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

Laboratorio 4 - Arboles de decisión

Integrantes: Matías Figueroa Contreras

Curso: Análisis de Datos

Profesor: Max Chacón Pacheco

Ayudante: Marcelo Álvarez

11 de Agosto 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 arboles de decisión los cuales permiten obtener relaciones entre las variables de un conjunto de datos (Hashemi-Pour & Lutkevich, 2024). El uso arboles de decisión 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 arboles de decisión, 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 C5.0, ajustando los parámetros que este tiene para conseguir buenas medidas de calidad al evaluar los arboles generados con el algoritmo.
- Identificar y describir las ramas distintivas en el árbol de decisión 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 secuencias de progresión de la enfermedad.

2. Marco Teórico

2.1. Arboles de decisión

Los árboles de decisión son modelos de predicción utilizados en estadística, minería de datos y aprendizaje automático. Estos modelos se emplean para dividir un conjunto de datos en subconjuntos basándose en pruebas lógicas sucesivas. Los árboles de decisión constan de nodos de decisión, nodos hoja y ramas que representan las decisiones o divisiones basadas en los atributos de los datos (IBM, 2024a). Dentro de estos existen algunos conceptos clave para la generación de los arboles de decision:

2.1.1. Entropía

La entropía es una medida de la incertidumbre o impureza en un conjunto de datos. En el contexto de los árboles de decisión, se utiliza para determinar la mejor manera de dividir los datos en subconjuntos homogéneos (IBM, 2024a). La entropía se define como:

$$\text{Entropía}(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

donde p_i es la proporción de elementos de la clase i en el conjunto de datos S.

2.1.2. Ganancia de Información

La ganancia de información es una métrica utilizada para seleccionar el atributo que mejor separa un conjunto de datos en clases. Se calcula como la reducción de la entropía después de dividir un conjunto de datos en función de un atributo (IBM, 2024a):

$$\text{Ganancia de Información}(S, A) = \text{Entropía}(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} \text{Entropía}(S_v) \tag{2}$$

donde S_v es el subconjunto de S donde el atributo A tiene el valor v.

2.1.3. Poda

La poda es una técnica utilizada para reducir el sobreajuste en los árboles de decisión. Existen dos tipos principales de poda (Numerentur, 2024):

- **Poda previa:** Detiene la construcción del árbol antes de que se haya ajustado completamente a los datos de entrenamiento.
- **Poda posterior:** Elimina partes del árbol después de que se haya construido, generalmente basado en una evaluación de su rendimiento en un conjunto de validación.

2.2. Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar la capacidad de generalización de un modelo. En la validación cruzada k-fold, los datos se dividen en k subconjuntos, y el modelo se entrena k veces, cada vez utilizando $k-1$ subconjuntos para el entrenamiento y el subconjunto restante para la validación. Este proceso ayuda a garantizar que el modelo no esté sobreajustado a un conjunto específico de datos de entrenamiento (Datascientest, 2024).

2.3. C5.0

C5.0 es un algoritmo avanzado para la generación de árboles de decisión, que es una mejora del algoritmo C4.5. Este algoritmo es conocido por su rapidez y eficacia en la clasificación de datos, y permite manejar grandes volúmenes de datos y trabajar con variables categóricas y continuas (IBM, 2024b).

2.4. Métricas de Calidad

Las métricas de calidad son esenciales para evaluar el rendimiento de un modelo de árbol de decisión. Las métricas utilizadas son las siguientes:

2.4.1. Precisión (Accuracy)

La precisión es la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el total de predicciones realizadas (Barrios, 2024). Se define como:

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

donde TP son los verdaderos positivos, TN los verdaderos negativos, FP los falsos positivos y FN los falsos negativos.

