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

Laboratorio 3 - Reglas de Asociación

Integrantes: Matías Figueroa Contreras
Curso: Análisis de Datos
Profesor: Max Chacón Pacheco
Ayudante: Marcelo Álvarez

9 de Junio 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 laboratorio, se hará uso de reglas de asociación las cuales permiten obtener relaciones entre las variables de un conjunto de datos (Hashemi-Pour & Lutkevich, 2024). El uso de las reglas de asociación en este contexto, ayudara a obtener patrones en la aparición de variables que se relacionen con el desarrollo esclerosis múltiple clínicamente definida (EMCD) en los pacientes del estudio.

1.1. Objetivos

El objetivo principal de este laboratorio es determinar y caracterizar los patrones de progresión desde SCA hacia EMCD, mediante el análisis de datos utilizando reglas de asociación, con el fin de identificar tempranamente a los pacientes con alto riesgo de evolución hacia la EMCD. Lo anterior conlleva lo siguiente:

- Aplicar el algoritmo apriori, seleccionando y definiendo los umbrales de las medidas de calidad relevantes a utilizar para la obtención de las reglas de asociación.
- Identificar patrones distintivos en los datos que puedan identificar relaciones comunes en 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. Reglas de Asociación

Las reglas de asociación son una técnica de minería de datos cuyo propósito es encontrar relaciones interesantes en grandes conjuntos de datos. Estas reglas ayudan a descubrir patrones frecuentes y asociaciones ocultas entre los elementos de una base de datos. Formalmente, una regla de asociación se expresa como una implicación de la forma $A \Rightarrow B$, donde A y B son conjuntos de ítems (Hashemi-Pour & Lutkevich, 2024).

2.1.1. Ítems y Transacciones

Un ítem es una unidad elemental de datos, como un producto en una tienda. Una transacción es un conjunto de ítems que ocurren juntos en una instancia del conjunto de datos, como todos los productos comprados en una sola venta. Por ejemplo, en un supermercado, una transacción podría ser la lista de productos comprados por un cliente en una visita (Hashemi-Pour & Lutkevich, 2024).

2.2. Medidas de Calidad

Para evaluar la calidad de las reglas de asociación, se utilizan varias métricas, siendo las más comunes el soporte, la confianza y el lift. Estas medidas permiten determinar la relevancia y utilidad de las reglas encontradas.

2.2.1. Soporte

El soporte (*support*) de una regla de asociación es una medida que indica la frecuencia con la que los ítems aparecen juntos en el conjunto de datos. Se define como la proporción del número de transacciones que contienen $A \cup B$ respecto al número total de transacciones (GeeksforGeeks, 2023). Matemáticamente, se expresa como:

$$\text{Soporte}(A \Rightarrow B) = \frac{\text{Número de transacciones que contienen } A \cup B}{\text{Número total de transacciones}}$$

2.2.2. Confianza

La confianza (*confidence*) de una regla de asociación es una medida de la certeza de que la ocurrencia del antecedente A lleva a la ocurrencia del consecuente B. Se calcula como la proporción del número de transacciones que contienen $A \cup B$ respecto al número de transacciones que contienen A (GeeksforGeeks, 2023). Formalmente, se define como:

$$\text{Confianza}(A \Rightarrow B) = \frac{\text{Número de transacciones que contienen } A \cup B}{\text{Número de transacciones que contienen } A}$$

2.2.3. Lift

El lift es una medida que evalúa la independencia entre A y B. Indica cuántas veces más frecuente es la ocurrencia conjunta de A y B en comparación con lo que se esperaría si A y B fueran independientes. Se calcula como:

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Confianza}(A \Rightarrow B)}{\text{Soporte}(B)}$$

Un lift mayor a 1 indica que A y B ocurren juntos más frecuentemente de lo que se esperaría por azar (Ali, 2023).

