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


# Laboratorio 1 - Análisis de Datos

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

21 de Abril 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.).

## 1.1. Objetivos

El objetivo principal de este laboratorio es comprender, identificar y analizar los factores que tienen relación con la conversión de CIS a EM sobre el conjunto de datos de pacientes atendidos en el Instituto Nacional de Neurología y Neurocirugía en Ciudad de México entre los años 2006 y 2010 («Conversion Predictors of CIS to Multiple Sclerosis», 2023). El análisis de estos datos conlleva:

- Identificar y describir las características clínicas y diagnosticas presentes en los datos, estableciendo como cada atributo puede estar relacionado con la progresión de SCA a EM.
- Utilizar técnicas de estadística descriptiva e inferencial para investigar las interacciones entre diferentes variables y su impacto en la evolución de la enfermedad.
- Construir y validar modelos predictivos que puedan estimar la probabilidad de conversión de SCA a EM, basados en los datos y patrones identificados.
- Obtención de conocimiento dado el análisis realizado al conjunto de datos.

# 2. Descripción del problema

El síndrome clínicamente aislado (SCA) representa un desafío significativo en el ámbito de la neurología debido a su naturaleza incierta y su potencial para desarrollarse en esclerosis múltiple (EM), una enfermedad crónica y progresiva del sistema nervioso central. El SCA es definido como un primer episodio neurológico que dura al menos 24 horas, con síntomas que sugieren inflamación o desmielinización, pero sin suficiente evidencia para cumplir con los criterios de diagnóstico de la EM. Los síntomas de SCA pueden variar ampliamente dependiendo de las áreas del sistema nervioso que estén afectadas, incluyendo, pero no limitándose a, problemas de visión, alteraciones sensoriales, y debilidad muscular o problemas de coordinación («Síndrome Clínicamente Aislado (CIS)», s.f.).

Por otro lado, la EM puede provocar una serie más extensa de síntomas físicos y cognitivos debido a la amplia gama de áreas que puede afectar dentro del sistema nervioso central. Estos incluyen fatiga severa, dificultades en la coordinación y el equilibrio, problemas de visión, alteraciones en el habla, temblores y, en etapas más avanzadas, deterioro cognitivo y cambios en la función psicológica, como depresión o cambios de humor («Esclerosis múltiple - Síntomas y causas», s.f.). La transición de SCA a EM puede significar una evolución hacia un estado más debilitante, lo que hace crucial la identificación temprana de aquellos pacientes con SCA que tienen un alto riesgo de desarrollar EM.

Dado lo anterior, identificar de manera temprana a los pacientes con SCA que probablemente evolucionarán a EM es vital para iniciar intervenciones terapéuticas que puedan retrasar la progresión de la enfermedad y mejorar la calidad de vida. Además, otorgar una mejor comprensión de los factores de riesgo asociados con la conversión de SCA a EM contribuiría a afinar las predicciones clínicas y a personalizar el manejo de la enfermedad.

En este laboratorio, el principal desafió consta de determinar con precisión cuales pacientes con SCA están en riesgo de llegar a desarrollar EM, empleando análisis para describir detalladamente las características del conjunto de datos y aplicar métodos estadísticos e inferenciales para encontrar patrones y correlaciones significativas entre las variables estudiadas para la comprensión de como las diferentes variables afectan la probabilidad de progresión a EM.

## 2.1. Descripción de la base de datos

La base de datos incluye información recogida de pacientes atendidos en el Instituto Nacional de Neurología y Neurocirugía en Ciudad de México entre 2006 y 2010 (Pineda & Flores Rivera, 2023). Este conjunto de datos cuenta con 273 instancias y un total de 19 variables relativas al estudio, en donde se encuentran factores demográficos, historial médico, presentaciones clínicas, síntomas y resultados de exámenes indicativos de EM. A continuación se especifican los campos y su relevancia del conjunto de datos en el estudio, además de una tabla resumen con las características de las variables 1:

1. **ID**: Identificador único del paciente.

2. **Age**: Edad del paciente en años. La edad puede influir en la probabilidad de desarrollar EM y en la progresión de la enfermedad. En donde, la EM generalmente se diagnostica entre los 20 y 40 años («Esclerosis múltiple - Síntomas y causas», s.f.).

3. **Schooling**: Años que el paciente estuvo en la escuela. Este factor puede ser relevante para un análisis de correlación entre nivel educativo y la salud del paciente (UNESCO, 2019).

4. **Gender**: Genero del paciente. la EM es más común en mujeres que en hombres («Esclerosis múltiple - Síntomas y causas», s.f.), por lo que este factor es relevante para el estudio y análisis de la evolución de la SCA.

5. **Breastfeeding**: Indica si el paciente fue amamantado cuando era bebe. La lactancia materna ha sido estudiada por sus posibles efectos protectores en el desarrollo de enfermedades auto-inmunes (Sara Collorone, 2022), por lo tanto es una variable a tener en cuenta en el estudio de la enfermedad.

6. **Varicella**: Refiere a si el paciente sufrió de varicela. Algunos estudios sugieren que ciertas infecciones virales podrían estar relacionadas con el riesgo de desarrollar EM (Bermudez et al., 2016).

7. **Initial_Symptoms**: Tipo de síntomas iniciales que el paciente experimento, codificados en varias categorías que combinan síntomas visuales, sensoriales y motores. El tipo

de síntoma inicial influye directamente en el diagnostico y pronostico de cualquier enfermedad.

