neurológico. Las resonancias magnéticas (MRIs) de este clúster muestran también una alta incidencia de resultados positivos en las áreas periventricular y cortical, indicando lesiones neurológicas extensas y activas, que son características de una enfermedad avanzada y activa (Peñailillo et al., 2019).

- **Clúster 2:** El Clúster 2 es notable por su diversidad en la presentación de síntomas y resultados de pruebas, compuesto en su mayoría por mujeres. La cantidad de personas que desarrollaron EMCD en este grupo es del 10.3 %, con un 6.6 % de individuos que no han desarrollado la enfermedad, indicando una distribución moderada de la enfermedad en comparación con otros clústers. Este grupo muestra una equilibrada mezcla de antecedentes de varicela, por lo que no es determinante para este clúster. Esta variabilidad también se ve reflejada en los resultados de pruebas diagnósticas, donde hay una distribución mixta en bandas oligoclonales, LLSEP, ULSEP, VEP y BAEP, lo que indica diferencias individuales en la actividad de la enfermedad y la afectación neurológica. Los resultados de las resonancias magnéticas (MRI) en este clúster muestran una diversidad similar, con una proporción parecida en resultados positivos y negativos en las áreas periventricular, cortical e infratentorial, y en la médula espinal.

- **Clúster 3:** Está compuesto exclusivamente por mujeres y se caracteriza por tener la menor prevalencia de desarrollo de SCA a EMCD, con solo un 4,4 % de los casos reportados como positivos frente a un 25,3 % que no presentan el desarrollo de la enfermedad, lo que no coincide con la literatura, la cual menciona que las mujeres tienden a desarrollar más EMCD («Esclerosis múltiple - Síntomas y causas», s.f.). Por otro lado, este grupo muestra una predominancia de síntomas polisintomáticos como todos los anteriores. Además, este clúster tiene una baja incidencia de antecedentes de varicela, lo que podría sugerir una menor exposición a ciertos estímulos inmunológicos tempranos que afecten el desarrollo de EMCD (Bermudez et al., 2016) como se menciono anteriormente. En términos de resultados diagnósticos, la mayoría de las pruebas, incluidas las bandas oligoclonales, LLSEP, ULSEP, VEP, y BAEP, son predominantemente negativas, lo

que indica una actividad limitada de la enfermedad en el sistema nervioso central. Las resonancias magnéticas (MRIs) corroboran estos hallazgos, mostrando principalmente resultados negativos en las áreas periventricular, cortical, infratentorial y de la medula espinal, lo que hace sentido según la literatura, que menciona que estos exámenes son concluyentes sobre el desarrollo de EMCD (Peñailillo et al., 2019).

Cluster	**Gender**	**Breastfeeding**	**Varicella**	**Initial_Symptom**	**Mono_or_Polysymptomatic**
0	Femenino	Si	Positivo	Sensorial y Motor	Polisintomático
1	Masculino	Si	Negativo	Motor	Polisintomático
2	Femenino	Si	Negativo	Sensorial y Motor	Polisintomático
3	Femenino	Desconocido	Negativo	Visual	Polisintomático

Tabla 3: Resumen variables categóricas parte 1

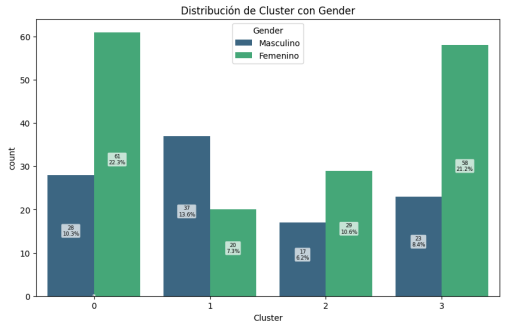


Figura 5: Distribución de variable Gender

Cluster	**Oligoclonal_Bands**	**LLSSEP**	**ULSSEP**	**VEP**	**BAEP**
0	Negativo	Negativo	Negativo	Negativo	Negativo
1	Negativo	Positivo	Positivo	Positivo	Negativo
2	Negativo	Positivo	Positivo	Negativo	Negativo
3	Negativo	Negativo	Negativo	Negativo	Negativo

Tabla 4: Resumen variables categóricas parte 2

Cluster	Periventricular_MRI	Cortical_MRI	Infratentorial_MRI	Spinal_Cord_MRI
0	Negativo	Negativo	Negativo	Negativo
1	Positivo	Positivo	Positivo	Negativo
2	Positivo	Positivo	Negativo	Positivo
3	Negativo	Negativo	Negativo	Negativo