2.3. Algoritmo Apriori

El algoritmo Apriori, propuesto por Rakesh Agrawal y Ramakrishnan Srikant en 1994 (colaboradores de Wikipedia, 2024), es un método reconocido en la minería de reglas de asociación. Este algoritmo comienza estableciendo un umbral de soporte mínimo que determina la frecuencia necesaria para que un ítem sea considerado dentro de un conjunto de ítems frecuentes. Los conjuntos que no cumplen con este criterio son eliminados.

Utilizando los ítems frecuentes, Apriori genera todas las combinaciones posibles, evaluando cuántas veces cada una aparece en la base de datos. Basándose en estas combinaciones, el algoritmo formula reglas de asociación, conservando únicamente aquellas que superan un umbral de confianza predeterminado, denominadas como reglas de asociación fuertes. Apriori adopta un enfoque de abajo hacia arriba, comenzando con ítems individuales y agrupándolos progresivamente en conjuntos más grandes a medida que identifica patrones

frecuentes. Implementa, además, un método de eliminar y reetiquetar para reducir eficientemente el espacio de búsqueda, excluyendo conjuntos de ítems infrecuentes. El resultado es una lista de reglas de asociación fuertes, aplicables en diversas áreas como el análisis de cestas de compras, sistemas de recomendación y detección de fraudes (Ali, 2023).

3. Obtención de Reglas

3.1. Pre-procesamiento

Antes de proceder a la generación de reglas de asociación, 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. Por otro lado, las variables *Breastfeeding*, *Mono_or_polysintomatic*, *Varicela* y *Oligoconal_Bands*, tenian datos catalogados como "Desconocido", a estos valores igual se les aplico imputación de datos. 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.

- *Mono_or_polysintomatic*: se imputo basado en los datos de síntomas iniciales, lo que esta directamente relacionado con esta variable.
- *Varicela* y *Oligoconal_Bands*: para ambas variables se decidido utilizar MICE (Multivariate Imputation by Chained Equations) mediante el método "predictive mean matching"(PMM) para imputar los datos faltantes, ya que este es capaz de conservar la distribución original de los datos y aprovechar la información de otras variables interrelacionadas en el conjunto de datos (Soni, 2023).

(B) **Eliminación de columnas**: Se excluyeron las variables *Initial_EDSS* y *Final_EDSS* del análisis debido a la ausencia completa de datos para el grupo 2 (no EMCD). Esta exclusión se debe a que estas variables representan una escala de discapacidad diseñada específicamente para pacientes con esclerosis múltiple clínicamente definida (EMCD), y su ausencia en el grupo no EMCD se explica porque estos individuos no desarrollaron EMCD y, por tanto, no se les realizaron evaluaciones para monitorear el desarrollo de la enfermedad. La omisión de estas variables se justifica porque no se pueden estimar las medidas de discapacidad en pacientes que no han progresado a EMCD. Además, se decidió eliminar la variable *Breastfeeding* debido a la alta incidencia de datos faltantes. Utilizar imputación artificial o incluir estos datos de manera incompleta podría introducir un sesgo significativo en el análisis. Así, para preservar la integridad y precisión del análisis en base a reglas de asociación, es adecuado excluir estas variables y centrarse en otras características clínicas y demográficas comunes a ambos grupos.

3.1.2. Discretización de variables numéricas

Dado que las reglas de asociación son generadas a partir de datos categóricos, se discretizaron las siguientes variables:

- **Edad:** Se seleccionaron los rangos que se ven reflejados en la tabla 1 se basa en lo revisado en Brichford, 2024. En el cual se menciona que el rango de los adultos jóvenes a edad media de entre 20 a 50 años, es cuando se diagnostica la mayoría de casos. Por otro lado, para el rango de los niños y adolescentes (edad menor a 20 año), es importante estudiar este rango ya que permitirá observar como la SCA evoluciona a

EMCD durante etapas criticas del desarrollo, además como se menciona en Brichford, 2024 hasta el 10 % de los casos se diagnostican en este grupo. Por ultimo para los adultos de 51 años en adelante, se muestra que la investigación presenta los desafíos diagnósticos en medio de cambios relacionados con la vejez y comorbilidades presentes dado este factor.