8. **Mono_or_Polysymptomatic**: Indica si el paciente presento un solo síntoma o multiples sintomas en el primer episodio (Monosinotmático o Polisintomático respectivamente). Esta variable afecta en la evaluación inicial y el seguimiento del paciente.

9. **Oligoclonal_Bands**: Presencia de bandas oligoclonales en el suero sanguíneo o el liquido cefalorraquídeo. Utilizado en el diagnostico de diversas enfermedades neurológicas y sanguíneas. Este es un indicador común en los pacientes con esclerosis múltiple clínicamente definida («Conversion Predictors of CIS to Multiple Sclerosis», 2023).

10. **LLSSEP**: Resultado de procedimiento llamado potencial evocado somatosensorial para extremidades inferiores, el cual evalúa las vías sensoriales que viajan desde las extremidades inferiores hasta el cerebro. Se mide la respuesta eléctrica del cerebro después de estimular de manera controlada los nervios de las piernas. Es útil para detectar daños o disfunciones en las vías nerviosas, el cual es un sintoma comun en pacientes con EM (Wagner et al., 2021).

11. **ULSSEP**: Resultado de procedimiento potencial evocado somatosensorial similar al LLSSEP, pero se enfoca en las extremidades superiores. Estimulando los nervios de los brazos y midiendo la respuesta cerebral. Este estudio es importante evaluar la integridad de las vías sensoriales que conectan los brazos con el cerebro, siendo un síntoma común en la EM (Wagner et al., 2021).

12. **VEP**: Resultado de examen llamado potencial evocado visual que mide la respuesta eléctrica del cerebro a estímulos visuales. Utilizado comúnmente para evaluar la función de la retina y el nervio óptico y asi detectar problemas en la vía visual. Esto es relevante para determinar un síntoma común presente en la EM («Conversion Predictors of CIS to Multiple Sclerosis», 2023).

13. **BAEP**: Resultado de procedimiento llamado potenciales evocados auditivos del tronco encefálico que mide como el cerebro procesa los sonidos. en donde, se registra la actividad eléctrica en el tronco cerebral en respuesta a estímulos auditivos («Conversion

Predictors of CIS to Multiple Sclerosis», 2023). Es util para evaluar la función del tronco cerebral y los nervios relacionados con la audición, factor que esta presente como un síntoma en la EM.

14. **Periventricular_MRI**: Variable que indica la presencia o ausencia de lesiones en la zona periventricular, que es el área alrededor de los ventrículos cerebrales. Las lesiones en esta región son muy comunes y son uno de los criterios para diagnosticar la EM, por lo tanto su detección puede ser un indicativo de alto riesgo de conversión de SCA a EM. Estas lesiones pueden interrumpir las vías que conectan diferentes partes del cerebro, afectando diversas funciones neurológicas (Peñailillo et al., 2019).

15. **Cortical_MRI**: Se refiere a la presencia o ausencia de lesiones en la corteza cerebral, esta es la capa externa del cerebro la cual esta involucrada en funciones de alto nivel como la memoria, la atención, la percepción, el pensamiento, el lenguaje y la conciencia. estas lesiones son indicativas de una posible progresión hacia etapas más severas de EM (Peñailillo et al., 2019).

16. **Infratentorial_MRI**: Esta variable determina si el paciente presenta lesiones en la región infratentorial del cerebro, que incluye el cerebelo y el tronco encefálico. Estas áreas contra funciones vitales como el equilibrio, la coordinación, el habla y la deglución (Peñailillo et al., 2019).

17. **Spinal_Cord_MRI**: Indica si el paciente presenta o no lesiones en la medula espinal. estas lesiones medulares son comunes en la EM y pueden provocar síntomas como debilidad muscular, alteraciones en la sensación, y problema con el control de la vejiga y el intestino (Peñailillo et al., 2019).

18. **initial_EDSS**: Estado inicial en el estudio, siguiendo la escala ampliada del estado de discapacidad. Este metodo cuantifica la discapacidad presente en pacientes, por lo que es util para evaluar en que grado se encuentra una persona con posible progreso a EM (Trust, 2022).

19. **final_EDSS**: Estado final en el estudio, siguiendo la escala ampliada del estado de discapacidad.

20. **Group**: Clasificación del paciente al final del estudio, indicando si el paciente a desarrollado esclerosis múltiple clínicamente definida o no. esta es la variable objetivo del estudio.

| **Nombre de la Variable** | **Tipo de Dato** | **Rango de Valores** |
|---|---|---|
| ID | Entero | [0, 272] |
| Age | Entero | [15, 77] (en años) |
| Schooling | Entero | [0, 25] (en años) |
| Gender | Nominal | 1 = masculino, 2 = femenino |
| Breastfeeding | Nominal | 1 = sí, 2 = no, 3 = desconocido |
| Varicella | Nominal | 1 = positivo, 2 = negativo, 3 = desconocido |
| Initial_Symptom | Nominal | 1 = visual, 2 = sensorial, 3 = motor, 4 = otro,<br>5 = visual y sensorial, 6 = visual y motor,<br>7 = visual y otro, 8 = sensorial y motor,<br>9 = sensorial y otro, 10 = motor y otro,<br>11 = visual, sensorial y motor,<br>12 = visual, sensorial y otro,<br>13 = visual, motor y otro,<br>14 = sensorial, motor y otro,<br>15 = visual, sensorial, motor y otro |
| Mono_or_Polysymptomatic | Nominal | 1 = monosintomático, 2 = polisintomático,<br>3 = desconocido |
| Oligoclonal_Bands | Nominal | 0 = negativo, 1 = positivo, 2 = desconocido |
| LLSSEP | Nominal | 0 = negativo, 1 = positivo |
| ULSSEP | Nominal | 0 = negativo, 1 = positivo |
| VEP | Nominal | 0 = negativo, 1 = positivo |
| BAEP | Nominal | 0 = negativo, 1 = positivo |
| Periventricular_MRI | Nominal | 0 = negativo, 1 = positivo |