Tabla 5: Resumen variables categóricas parte 3

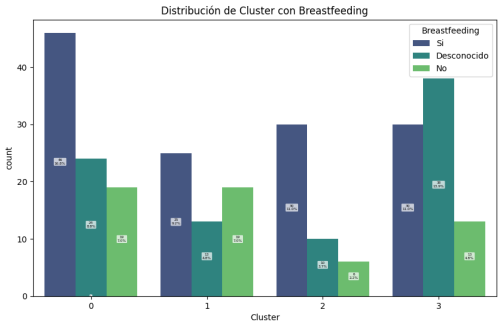
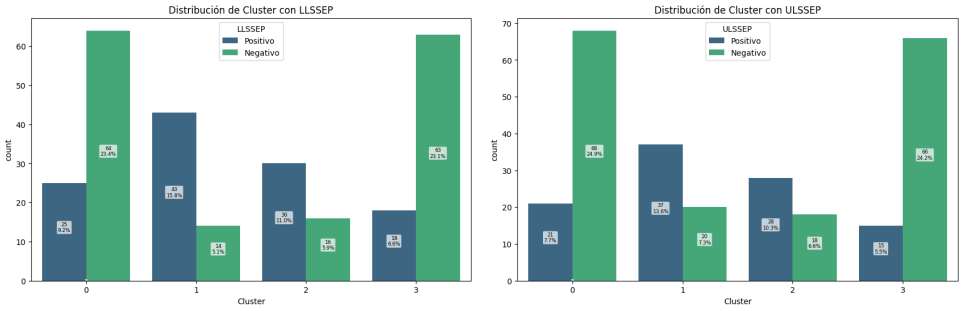
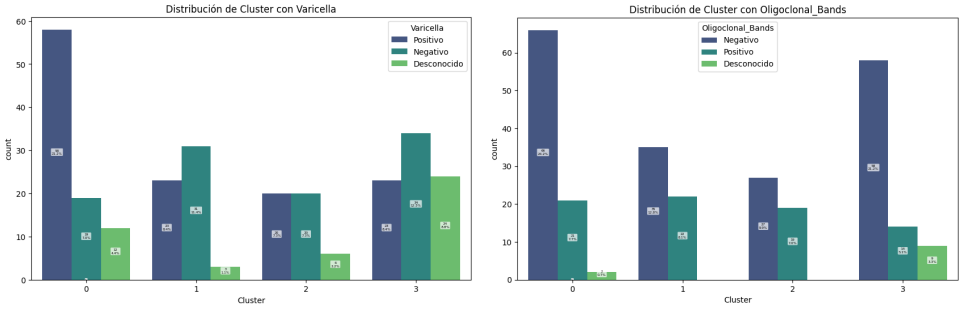


Figura 6: Distribución de variable Breastfeeding



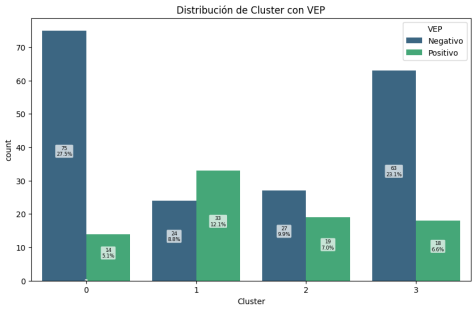
(a) Distribución de LLSSEP (b) Distribución de ULSSEP