Descripción	**Rango de Edad (años)**
Niños y Adolescentes	Menor a 20
Adultos jóvenes a edad media	20-50
Adultos de edad media a mayor	Mayor a 50

Tabla 1: Descripción de los Rangos de Edad

- **Escolaridad:** Dado que las muestras fueron realizadas en México, se decidió utilizar los niveles de escolaridad de este país («Escolaridad. Cuéntame de México», s.f.), generando así las categorías presentes en la tabla 2

Nivel de Escolaridad	**Años de Duración**
Sin Educación Formal	0 años
Educación Primaria	1-6 años
Educación Secundaria	7-9 años
Educación Media Superior	10-12 años
Educación Superior	13-17 años
Posgrado	18-25 años

Tabla 2: Descripción de los Niveles de Escolaridad

3.1.3. Codificación de variables categóricas

Para aplicar el algoritmo Apriori, se transformaron las variables categóricas *Varicella*, *Oligoclonal_Bands*, *LLSSEP*, *ULSSEP*, *VEP*, *BAEP*, *Periventricular_MRI*, *Cortical_MRI*, *Infratentorial_MRI* y *Spinal_Cord_MRI* en variables booleanas, dado que estas indicaban 'positivo' o 'negativo' para ciertos exámenes o antecedentes médicos. Por otra parte,

la variable *Initial_Symptom*, que contenía combinaciones de presencia de síntomas entre *Sensorial*, *Motor*, *Visual* y *Otro*, fue transformada en múltiples columnas booleanas dentro del conjunto de datos para reflejar la presencia o ausencia de cada síntoma en particular. Por ultimo, la variable *group* se separo en las variables booleanas *EMCD* y *No_EMCD*, facilitando así la identificación de asociaciones tanto en los casos que evolucionaron hacia EMCD como en aquellos que no presentaron esta progresión al aplicar el algoritmo.

3.2. Generación de Reglas

Para la generación de reglas de asociación, se fijaron como consecuente las variables generadas *EMCD* y *No_EMCD*, con el objetivo de estudiar cómo los antecedentes influyen sobre estas variables. Usando el algoritmo Apriori se obtuvo lo siguiente:

3.2.1. Pacientes que desarrollaron EMCD

Se decidió establecer un valor de soporte de 0.1 (10 %) considerando que las características y síntomas que se evalúan en relación con la progresión a EMCD no necesariamente serán predominantes entre todos los pacientes. Un soporte del 10 % asegura que las reglas identificadas no sean excepcionalmente raras y posean suficiente relevancia estadística para ser consideradas significativas en un contexto de estudio clínico. Por otro lado, se seleccionó un valor de confianza del 0.9 (90 %) para garantizar que las reglas de asociación sean altamente confiables en predecir la transición hacia EMCD ante la aparición de ciertos síntomas o combinaciones de estos. Los resultados obtenidos se muestran en la tabla 3, los cuales están ordenados por el valor de lift que indica la fuerza y utilidad de cada regla dentro del estudio, reflejando cómo de probable es que el consecuente ocurra dado el antecedente en comparación con su frecuencia esperada de manera independiente (Ali, 2023). Esta métrica ayuda a identificar aquellas asociaciones que son de especial interés clínico por su mayor impacto predictivo.

Regla	**Soporte (%)**	**Confianza (%)**	**Lift**
{Periventricular MRI, Infratentorial MRI, Motor} ⇒ {EMCD}	13.55	94.87	2.0720
{Polisintomático, Periventricular MRI, Infratentorial MRI, Motor} ⇒ {EMCD}	11.36	93.94	2.0516
{Periventricular MRI, Infratentorial MRI, Age 20-50, Motor} ⇒ {EMCD}	10.99	93.75	2.0475
{LLSSEP, Infratentorial MRI, Motor} ⇒ {EMCD}	10.26	93.33	2.0384
{Periventricular MRI, Infratentorial MRI, Otro} ⇒ {EMCD}	11.72	91.43	1.9968
{ULSSEP, Periventricular MRI, Infratentorial MRI} ⇒ {EMCD}	10.99	90.91	1.9855
{Periventricular MRI, Age 20-50, Educación Superior} ⇒ {EMCD}	10.62	90.62	1.9793
{Polisintomático, Periventricular MRI, Infratentorial MRI, Otro} ⇒ {EMCD}	10.62	90.62	1.9793
{Polisintomático, Periventricular MRI, Infratentorial MRI, Visual} ⇒ {EMCD}	10.26	90.32	1.9726
{LLSSEP, ULSSEP, Periventricular MRI, Infratentorial MRI} ⇒ {EMCD}	10.26	90.32	1.9726