| **Nombre de la Variable** | **Tipo de Dato** | **Rango de Valores** |
|---|---|---|
| Cortical_MRI | Nominal | 0 = negativo, 1 = positivo |
| Infratentorial_MRI | Nominal | 0 = negativo, 1 = positivo |
| Spinal_Cord_MRI | Nominal | 0 = negativo, 1 = positivo |
| Initial_EDSS | Entero | Puntuación en la escala al diagnóstico inicial |
| Final_EDSS | Entero | Puntuación en la escala al final del estudio |
| group | Nominal | 1 = EMCD, 2 = no EMCD |

Tabla 1: Descripción de Variables del Conjunto de Datos

# 3. Análisis Estadístico e Inferencial

## 3.1. Análisis Descriptivo

### 3.1.1. Resumen descriptivo datos numéricos

Para obtener el resumen descriptivo de las variables numéricas se separaron los grupos de pacientes con esclerosis múltiple clínicamente definida (EMCD) y los que no (No EMCD), teniendo así las tablas 2 y 3 respectivamente.

| | **count** | **mean** | **std** | **min** | **25 %** | **50 %** | **75 %** | **max** |
|---|---|---|---|---|---|---|---|---|
| **Age** | 125.0 | 34.84 | 11.417587 | 15.0 | 27.0 | 34.0 | 41.0 | 70.0 |
| **Schooling** | 125.0 | 16.024 | 4.104411 | 0.0 | 15.0 | 15.0 | 20.0 | 25.0 |
| **Initial_EDSS** | 125.0 | 1.36 | 0.587504 | 1.0 | 1.0 | 1.0 | 2.0 | 3.0 |
| **Final_EDSS** | 125.0 | 1.448 | 0.65323 | 1.0 | 1.0 | 1.0 | 2.0 | 3.0 |

Tabla 2: Resumen estadístico para los pacientes que desarrollaron EMCD.

| | **count** | **mean** | **std** | **min** | **25 %** | **50 %** | **75 %** | **max** |
|---|---|---|---|---|---|---|---|---|
| **Age** | 148.0 | 33.405405 | 10.847011 | 16.0 | 25.0 | 32.5 | 39.25 | 77.0 |
| **Schooling** | 147.0 | 14.455782 | 4.2414 | 6.0 | 12.0 | 15.0 | 20.0 | 25.0 |
| **Initial_EDSS** | 0.0 | NA | NA | NA | NA | NA | NA | NA |
| **Final_EDSS** | 0.0 | NA | NA | NA | NA | NA | NA | NA |

Tabla 3: Resumen estadístico para los pacientes que NO desarrollaron EMCD.

A partir de las tablas 2 y 3 se tiene lo siguiente:

- **Age:** Ambos grupos, los pacientes que desarrollaron EMCD y los que no, presentan una amplia gama de edades. Sin embargo, la media de edad es ligeramente mayor en el grupo que desarrolló EMCD, lo que podría indicar una tendencia hacia la conversión de SCA a EM en pacientes de mayor edad. Por otro lado, la variabilidad en las edades

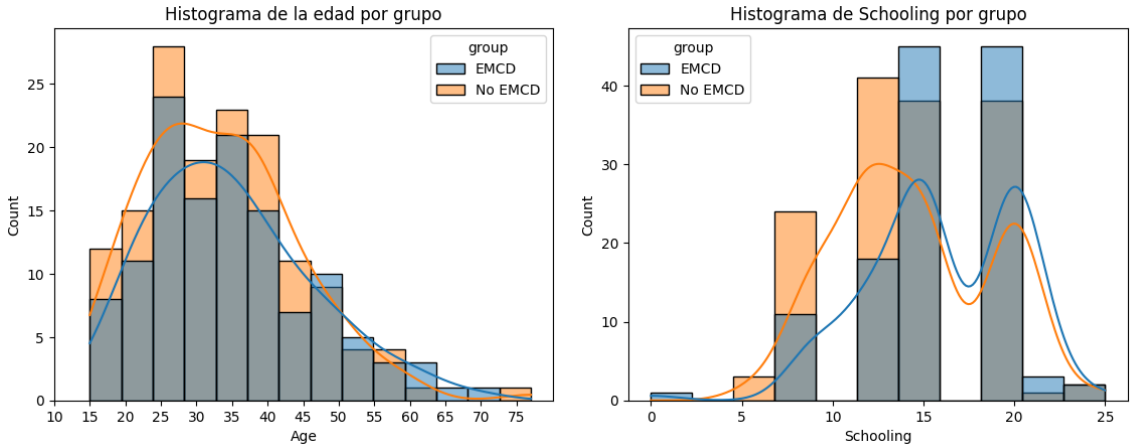
es ligeramente más alta en el grupo que no desarrollo EMCD, lo que sugiere una mayor heterogeneidad en este grupo respecto a la edad.