2.4.2. Sensibilidad (Recall o Tasa de Verdaderos Positivos)

La sensibilidad mide la capacidad del modelo para identificar correctamente las instancias positivas (Barrios, 2024). Se define como:

$$\text{Sensibilidad} = \frac{TP}{TP + FN} \tag{4}$$

2.4.3. Especificidad (Tasa de Verdaderos Negativos)

La especificidad mide la capacidad del modelo para identificar correctamente las instancias negativas (Barrios, 2024). Se define como:

$$\text{Especificidad} = \frac{TN}{TN + FP} \tag{5}$$

2.4.4. Valor F1

El valor F1 es la media armónica de la precisión y la sensibilidad, proporcionando un balance entre ambas (Barrios, 2024). Se define como:

$$\text{Valor F1} = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \tag{6}$$

3. Obtención del Árbol

3.1. Pre-procesamiento

Antes de proceder con la obtención de arboles de decisión, se realizo un pre-procesamiento de los datos. Durante esta fase se realizo el mismo procedimiento de limpieza de datos y discretización de variables numéricas que el laboratorio 3 (Contreras, 2024), y se vario ligeramente en el proceso de codificación, teniendo en cuenta el procedimiento necesario a realizar sobre las variables para trabajar con el algoritmo C50. En resumen se imputaron los valores de las variables *Schooling*, *Initial_Symptom*, *Mono_or_polysintomatic*, *Varicela* y *Oligoconal_Bands*, según la naturaleza de cada variable buscando mantener la integridad del conjunto de datos. Por otro lado, se excluyeron las variables *Initial_EDSS* y *Final_EDSS* debido a la ausencia completa de datos para la clase No EMCD. Ademas, se elimino la variable *Breastfeeding* debido a la alta incidencia de datos faltantes. Por otra parte, se discretizaron las variables numéricas **Edad** y **Escolaridad**, teniendo los resultados en las tablas 1 y 2 respectivamente. Por ultimo, se realizo el mismo proceso para la variable *Initial_Symptom* que en el laboratorio 3 (Contreras, 2024), transformando la combinación de síntomas que contenía esta variable entre *Sensorial*, *Motor*, *Visual* y *Otro*, en múltiples columnas booleanas dentro del conjunto de datos para reflejar la presencia o ausencia de cada síntoma. Luego, todas las variables luego del pre-procesamiento anterior fueron transformado al tipo de dato factor, con el fin de poder aplicar el algoritmo C50 y así generar el árbol de decisión.

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

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

3.2. Generación del árbol de decisión

Para la generación y evaluación del árbol de decisión, se dividió el conjunto de datos en dos partes: un conjunto de entrenamiento, siendo el 80 % del total, y un conjunto de prueba con el 20 % restante. Esta división se realizó utilizando la función *createDataPartition* de la librería *caret* en R, la cual asegura mantener la proporción de pacientes que desarrollaron EMCD y los que no, garantizando así que tanto el conjunto de entrenamiento como el de prueba sean representativos del conjunto de datos original en términos de distribución de las clases de la variable respuesta (Kili Technology, 2024).

Inicialmente, se generó un modelo preliminar utilizando la función *C5.0* de la librería *C50*, sin aplicar técnicas avanzadas de ajuste para establecer un modelo base. Para mejorar este modelo, se aplicó el proceso de validación cruzada con la función *train* de la librería *caret*, utilizando el método *C5.0*. Este enfoque permitió obtener un árbol de decisión con mejores medidas de calidad al ser evaluado en el conjunto de prueba.

3.2.1. Selección y justificación de medidas de calidad

Para evaluar la efectividad de los árboles de decisión generados, se seleccionaron varias medidas de calidad centradas en la precisión y la capacidad de generalización del modelo. Las métricas elegidas incluyen la precisión (*Accuracy*), sensibilidad (*Sensitivity*), especificidad (*Specificity*), y el valor F1. La precisión fue especialmente prioritaria dado que proporciona una medida directa del porcentaje de predicciones correctas sobre el total de

casos, lo que es crítico en contextos clínicos donde la correcta clasificación de los estados de enfermedad es esencial (Michelli, 2024).

Las métricas se calcularon utilizando una matriz de confusión, que permite una visualización detallada del rendimiento del modelo al diferenciar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Estas métricas permiten evaluar no solo cuán preciso es el modelo, sino también cómo se comporta en términos de evitar falsos diagnósticos (falsos positivos y falsos negativos), lo cual es crucial en este contexto que busca clasificar pacientes que puedan desarrollar o no EMCD (Barrios, 2024).

3.2.2. Resultados obtenidos

Los resultados de los modelos se visualizan en las figuras 1 y 2, mientras que las métricas de calidad obtenidas en el conjunto de prueba se presentan en la tabla 3.