Tabla 3: Resultados de Reglas de Asociación EMCD usando el algoritmo Apriori

3.2.2. Pacientes que NO desarrollaron EMCD

Para las reglas donde el consecuente es *No_EMCD*, se estableció un umbral de soporte más bajo (6 %). Este ajuste se realizó para abarcar un espectro más amplio de síntomas o características clínicas posiblemente vinculadas a no desarrollar EMCD. Dado que estos casos pueden ser más variables y menos definidos para el no desarrollo a EMCD. Un umbral de soporte más bajo permite descubrir relaciones menos frecuentes pero igualmente importantes, que podrían no detectarse utilizando umbrales más altos. Por otra parte, se

seleccionó un nivel de confianza del 75 %. Este valor más bajo, en comparación con el utilizado para los pacientes que sí desarrollan EMCD, se seleccionó para reflejar un mayor grado de incertidumbre y una diversidad más amplia de factores entre los pacientes que no progresan hacia EMCD. Al igual que en el caso anterior, las reglas están organizadas en función del valor de lift, y los resultados correspondientes se pueden consultar en la tabla 4.

Regla	**Soporte (%)**	**Confianza (%)**	**Lift**
{Monosintomático, Sensorial} ⇒ {No_EMCD}	6.96	90.48	1.6689
{Femenino, Polisintomático, Age 20-50, Educación Media Superior} ⇒ {No_EMCD}	6.23	80.95	1.4932
{Femenino, Polisintomático, Educación Media Superior} ⇒ {No_EMCD}	7.33	80.00	1.4757
{Femenino, Educación Media Superior} ⇒ {No_EMCD}	9.52	78.79	1.4533
{Femenino, Age 20-50, Educación Media Superior} ⇒ {No_EMCD}	8.06	78.57	1.4493

Tabla 4: Resultados de Reglas de Asociación No EMCD usando el algoritmo Apriori

4. Análisis de Resultados

A continuación se profundiza en las reglas de asociación identificadas para investigar qué factores contribuyen a la progresión hacia la esclerosis múltiple clínicamente definida (EMCD) y cuáles están asociados con la no progresión de la enfermedad.

4.1. Análisis para pacientes que desarrollaron EMCD

Uno de los patrones más destacados revelados por las reglas de asociación involucra la presencia de lesiones en las resonancias magnéticas (MRI) tanto periventricular como infratentorial, asociadas con síntomas motores en los pacientes. La investigación previa ha indicado que las lesiones periventriculares son un marcador temprano y potente de la EM, dado que sugieren una probabilidad incrementada de inflamación crónica dentro del cerebro, particularmente en las áreas alrededor de los ventrículos cerebrales, que son críticas para el procesamiento y la transmisión de señales neuronales (Peñailillo et al., 2019). De manera similar, las lesiones infratentoriales, que afectan el tronco cerebral y el cerebelo, están asociadas con disfunciones motoras y de coordinación, que son síntomas frecuentes en pacientes de EM avanzada (Peñailillo et al., 2019). La literatura específica indica que la combinación de hallazgos en estas regiones de la MRI aumenta significativamente la probabilidad de progresión a EM (Peñailillo et al., 2019), lo cual coincide con la alta confianza y el lift observadosen las reglas de asociación que incluyen esta combinación específica de antecedentes, la cual encabeza los resultados. En estos resultados, se presenta un valor de lift destacado de 2.0720, junto con un nivel de confianza alto del 94.87 % y un soporte del 13.55 %. Reflejando así, la relevancia clínica de estos factores de riesgo y su potencial para identificar a aquellos pacientes con mayor riesgo de desarrollar EMCD.