- **Schooling:** Los datos muestran que los pacientes que desarrollaron EMCD tienden a tener un nivel de escolaridad promedio más alto. Caso que no se condice con lo descrito en la literatura (UNESCO, 2019). De cualquier modo, es un análisis preliminar que se debe trabajar en apartado de analisis diferencial, para verificar el nivel de significancia de esta variable dentro del desarrollo de SCA a EM

- **Cambio en nivel de EDSS:** Dado que se tienen datos registrado solo para el grupo con EMCD, solo podemos observar que para este grupo existe un leve aumento en la media del EDSS final en relación con el inicial, lo que sugiere un desarrollo en el nivel de discapacidad según este método de medición.

### 3.1.2. Histogramas

A continuación se realiza un análisis de los histogramas 1 de las edades y años de escolaridad divido en los grupos EMCD y No EMCD, cabe destacar que no se tomo en consideración las variables de EDSS, dado que existen solo datos para el grupo con EMCD:




(a) Histograma Edad (b) Histograma Escolaridad

Figura 1: Histogramas variables numéricas

- **Edad:** En el grupo con EMCD, la edad muestra una distribución binomial negativa, con un pico en el segmento de 30-35 años. Esta distribución sugiere un número mayor de diagnósticos de EMCD en el rango de 25-40 años. Por otra parte, la distribución en el grupo sin EMCD exhibe una mayor uniformidad en la frecuencia a través de un rango de edad más amplio, con un leve incremento en la proporción de individuos más jóvenes, particularmente entre los 20 y 25 años, lo que podría reflejar la edad de aparición de síntomas iniciales en estos pacientes.

- **Escolaridad:** La escolaridad en el grupo con EMCD se distribuye de manera bimodal, indicando dos grupos predominantes: uno que ha completado la educación secundaria, con un pico en 15 años de estudio, y otro con educación universitaria, marcado por el pico alrededor de los 20 años. Esto sugiere diversidad en los niveles de educación alcanzados por los pacientes con EMCD. Por su parte, el grupo sin EMCD comparte esta bimodalidad, pero con una ligera tendencia hacia niveles de escolaridad más bajos, lo que podría implicar diferencias en el perfil de acceso a la educación entre los grupos.

### 3.1.3. Frecuencia datos categóricos

A partir de los gráficos de las figuras 2, 3, 5, 4 y 6 para la variables categóricas dividiendo en los grupos de EMCD y No EMCD, se presenta el siguiente análisis en base a la frecuencia de pacientes en cada grupo y categoría:

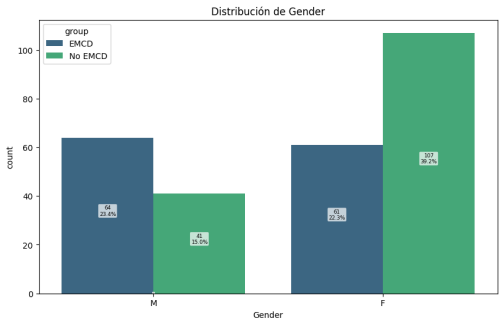



Figura 2: Distribución de variable Gender

- **Genero:** La distribución de pacientes muestra que la proporción del genero masculino en relación con el femenino presenta un mayor desarrollo de SCA a EM a diferencia de

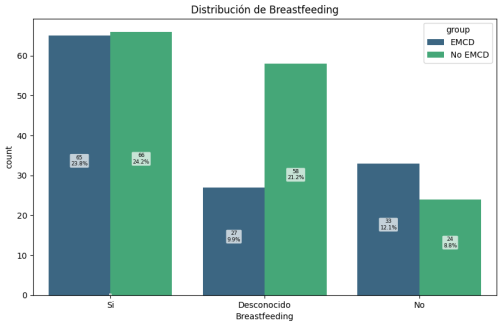
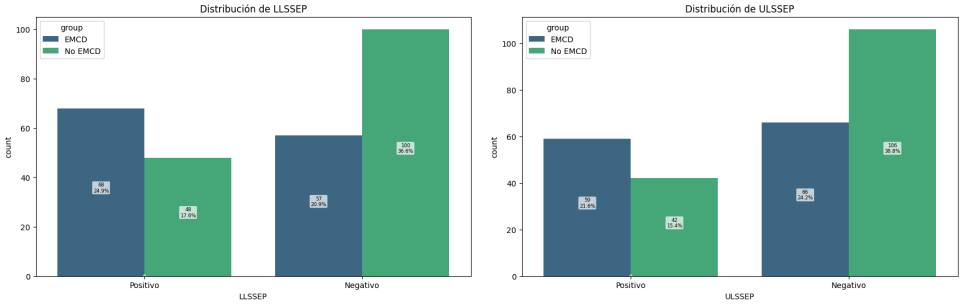
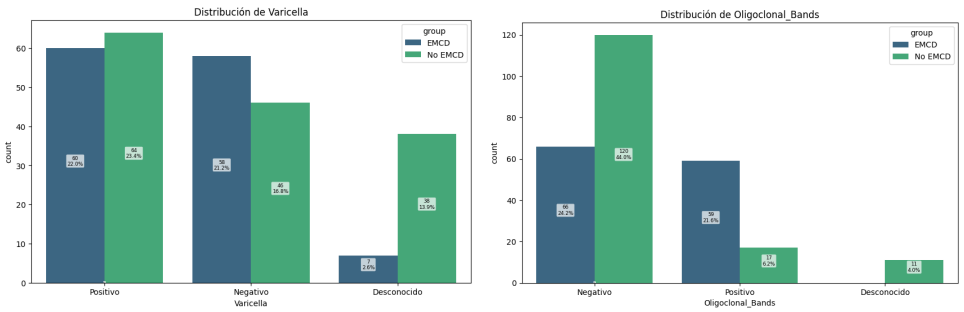