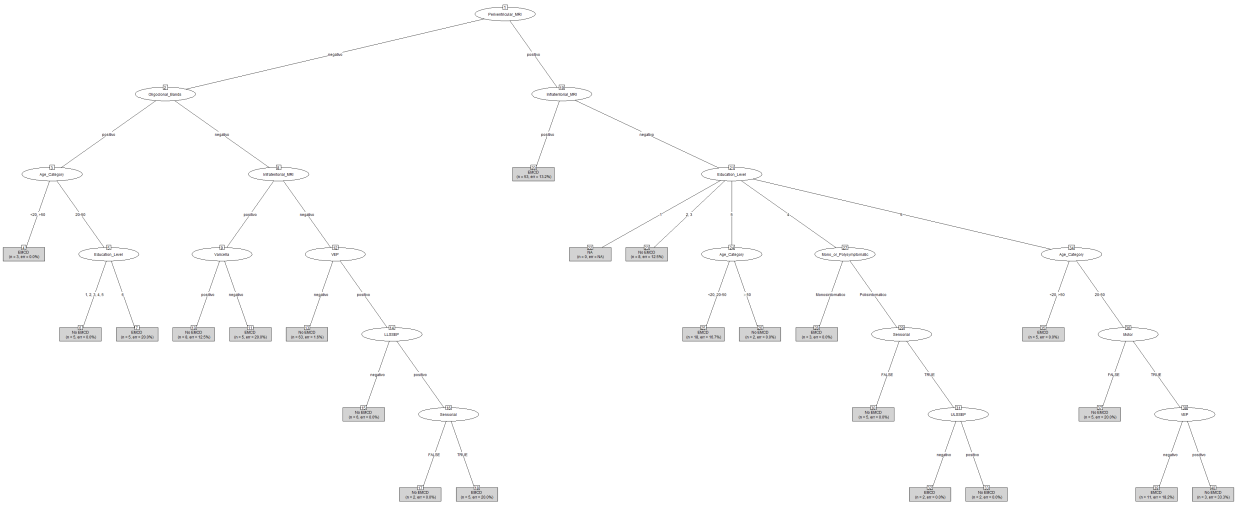

Figura 1: Árbol de decisión generado sin validación cruzada

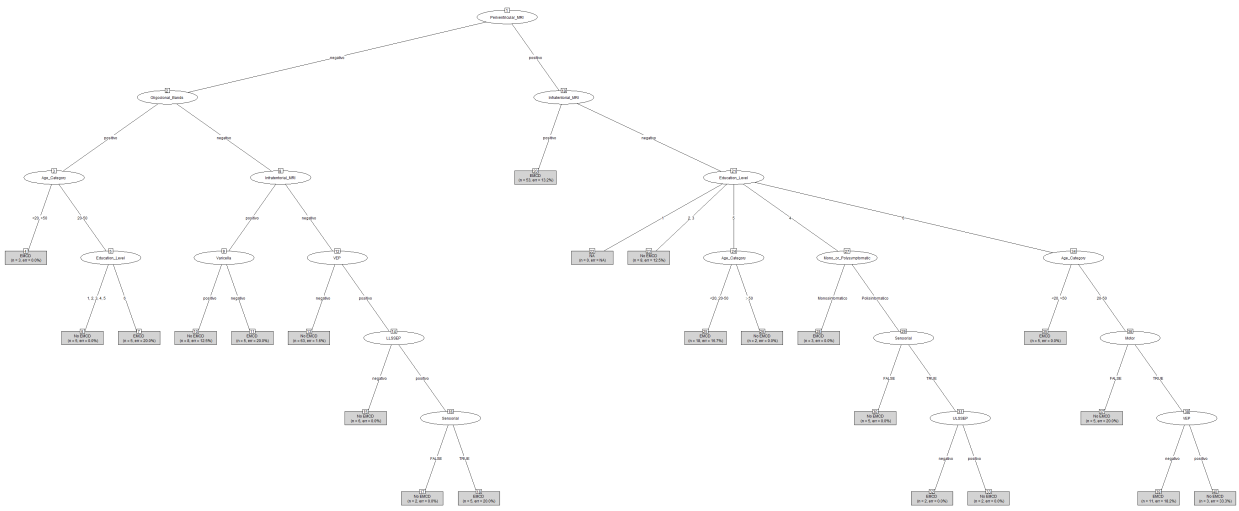

Figura 2: Árbol de decisión mejorado con validación cruzada

Métrica	Valor Sin VC	Valor Con VC
Precisión	0,78	0,815
Sensibilidad	0,68	0,72
Especificidad	0,862	0,897
Valor F1	0,739	0,783