Además, la inclusión de síntomas visuales y motores en combinación con hallazgos MRI sugiere un vínculo entre la expresión de síntomas específicos y la transición a EMCD. Estudios anteriores han mostrado que síntomas como trastornos de la visión y problemas motores son indicativos de una afección extensa y posiblemente agresiva del sistema nervioso central, que en última instancia puede culminar en una diagnosis de EM (Peñailillo et al., 2019). Estos síntomas, cuando son registrados en conjunto con signos de problemas encon-

trados en exámenes (Peñailillo et al., 2019 y Wagner et al., 2021), refuerzan la predicción de una evolución desfavorable de la SCA hacia EMCD, como reflejan las altas medidas de lift y confianza en las reglas obtenidas.

El impacto de la edad y la educación también se refleja en las reglas de asociación, enlazando la prevalencia de la EM principalmente en adultos jóvenes a media edad, un fenómeno que coincide con estudios revisados (Brichford, 2024). Además, la relación entre un mayor nivel de educación y la progresión a EMCD podría interpretarse bajo la opción de un mejor acceso a recursos de salud y una mayor capacidad de reconocimiento de los síntomas iniciales de la enfermedad, facilitando diagnósticos tempranos y posiblemente mejores resultados a largo plazo. Esta asociación es interesante porque resalta cómo factores socio económicos y educativos pueden influir en la detección y manejo de una enfermedad neurológica compleja como la EM.

En el contexto de las pruebas de potenciales evocados (LLSSEP y ULSSEP), su inclusión en las reglas con alto lift y confianza, sugiere que las disfunciones en las vías sensoriales, detectadas a través de estas técnicas, son predictores tempranos de daño neural en la EM. Las pruebas de potenciales evocados son capaces de identificar anormalidades clínicas antes que los síntomas se manifiesten claramente, proveyendo así un medio de anticipar el desarrollo de EM en pacientes de SCA (Wagner et al., 2021). Estudios anteriores han corroborado la utilidad de estas pruebas en la evaluación temprana de la EM, reforzando la validez de las reglas generadas (Wagner et al., 2021).

4.2. Análisis para pacientes que NO desarrollaron EMCD

Una de las asociaciones más significativas identificadas es la que involucra síntomas iniciales monosintomáticos junto con síntomas de tipo sensorial. Esta combinación, con un alto valor de confianza (90.48 %), sugiere que los pacientes que presentan un único síntoma sensorial tienen menor probabilidad de progresar hacia EMCD. Esta observación puede estar alineada con estudios previos que sugieren que presentaciones iniciales menos complejas y más específicas, como problemas sensoriales aislados, pueden no siempre correlacionarse con la acumulación de daño neurológico a largo plazo que caracteriza a la EM («Esclerosis múltiple - Síntomas y causas», s.f.).

Otro patrón destacado en las reglas se relaciona con factores demográficos y de estilo de vida, como pertenecer al genero femenino, tener entre 20 y 50 años y haber alcanzado un nivel de educación media superior. Estas reglas, especialmente aquellas que incluyen la combinación de género femenino y nivel de educación media superior, podrían reflejar un acceso más efectivo a recursos de salud y estrategias de manejo de síntomas que impiden la progresión de la enfermedad. Curiosamente, la literatura ha mostrado que, aunque la EM es más común en mujeres («Esclerosis múltiple - Síntomas y causas», s.f.), la variabilidad en la presentación y evolución de los síntomas puede influir en cómo cada individuo responde a la enfermedad, lo que puede estar en parte modulado por factores socio económicos y educativos.

La inclusión de la categoría de edad entre 20 y 50 años en varias de estas reglas apoya la idea de que durante estas edades se pueden establecer patrones de vida y manejo de la salud que potencialmente contribuyen a un mejor pronóstico en caso de manifestaciones iniciales de la enfermedad. Estos hallazgos podrían estar indicando que la interacción entre el nivel de educación y la atención médica durante este período crítico de edad puede desempeñar un papel crucial en mitigar la progresión de SCA a EMCD. Sin embargo, es importante destacar que, según la literatura, el desarrollo de EM es más común precisamente en este rango de edad (Brichford, 2024), lo que sugiere que existan otros factores no capturados en este estudio que influyen significativamente en la evolución de la enfermedad.