Figura 3: Distribución de variable Breastfeeding




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

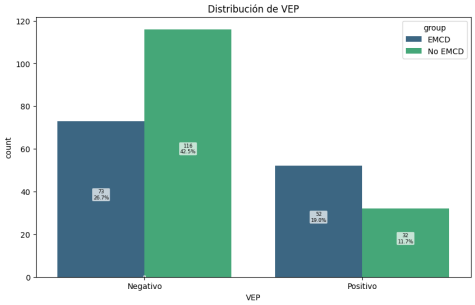
Figura 4: Distribuciones de resultados exámenes SSEP



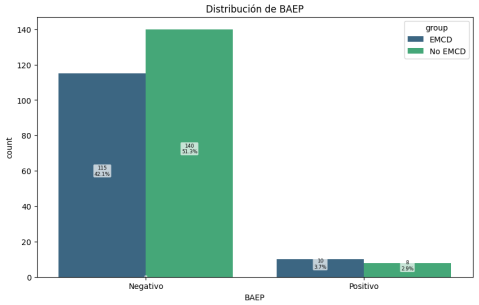

(a) Distribución de Varicella




(b) Distribución de Oligoclonal_Bands




(c) Distribución de VEP



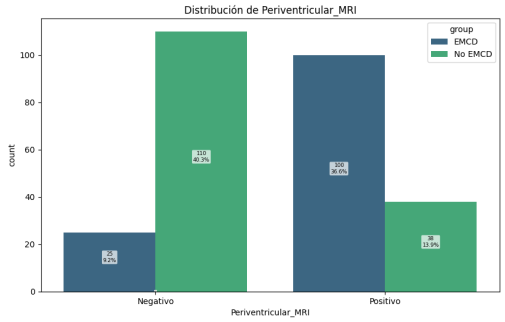

(d) Distribución de BAEP

Figura 5: Distribuciones de resultados de diversos exámenes

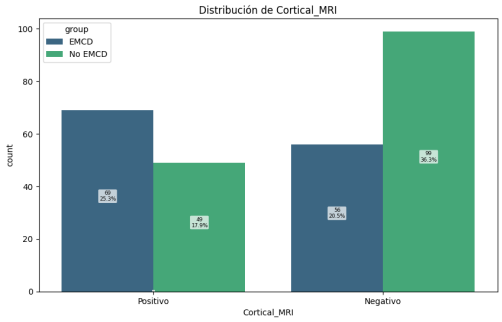
la literatura.

- **Lactancia materna:** En el gráfico sobre la lactancia materna se presenta una distribución similar para el caso de los pacientes que si fueron amamantados entre los grupos EMCD y No EMCD, teniendo una pequeño porcentaje mas alto en el grupo No EMCD. La categoría 'Desconocido' es más frecuente en el grupo No EMCD, mientras que la proporción de quienes no fueron amamantados es mayor en el grupo EMCD, lo que coincide con la investigación (Sara Collorone, 2022) revisada, esto podría ofrecer un punto de investigación sobre el posible rol de la lactancia materna en el desarrollo de EM, sobre este conjunto de datos.

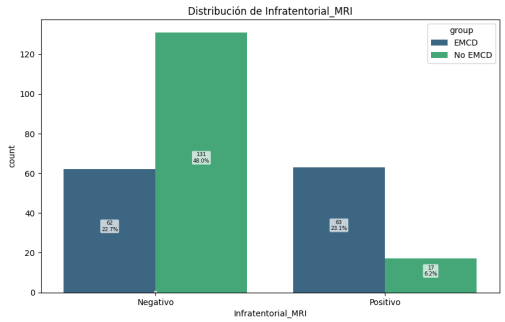
- **Varicela:** El historial de varicela muestra una distribución similar entre aquellos con un historial positivo en ambos grupos, pero hay una diferencia notable en aquellos con un historial negativo, siendo menor en el grupo No EMCD. Además se observa que la diferencia entre positivo y negativo para el grupo No EMCD se puede tomar en consideración para estudio posterior, a pesar de que no coincide con la literatura



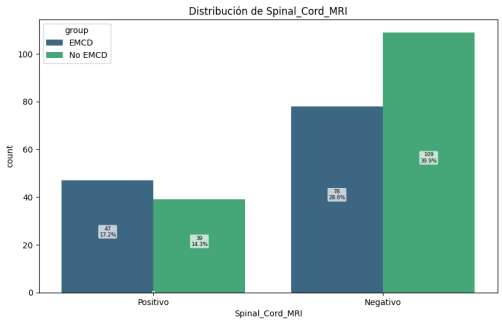

(a) Distribución de Periventricular_MRI




(b) Distribución de Cortical_MRI




(c) Distribución de Infratentorial_MRI




(d) Distribución de Spinal_Cord_MRI

Figura 6: Distribuciones de MRIs

(Bermudez et al., 2016)