Figura 7: Distribuciones de resultados exámenes SSEP

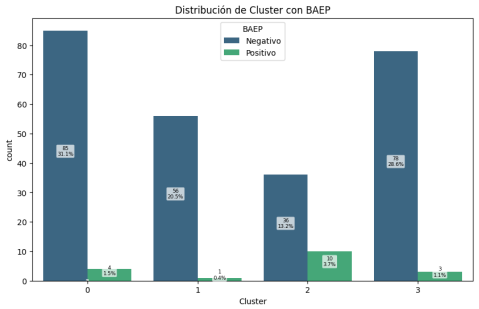


(a) Distribución de Varicella

(b) Distribución de Oligoclonal_Bands



(c) Distribución de VEP

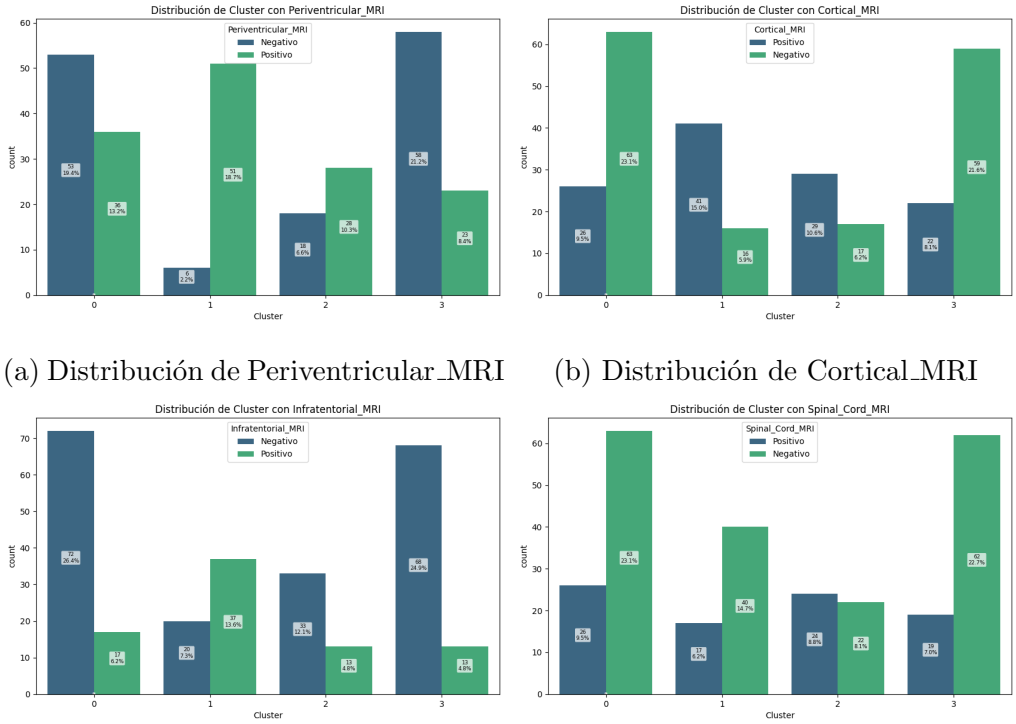


(d) Distribución de BAEP

Figura 8: Distribuciones de resultados de diversos exámenes

3.5.1. Análisis general

El análisis de los clústers proporciona una visión exhaustiva de cómo la Esclerosis Múltiple Clínicamente Definida (EMCD) varía según demografía, síntomas, y resultados médicos. El **Clúster 0**, predominantemente compuesto por mujeres jóvenes, bien educadas, que muestran una tasa relativamente baja de EMCD (12.1 %), como se refleja en sus mayoritariamente negativos resultados de pruebas diagnósticas y MRI. Por contraste, el **Clúster 1** muestra una alta prevalencia de EMCD (19 %) en hombres con menor nivel educativo y síntomas más severos, acompañados de significativos hallazgos positivos en MRI, indicando un desarrollo de SCA a EMCD, posiblemente exacerbada por limitado acceso a intervenciones tempranas y preventivas. **Clúster 2** es notable por su diversidad, con una edad media más alta y variabilidad en la educación, mostrando una prevalencia moderada de EMCD (10.3 %) y resultados mixtos en pruebas, este clúster no representa a un grupo en común y tiene valores variados para cada variable. Finalmente, el **Clúster 3**, también compuesto mayoritariamente por mujeres pero con la menor incidencia de EMCD (4.4 %) y la mayor proporción sin EMCD



(a) Distribución de Periventricular_MRI (b) Distribución de Cortical_MRI

(c) Distribución de Infratentorial_MRI (d) Distribución de Spinal_Cord_MRI

Figura 9: Distribuciones de MRIs

(25.3 %), presenta indicadores de un manejo efectivo o menor susceptibilidad genética a la enfermedad, apoyado por pruebas diagnósticas mayormente negativas.

4. Conclusiones

El estudio realizado sobre el Síndrome Clínicamente Aislado (SCA) y su posible evolución hacia la Esclerosis Múltiple Clínicamente Definida (EMCD) ha proporcionado conocimiento valioso mediante el uso del algoritmo de clustering K-Prototypes, aprovechando eficientemente la diversidad de datos numéricos y categóricos. El análisis exhaustivo comenzó con la adecuada limpieza, imputación y normalización de datos, asegurando una base sólida para el agrupamiento.

Los resultados del estudio mostraron cómo diferentes clústers representan variaciones significativas en términos de desarrollo de EMCD, influenciados por factores demográficos y clínicos. Se identificaron cuatro clústers distintos, cada uno con características y tendencias particulares en la progresión de SCA a EMCD.