Tabla 3: Comparación de métricas de calidad entre los modelos

4. Análisis de Resultados

Se puede observar en el árbol de decisión generado, que en los primeros nodos se encuentran resultados de exámenes teniendo las resonancias magnéticas periventriculares e infratentoriales, y el examen de bandas oligoclonales, estos exámenes son utilizados como indicadores comunes en el diagnostico de la EM (Peñailillo et al., 2019) y («Esclerosis múltiple - Síntomas y causas», s.f.), concordando así con la literatura en el hecho de que estén como primeros nodos de decisión en el árbol generado, ya que son los que más y mejor separan los grupos de pacientes que desarrollaron EMCD y los que no.

Se encontró la regla que lleva al mayor número de pacientes evaluados es:

```
Periventricular_MRI = negativo -> Oligoclonal_Bands = negativo ->
   Infratentorial_MRI = negativo -> VEP = negativo: No EMCD (63/1).
```

Esta regla evalúa un total de 63 pacientes, con solo 1 error, lo que implica una precisión del 98.41 % en el conjunto de entrenamiento. Estos resultados indican que cuando los exámenes de resonancia magnética periventricular, bandas oligoclonales, resonancia infratentorial y VEP son negativos, hay una alta probabilidad de predecir correctamente la ausencia de EMCD en los pacientes. Este conjunto de resultados negativos proporciona un diagnóstico robusto para descartar EMCD, como se ha revisado en laboratorios anteriores y como lo indica la literatura, estos exámenes son indicativos claros para determinar si una persona con SCA desarrollo o no EMCD (Peñailillo et al., 2019) y («Esclerosis múltiple - Síntomas y causas», s.f.).

Otra regla con un alto nivel de precisión y un número significativo de pacientes evaluados es:

```
Periventricular_MRI = negativo -> Oligoclonal_Bands = negativo ->
   Infratentorial_MRI = positivo -> Varicella = positivo: No EMCD
   (8/1).
```

Esta regla evalúa a 8 pacientes y presenta un solo error, lo que resulta en una precisión del 87.5 % en el conjunto de entrenamiento. Esta combinación específica de resultados diagnósticos sigue la línea que se ha estado mencionando en este análisis, en donde el resultado de los exámenes son claves para determinar el desarrollo o no de EMCD. En esta regla, muestran

una alta probabilidad de no desarrollar EMCD, para los pacientes que tienen una resonancia magnética periventricular negativa y bandas oligoclonales negativas, junto con una resonancia infratentorial positiva y un historial de varicela positiva. Este último valor de varicela positiva, no coincide con la literatura, la cual señala que el haber tenido enfermedades infecciosas, pueden afectar en el desarrollo y la severidad del la EMCD (Bermudez et al., 2016). Lo anterior, puede ser el factor que afecte en la precisión de la regla generada.

Otra regla destacada es:

```
Periventricular_MRI = positivo -> Infratentorial_MRI = positivo:
   EMCD (53/7).
```

Esta regla evalúa un total de 53 pacientes, con 7 errores, lo que resulta en una precisión del 86.79 % en el conjunto de entrenamiento. Esta regla muestra que cuando tanto la resonancia magnética periventricular como la infratentorial son positivas, hay una alta probabilidad de que el paciente desarrolle EMCD. La presencia de lesiones en ambas áreas cerebrales es un fuerte indicador de la progresión hacia EMCD, lo cual es consistente con la literatura, la cual señala que el resultado positivo de estos exámenes de resonancia magnética son claves para el diagnóstico de EM (Peñailillo et al., 2019). Sin embargo, el margen de error del 13.21 % indica que hay casos en los que esta combinación de resultados puede no ser suficiente para un diagnóstico concluyente, lo que podría deberse a variaciones individuales en la presentación de la enfermedad o a la presencia de otras condiciones que afecten ha esta regla generada.