- **Síntomas Iniciales:** Con respecto a los síntomas iniciales, podemos observar que los pacientes con el tipo 11, 12, 13, 14 y 15 presentan una gran diferencia, teniendo una mayor cantidad de pacientes que desarrollo EM, con esos síntomas iniciales. Por otro lado, en su mayoría los pacientes con los síntomas iniciales 1 y 2 no desarrollaron EM. Los síntomas iniciales restantes están más equilibrados en cuanto a cantidad de pacientes con EMCD y sin EMCD. En base a lo anterior, podemos observar que los síntomas iniciales podrían ser una variable interesante a tener en cuenta para el análisis posterior.

- **Mono o Polisintomático:** En ambos grupos, la presencia de síntomas polisintomáticos es más común que los monosintomáticos, lo cual podría indicar que la presentación polisintomática no es un factor distintivo en la progresión a EMCD. Por otro lado, para los pacientes monosintompaticos se observa una mayor frecuencia en el grupo No EMCD, lo que podría indicar que aquellos que presentan un solo síntoma en su evento de SCA estén menos cercanos a desarrollar EM. Por ultimo, la categoría 'Desconocido' es marginal en ambos grupos.

- **Bandas Oligoclonales:** La presencia de bandas oligoclonales es ampliamente mayor en el grupo EMCD. Esto es coincide con los estudios, ya que la presencia de bandas oligoclonales se asocia comúnmente con EM («Conversion Predictors of CIS to Multiple Sclerosis», 2023).

- **LLSSEP y ULSSEP:** Los resultados de LLSSEP (Potenciales Evocados Somatosensoriales de Extremidades Inferiores) muestran una diferencia clara entre los grupos, con una mayor proporción de resultados negativos en el grupo No EMCD. Esto podría sugerir que anormalidades en LLSSEP están asociadas con la progresión a EMCD. Una observación similar se presenta en ULSSEP (Potenciales Evocados Somatosensoriales de Extremidades Superiores), con mayor prevalencia de resultados negativos en el grupo No EMCD, reflejando posiblemente la misma tendencia. Lo anterior tiene relación con estudios previos, ya que el diagnostico positivo de estos estudios, representa un

síntoma común en la EM (Wagner et al., 2021).

- **VEP y BAEP:** Los resultados del VEP (Potenciales Evocados Visuales) muestran una mayor proporción de resultados negativos en el grupo No EMCD. Por otro lado, los resultados del BAEP (Potenciales Evocados Auditivos del Tronco Cerebral) muestran una amplia mayoría de resultados negativos en el grupo No EMCD, lo que podría indicar que anormalidades en el estudio VEP y BAEP no son comunes en pacientes que no desarrollan EMCD, como se encuentra en la literatura, ya que el resultados positivo de estos exámenes se relaciona con un síntoma común en la EM («Conversion Predictors of CIS to Multiple Sclerosis», 2023).

- **Imágenes por resonancia magnética (MRIs):** Para los resultados de imagenología, se tiene que los resultados negativos de lesiones en las diversas áreas están asociados en amplia mayoría con el grupo que no desarrolla EM. Además, los resultados positivos de lesiones están relacionados principalmente con el grupo que desarrollo EM. En síntesis, la presencia de lesiones en el área del cerebro o en la medula espinal, podría indicar en coincidencia con investigaciones, que el paciente desarrollara finalmente EM.

## 3.2. Análisis Inferencial

### 3.2.1. Pruebas estadísticas variables numéricas

Dado que las pruebas de normalidad indicaron que las distribuciones de las variables numéricas como la edad y la escolaridad no cumplen con los supuestos de normalidad, se optó por aplicar la prueba de Mann-Whitney U. Esta prueba no paramétrica permite comparar las medianas entre dos grupos independientes, en este caso, el grupo de pacientes con Esclerosis Múltiple Clínicamente Definida (EMCD) y el grupo sin EMCD. Las hipótesis planteadas para estas pruebas son las siguientes:

- **Hipótesis nula ($H_0$):** No existe una diferencia significativa en las medianas de la variable numérica entre los pacientes que desarrollaron EM y los que no.

- **Hipótesis alternativa ($H_A$):** Existe una diferencia significativa en las medianas de la variable numérica entre los pacientes que desarrollaron EM y los que no.

| Variable | u-statistic | p-value |
|---|---|---|
| Age | 9834.5 | 0.368620 |
| Schooling | 11366.0 | 0.000751 |

Tabla 4: Resultados de la prueba de Mann-Whitney U para las variables numéricas.

Según los resultados obtenidos en la tabla 4, se observa lo siguiente:

- **Edad:** Se falla en rechazar la hipótesis nula, dado que el p-value es mayor a 0.05, entonces no hay una diferencia significativa en las medianas de las edades entre el grupo con EMCD y sin EMCD. Por lo tanto, no se puede afirmar que la edad sea un factor determinante para la presencia de EMCD.
- **Escolaridad:** Se rechaza la hipótesis nula en favor de la alternativa con una confianza del 95 %, dado que el p-value es menor a 0.05, entonces hay una diferencia significativa en las medianas en la escolaridad entre el grupo con EMCD y sin EMCD. Por lo tanto, se puede afirmar que la escolaridad es un factor determinante para la presencia de EMCD.