4.3. Comparación con laboratorios anteriores

En los laboratorios previos, se han utilizado métodos variados para analizar las relaciones entre distintas variables y la enfermedad. El primer laboratorio utilizó técnicas de análisis estadístico descriptivo e inferencial para explorar diferencias en variables entre grupos que desarrollaron o no EMCD. Estas técnicas permitieron identificar la posible influencia de variables como edad, escolaridad y resultados de pruebas clínicas entre los pacientes que desarrollaron EM (Contreras, 2024a). En el segundo laboratorio, se implementó un análisis de clúster para agrupar a los pacientes según características similares y observar cómo estas agrupaciones se correlacionaban con la presencia o ausencia de EMCD, proporcionando una comprensión más profunda de cómo los síntomas y los antecedentes clínicos podrían agruparse

en patrones que influyen en el riesgo de desarrollar EMCD (Contreras, 2024b).

El análisis actual, centrado en las reglas de asociación, coincide con los hallazgos de los estudios previos en varios aspectos importantes. Todos los métodos han subrayado la importancia de variables como los síntomas motores y visuales, y los resultados de MRI, estableciendo fuertes vínculos entre estas características y la progresión hacia EMCD. Además, tanto los clústeres como las reglas de asociación han resaltado la relevancia de la edad y la escolaridad, encontrando que el rango de edad de 20 a 50 años y un mayor nivel de educación se asocian con patrones específicos de progresión de la enfermedad.

Un hallazgo interesante es que, a pesar de las consistencias observadas en varios parámetros, todos los laboratorios mostraron diferencias con respecto a la literatura en cuanto al género de los pacientes que con más frecuencia desarrollaron EM. Contrario a lo que reportan estudios donde las mujeres muestran una prevalencia mayor de la enfermedad (Brichford, 2024), los análisis desarrollados no corroboraron que el género femenino tuviera una incidencia más alta de EMCD, sino al contrario. Este resultado indica que podrían existir características específicas en la muestra o en el entorno clínico analizado que alteren este patrón.

5. Conclusiones

El análisis realizado mediante reglas de asociación brindó el conocimiento de una herramienta efectiva para explorar y descubrir patrones potenciales en la progresión del síndrome clínicamente aislado (SCA) hacia la esclerosis múltiple clínicamente definida (EMCD). El método de reglas de asociación, aplicado a través del algoritmo Apriori, permitió establecer reglas con altos niveles de confianza y soporte, proporcionando un marco confiable para interpretar cómo ciertas combinaciones de síntomas y resultados de exámenes influyen en la evolución de la enfermedad.

Uno de los hallazgos más notables del análisis fue la asociación entre la presencia de lesiones en las resonancias magnéticas, especialmente periventriculares e infratentoriales, y el desarrollo de EMCD. Estos hallazgos radiológicos, junto con síntomas motores, ofrecieron un alto grado de confiabilidad en la predicción de progresión a EMCD, destacados por altas métricas de soporte y confianza. Además, las reglas derivadas sugieren que los pacientes que exhiben síntomas iniciales polisintomáticos junto con ciertos patrones en resonancias magnéticas tienen un riesgo considerablemente mayor de progresar hacia EMCD. Esto refuerza la importancia de una evaluación detallada y temprana en pacientes con SCA para identificar aquellos con alto riesgo de desarrollar EMCD.

Por otro lado, el análisis de pacientes que no desarrollaron EMCD identificó patrones importantes. En particular, la identificación de síntomas iniciales monosintomáticos de tipo sensorial sugiere que presentaciones menos complejas podrían estar asociadas con un menor riesgo de progresión. Ademas, tenia asociado pacientes de género femenino, contradiciendo la literatura que indican que en su mayoria estas son las que desarrollan EMCD. Esto podría indicar la influencia de factores específicos no capturados en la muestra analizada. Además, la consideración de factores demográficos y de estilo de vida en el análisis sugiere que el acceso a la educación y recursos de salud adecuados juega un papel crucial en prevenir el desarrollo de EM.