Aunque el uso del algoritmo K-Prototypes ha permitido segmentar y analizar los datos los resultados han revelado áreas de mejora. El puntaje de silueta muy bajo y el índice Davies-Bouldin alto indican una superposición significativa entre los clusters, sugiriendo la necesidad de refinar el proceso de clustering.

En términos de aspectos positivos, el estudio ha destacado la importancia de considerar una amplia gama de características, desde datos demográficos hasta resultados médicos, para comprender mejor la progresión del SCA. La metodología adoptada, particularmente la elección del algoritmo y las métricas de disimilitud adecuadas para datos mixtos, ha demostrado ser adecuada para manejar la complejidad de los datos.

Para futuros estudios, se recomienda explorar otros métodos de clustering o ajustar la técnica actual para mejorar la calidad de los clusters. Además, podría ser beneficioso revaluar las variables eliminadas o incorporar nuevas variables que podrían proporcionar más conocimiento sobre los factores que influyen en la progresión de la enfermedad.

Este estudio deja ver la complejidad del SCA y su potencial evolución hacia EMCD, y demuestra cómo un enfoque en el análisis de datos puede ayudar a descubrir patrones significativos que son cruciales para la intervención temprana de la enfermedad.

Referencias

Bermudez, V., Castrejon, R., Torres, K., Flores, J., Flores, M., & Vicente Madrid, C. H. (2016). Papel de las enfermedades infecciosas en el desarrollo de la esclerosis múltiple: evidencia científica. *PMC*, 40-48. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7154617/

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, *36*(7), 10223-10228. https://doi.org/https://doi.org/10.1016/j.eswa.2009.01.060

Centroid [Accedido el 12/05/2024]. (s.f.).

Conversion Predictors of CIS to Multiple Sclerosis [Último acceso: 2024-04-14]. (2023). *Kaggle*. https://www.kaggle.com/datasets/desalegngeb/conversion-predictors-of-cis-to-multiple-sclerosis/data

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224-227.

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

Esclerosis múltiple - Síntomas y causas [Último acceso: 2024-04-14]. (s.f.). *Mayo Clinic*. https://www.mayoclinic.org/es/diseases-conditions/multiple-sclerosis/symptoms-causes/syc-20350269

Euclidean distance [Accedido el 12/05/2024]. (s.f.).

Harmouch, M. (2023). 17 Types of Similarity and Dissimilarity Measures Used in Data Science [Accessed: 2023-05-12].

Huang, Z. (1998). Extensions of the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*. https://cse.hkust.edu.hk/~qyang/537/Papers/huang98extensions.pdf

Institute, I. (2023). ¿Qué es el Clustering? [Accedido el 12/05/2024]. https://immune.institute/blog/que-es-el-clustering/

Mobility, S. (2019). *Clustering: Cómo obtener agrupaciones inherentes en los datos* [Accedido el 12/05/2024]. https://slashmobility.com/blog/2019/07/clustering-como-obtener-agrupaciones-inherentes-en-los-datos/

Overlap distance [Accedido el 12/05/2024]. (s.f.).

Peñailillo, E., Zerega, M., Elizabeth Guerrero, E. C., Uribe, R., Cárcamo, C., Arraño, L., Bravo, S., & Cruz, J. (2019). Ensayo pictórico: Diagnóstico diferencial radiológico en Esclerosis Múltiple. *Revista chilena de radiología*, *25*, 5-18. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-93082019000100005&nrm=iso

Sara Collorone, A. T. T., Srikirti Kodali. (2022). The protective role of breastfeeding in multiple sclerosis: Latest evidence and practical considerations. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.1090133

Síndrome Clínicamente Aislado (CIS) [Último acceso: 2024-04-14]. (s.f.). *National Multiple Sclerosis Society*. https://www.nationalmssociety.org/es/que-es-esclerosis-multiple/tipos-de-esclerosis-multiple/sindrome-clinicamente-aislado#:~:text=El%20s%C3%ADndrome%20cl%C3%ADnicamente%20aislado%20es,esclerosis%20m%C3%BAltiple%20en%20el%20futuro.

sklearn.metrics.davies_bouldin_score [Accedido el 12/05/2024]. (2024). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

sklearn.metrics.silhouette_score [Accedido el 12/05/2024]. (2024). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Trust, M. (2022). Expanded Disability Status Scale (EDSS). https://mstrust.org.uk/a-z/expanded-disability-status-scale-edss

UNESCO. (2019). Education and health: The role of cognitive skills [Accessed: 2023-04-19]. https://unesdoc.unesco.org/ark:/48223/pf0000381728