### 3.2.2. Pruebas estadísticas variables categóricas

Para analizar la relación entre las variables categóricas y la transición de SCA a EM, se emplearon métodos estadísticos adecuados según el tipo de variable. Para las variables dicotómicas, se optó por la prueba exacta de Fisher, dada su precisión en muestras pequeñas y su capacidad para manejar tablas de contingencia 2x2. En el caso de variables categóricas con más de dos categorías, se utilizaron pruebas de Chi-cuadrado, que ayudan a determinar si las diferencias en las distribuciones de frecuencias son estadísticamente significativas. Las hipótesis planteadas para estas pruebas estadísticas son las siguientes:

- **Hipótesis nula ($H_0$):**No existe relación entre la variable categórica y el desarrollo de EMCD
- **Hipótesis alternativa ($H_A$):** Existe relación entre la variable categórica y el desarrollo de EMCD

| Variable | Prueba | Chi2/Odds Ratio | p-value | dof | Sign. |
|---|---|---|---|---|---|
| Gender | Fisher | 2.738105 | 1.002996e-04 | 1 | ** |
| Breastfeeding | Chi2 | 10.874022 | 4.352472e-03 | 2 | ** |
| Varicella | Chi2 | 21.081106 | 2.644210e-05 | 2 | ** |
| Initial_Symptom | Chi2 | 57.108698 | 3.745727e-07 | 14 | ** |
| Mono_or_Polysymptomatic | Chi2 | 7.684718 | 2.144296e-02 | 2 | * |
| Oligoclonal_Bands | Chi2 | 48.292996 | 3.260687e-11 | 2 | ** |
| LLSSEP | Fisher | 0.402353 | 3.491561e-04 | 1 | ** |
| ULSSEP | Fisher | 0.443236 | 1.637113e-03 | 1 | ** |
| VEP | Fisher | 0.387268 | 5.783209e-04 | 1 | ** |
| BAEP | Fisher | 0.657143 | 4.660641e-01 | 1 | |
| Periventricular_MRI | Fisher | 0.086364 | 8.699969e-20 | 1 | ** |
| Cortical_MRI | Fisher | 0.401698 | 3.575146e-04 | 1 | ** |
| Infratentorial_MRI | Fisher | 0.127711 | 1.202655e-12 | 1 | ** |
| Spinal_Cord_MRI | Fisher | 0.593793 | 5.065238e-02 | 1 | |

Tabla 5: Resultados de las pruebas estadísticas para variables categóricas.

Dada la tabla 5 que contiene los resultados de las pruebas se identifica lo siguiente:

- Las variables BAEP y Spinal_Cod_MRI tienen un p-value mayor a 0.05, por lo que se falla en rechazar la hipótesis nula, es decir, no existe relación significativa entre estas variables y el desarrollo de EMCD.
- Por el contrario, las variables Gender, Breastfeeding, Varicella, Initial_Symptoms, Mono_or_Polysymptomatic, Oligoclonal_Bands, LLSSEP, ULSSEP, VEP, Periventicular_MRI, Cortical_MRI e Infratentorial_MRI, presentan un p-value menor a 0.05, por lo que se rechaza la hipótesis nula en favor de la hipótesis alternativa, es decir, existe relación significativa entre estas variables y el desarrollo de EMCD.

### 3.2.3. Modelo de Regresión Logística

Para desarrollar el modelo de regresión logística, inicialmente se examinó la multicolinealidad entre las variables involucradas mediante el cálculo del Factor de Inflación de la Varianza (VIF). Se establecieron dos umbrales para la selección de variables: un VIF superior a 10 y otro superior a 5. A partir de esta evaluación, en la primera instancia, se eliminaron cinco variables que mostraron alta multicolinealidad: Gender, Age, Schooling, Varicella y Mono_or_Polysymptomatic. Tras eliminar estas variables, se procedió a construir el modelo de regresión logística, ya que en segunda instancia no superaban el umbral de VIF superior a 5. La evaluación del modelo incluyó la generación de la matriz de confusión 7, la curva ROC 8 y el cálculo de diversas métricas 6 sobre el conjunto de datos de validación.

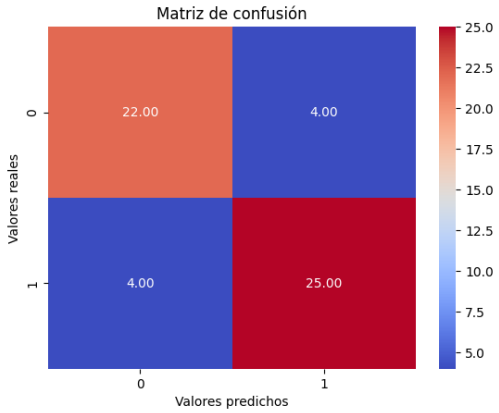



Figura 7: Matriz de confusión datos validación

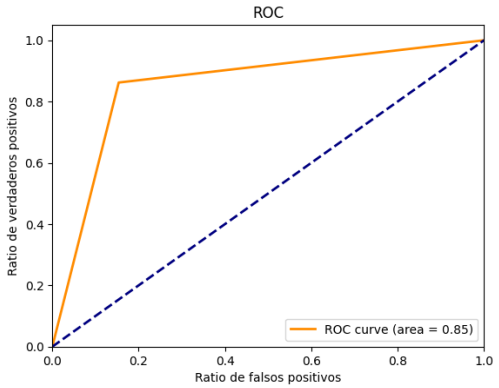



Figura 8: ROC datos validación

| Clase | Precisión | Recall | F1-Score |
|---|---|---|---|
| 1 | 0.85 | 0.85 | 0.85 |
| 2 | 0.86 | 0.86 | 0.86 |