Otra regla significativa es:

```
Periventricular_MRI = positivo -> Infratentorial_MRI = negativo ->
   Education_Level in {2,3}: No EMCD (8/1).
```

Esta regla evalúa a 8 pacientes y presenta 1 error, resultando en una precisión del 87.5 % en el conjunto de entrenamiento. Esta combinación de características muestra que, a pesar de una resonancia magnética periventricular positiva, la ausencia de lesiones infratentoriales junto con un nivel educativo específico entre 1 y 9 años de educación (Educación Primaria y Secundaria) puede ser un indicador confiable para predecir la ausencia de EMCD.

La regla:

```
Periventricular_MRI = positivo -> Infratentorial_MRI = negativo ->
```

```
Education_Level = 5 -> Age_Category in {<20,20-50}: EMCD (18/3)
```

evalúa a 18 pacientes con 3 errores, resultando en una precisión del 83.33 % en el conjunto de entrenamiento. Esta combinación específica de características sugiere una fuerte correlación con el desarrollo de EMCD, aunque con un margen de error mayor que las reglas previamente mencionadas. Este resultado indica que los pacientes con una resonancia periventricular positiva, resonancia infratentorial negativa, un nivel educativo de 5 (correspondiente a educación superior no universitaria) y que se encuentren en las categorías de edad menores de 20 o entre 20 y 50 años, tienen una alta probabilidad de desarrollar EMCD.

Existen otras reglas con precisiones altas de un 100 % en el conjunto de entrenamiento, pero con menor cantidad de pacientes evaluados en cada una. indicando que el modelo es capaz de identificar correctamente en subgrupos mas pequeños o específicos de pacientes si desarrollaron o no EMCD.