El uso de herramientas de minería de datos ha sido un punto fuerte, permitiendo una comprensión de la dinámica de la enfermedad a partir de los datos. La metodología de pre-procesamiento de datos aseguró la calidad y integridad del conjunto de datos, lo se vio

reflejado en el análisis y hallazgos encontrados.

Sin embargo, el estudio presenta oportunidades de mejora. La inclusión de un conjunto más amplio de variables clínicas podría proporcionar una visión más global de los factores que influyen en la progresión a EMCD. Es importante destacar que el conjunto de datos original presentaba ausencias en ciertas variables claves, lo que llevó a la eliminación de algunas de estas del análisis. Incorporar estrategias efectivas para manejar o recuperar estos datos faltantes podría permitir una evaluación más completa y detallada de todos los factores potenciales. Además, la validación de reglas de asociación en estudios adicionales sobre otros conjuntos de datos es crucial para confirmar los hallazgos encontrados en diferentes entornos y poblaciones.

En comparación con los métodos aplicados en estudios anteriores, como el análisis estadístico descriptivo e inferencial y el análisis de clúster, el uso de reglas de asociación en este laboratorio ha proporcionado un enfoque más directo y estructurado para identificar relaciones significativas entre síntomas y progresión hacia la esclerosis múltiple. La gran mayoría de estos hallazgos no solo están alineados con la literatura existente, sino que también amplían la comprensión de cómo las presentaciones iniciales y factores demográficos pueden influir en la evolución de la enfermedad. Esta metodología ha demostrado ser especialmente útil en el análisis de los patrones de progresión del síndrome clínicamente aislado (SCA) hacia EMCD, proporcionando conocimiento valiosos para la intervención temprana de esta enfermedad.

Referencias

Ali, M. (2023, enero). Association Rule Mining in Python [Último acceso el 9 de junio de 2024]. https://www.datacamp.com/tutorial/association-association-rule-mining-python

Brichford, C. (2024, junio). How Age Affects Multiple Sclerosis Symptoms and Progression. https://www.everydayhealth.com/multiple-sclerosis/symptoms/multiple-sclerosis-age-progression/

colaboradores de Wikipedia. (2024, mayo). Algoritmo apriori. https://es.wikipedia.org/wiki/Algoritmo_apriori

Contreras, M. F. (2024a). Laboratorio 1 - Análisis de Datos [Realizado el 21 de abril de 2024].

Contreras, M. F. (2024b). Laboratorio 2 - Análisis de Datos [Realizado el 12 de mayo de 2024].

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

Escolaridad. Cuéntame de México. (s.f.). https://www.cuentame.inegi.org.mx/poblacion/escolaridad.aspx

GeeksforGeeks. (2023, enero). What is Support and Confidence in Data Mining? https://www.geeksforgeeks.org/what-is-support-and-confidence-in-data-mining/

Hashemi-Pour, C., & Lutkevich, B. (2024, abril). association rules. https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining#:~:text=Association%20rules%20are%20if%2Dthen,in%20various%20types%20of%20databases.

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

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.

Soni, B. (2023). Topic:9 MICE or Multivariate Imputation with Chain-Equation. https://medium.com/@brijesh_soni/topic-9-mice-or-multivariate-imputation-with-chain-equation-f8fd435ca91#:~:text=MICE%20stands%20for%20Multivariate%20Imputation,produce%20a%20final%20imputed%20dataset.

Wagner, A. K., Franzese, K., Weppner, J. L., Kwasnica, C., Galang, G. N., Edinger, J., & Linsenmeyer, M. (2021). 43 - Traumatic Brain Injury (D. X. Cifu, Ed.; Sixth Edition), 916-953.e19. https://doi.org/https://doi.org/10.1016/B978-0-323-62539-5.00043-6