Tabla 6: Métricas de rendimiento del modelo de regresión logística

- **Matriz de confusión:** muestra que el modelo predijo correctamente 22 casos de presencia de EMCD y 25 casos sin presencia, mientras que hubo 4 falsos positivos y 4 falsos negativos. Esto indica un buen balance en la capacidad del modelo para identificar ambas clases. Además, no parece haber un sesgo hacia una clase en particular.

- **Curva ROC:** El modelo demuestra una capacidad de clasificación robusta con un área bajo la curva (AUC) de 0.85. Un AUC de 1 indicaría un modelo perfecto, mientras que un AUC de 0.5 sugiere un desempeño no mejor que el azar (Narkhede, 2021). Por ende, un AUC de 0.85 refleja una buena habilidad del modelo para diferenciar entre pacientes que desarrollaron EMCD y quienes no.

- **Modelo de regresión logística:** ha demostrado un rendimiento consistente y equilibrado en la clasificación de ambas categorías, con una precisión, un recall y un F1-Score de aproximadamente 0.85 para la clase 1 (EMCD) y ligeramente superior, 0.86, para la clase 2 (No EMCD). Estos valores indican que el modelo tiene una alta tasa de aciertos en sus predicciones y es eficaz en identificar correctamente los casos positivos de cada clase, asegurando un balance saludable entre la precisión y la sensibilidad y confirmando su fiabilidad en la discriminación de las clases.

En conjunto, estos resultados sugieren que el modelo de regresión logística está distinguiendo bien entre los pacientes que desarrollaron EMCD y quienes no. No muestra un sesgo significativo hacia una clase en particular y demuestra un buen equilibrio entre la precisión y la sensibilidad para ambas clases. Sin embargo, se podría llegar a mejorar el modelo investigando los casos específicos de faltosos positivos y negativos, para entender mejor en que circunstancias se confunde entre pacientes que desarrollaron EMCD y quienes no.

# 4. Conclusiones

Los resultados del análisis estadístico e inferencial realizado en este estudio proveen una comprensión profunda de las características clínicas y factores de riesgo asociados con la conversión de SCA a EM. Mediante el uso de técnicas de estadística descriptiva, pruebas estadísticas y modelado predictivo, se ha podido identificar varias variables significativas que influyen en la progresión de SCA a EM, lo cual es esencial para la intervención temprana de los pacientes.

El análisis descriptivo reveló diferencias notables en variables como edad, escolaridad, y resultados de pruebas clínicas entre los pacientes que desarrollaron EM y los que no. En particular, la escolaridad mostró una relación significativa con la conversión de SCA a EM, sugiriendo que niveles más altos de educación podrían estar asociados con un mayor riesgo de desarrollar EM. Además, los análisis inferenciales indicaron que variables como la presencia de bandas oligoclonales, resultados de MRIs, y síntomas iniciales están fuertemente relacionadas con la progresión de la enfermedad.

La aplicación de un modelo de regresión logística permitió estimar la probabilidad de conversión de SCA a EM con una precisión y sensibilidad considerables. Este modelo destacó por su capacidad para manejar eficazmente los desafíos del diagnóstico precoz de la EM, proporcionando una herramienta valiosa para los clínicos en la toma de decisiones. Sin embargo, es crucial validar este modelo con datos actuales para asegurar su relevancia y efectividad en el tiempo. Además, sería beneficioso evaluar el rendimiento del modelo en diferentes contextos y poblaciones para entender su aplicabilidad y limitaciones en escenarios variados.

La relevancia de este estudio reside en su contribución al entendimiento del SCA y su potencial conversión a EM, subrayando la importancia de una evaluación temprana y continua de los pacientes con SCA. Los hallazgos sugieren que, además de los tratamientos clínicos, factores socioeducativos y demográficos deben ser considerados en el manejo y seguimiento de estos pacientes.

En conclusión, este estudio aporta evidencia significativa que refuerza la identificación temprana y precisa de los pacientes con SCA que tienen un alto riesgo de desarrollar

EM. Este enfoque mejora la capacidad de monitorizar y gestionar de manera efectiva estos casos, potencialmente influenciando de manera positiva su trayectoria clínica y calidad de vida. A futuro, sería recomendable expandir este análisis a cohortes más grandes y diversas para validar y refinar los modelos predictivos desarrollados, garantizando su aplicabilidad en diferentes contextos poblacionales.

# 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/

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

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

Narkhede, S. (2021). Understanding AUC-ROC Curve [Último acceso: 21 de abril de 2024]. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

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

Pineda, B., & Flores Rivera, J. D. J. (2023). Conversion predictors of Clinically Isolated Syndrome to Multiple Sclerosis in Mexican patients: a prospective study. https://doi.org/10.17632/8wk5hjx7x2.1

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.

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

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