Por ejemplo, reglas como:

- ```
  Periventricular_MRI = negativo -> Oligoclonal_Bands = positivo
      -> Age_Category in {<20,>50}: EMCD (3)
  ```

- ```
  Periventricular_MRI = positivo -> Infratentorial_MRI = negativo
      -> Education_Level = 4 -> Mono_or_Polysymptomatic =
      Monosintomatico: EMCD (3)
  ```

- ```
  Periventricular_MRI = negativo -> Oligoclonal_Bands = negativo
      -> Infratentorial_MRI = negativo -> VEP = positivo -> LLSSEP
      = negativo: No EMCD (6)
  ```

- ```
  Periventricular_MRI = positivo -> Infratentorial_MRI = negativo
      -> Education_Level = 6 -> Age_Category in {<20,>50}: EMCD (5)
  ```

- ```
  Periventricular_MRI = negativo -> Oligoclonal_Bands = positivo
      -> Age_Category = 20-50 -> Education_Level in {1,2,3,4,5}: No
       EMCD (5)
  ```
  ```

son ejemplos de reglas con precisión del 100 %. Aunque evalúan un número reducido de pacientes, la capacidad del modelo para predecir correctamente en estos escenarios específicos refuerza la robustez del árbol de decisión. Lo anterior, se debe a que toma más factores para separar grupos de pacientes, teniendo en cuenta factores como la categoría de edad y el nivel educativo. Para la categoría de edad es posible observar que para las reglas que poseen esta variable, se encuentra que las personas menores a 20 y mayores a 50 de años de edad, desarrollan EMCD, y en el caso contrario no desarrollan la enfermedad. Lo anterior, se contradice con la literatura presente sobre la progresión de SCA a EMCD, la cual señala que la mayor cantidad de personas es diagnosticada entre los 20 y 50 años (Brichford, 2024). Por el lado del nivel educativo, no se observan patrones marcados como para señalar algo concreto. Por otro lado, se observa que en general cuando se tiene mayor cantidad de exámenes positivos sobre problemas que provocan los síntomas en los pacientes, se tiende a llegar a un desarrollo de EMCD, como se ha visto en reglas anteriores, fortaleciendo lo mencionado en la literatura sobre el resultado positivo de estos exámenes (Peñailillo et al., 2019) y (Wagner et al., 2021).

En conjunto, estas reglas adicionales resaltan la capacidad del modelo para realizar predicciones precisas incluso en subconjuntos pequeños de pacientes. La alta precisión observada en muchas de estas reglas sugiere que el modelo está bien ajustado para identificar patrones en los datos, y que puede ser utilizado para proporcionar resultados teniendo en cuenta las distintas variables que se presentan en el conjunto de datos de los pacientes.

4.1. Comparación con el laboratorio anterior

En el laboratorio previo, se hizo uso de reglas de asociación usando el algoritmo Apriori, en donde se pudo identificar combinaciones específicas de síntomas y resultados de exámenes que tienen un impacto en la progresión de SCA a EMCD. Al comparar los resultados obtenidos, se observa que coinciden en las variables que se identificaron como importantes para predecir la progresión de SCA a EMCD. En ambos, las resonancias magnéticas, específicamente las lesiones periventriculares e infratentoriales, se destacaron consistentemente como factores clave en la determinación de la progresión hacia EMCD. Esta coincidencia muestra la relevancia de estos exámenes en la identificación temprana de la progresión a EMCD,

coincidiendo con la literatura médica existente (Peñailillo et al., 2019).

En el laboratorio de reglas de asociación, las combinaciones de síntomas y resultados de exámenes que involucraban estas lesiones fueron recurrentes entre las reglas con mayor confianza y lift. Estas reglas sugieren que los pacientes con ciertas configuraciones de resultados de resonancia magnética y síntomas específicos tienen una alta probabilidad de progresar hacia EMCD, lo que refuerza la importancia de estos factores en la predicción de la enfermedad (Contreras, 2024).

Por otro lado, el análisis mediante árboles de decisión también destacó estas mismas variables como nodos clave en el árbol, indicando que son determinantes principales en la clasificación de los pacientes. El árbol de decisión no solo confirmó la importancia de las resonancias magnéticas periventriculares e infratentoriales, sino que también organizó estas variables de manera jerárquica, tomando estas variables como los primeros puntos de decisión al predecir la progresión a EMCD.

En términos de la cantidad de reglas generadas, En el laboratorio basado en reglas de asociación se obtuvieron un total de 15 reglas, seleccionadas en función de umbrales específicos de confianza y soporte (Contreras, 2024). Estas reglas se centraron en identificar combinaciones de síntomas y resultados de exámenes con una alta probabilidad de estar asociadas con la progresión a EMCD. En contraste, el árbol de decisión generó un total de 22 reglas derivadas de sus ramas. Este mayor número de reglas en el enfoque de árboles de decisión se debe a la naturaleza jerárquica del modelo, que explora múltiples rutas posibles a través de las variables para clasificar a los pacientes y a los umbrales establecidos para las reglas de asociación.

En cuanto a la información que entrega cada enfoque, existen diferencias en como se pueden interpretar y utilizar los resultados. En cuanto a las reglas de asociación, estas ofrecen combinaciones especificas de las variables para asociarlas directamente a la progresión a EMCD o no progresión. Esto permite descubrir patrones específicos, limitándose a esas combinaciones exactas, sin ofrecer una jerarquía clara de la importancia de las variables en general. Por lo que, es útil para identificar relaciones directas entre variables especificas, pero no proporciona una visión global de como se relacionan todas las variables dentro del conjunto de datos. Por el contrario, el árbol de decisión organiza la información de manera jerárquica,

mostrando no solo cuáles son las variables importantes, sino también cómo estas interactúan secuencialmente para llevar a la decisión de si el paciente desarrollara o no EMCD. De esta forma no solo se destacan las variables clave, sino que también muestra la importancia de cada variable cuando se va avanzando en el árbol de decisión. Esto permite un panorama más completo y estructurado, y no solo las relaciones especificas entre variables.

En cuanto a la forma de presentar los resultados, ambos enfoques adoptan estrategias distintas. Las reglas de asociación en el caso del laboratorio anterior fueron presentadas en tablas, que detallan cada regla junto con sus métricas de soporte, confianza y lift. Este formato permite una comparación directa de la relevancia y fuerza de cada regla facilitando la identificación de las combinaciones más importantes. Sin embargo, esta presentación puede resultar incomoda de ver y analizar, ya que se tiene que ir revisando cada regla por separado en base a las variables presentadas de algún paciente.

Por otro lado, para el árbol de decisión se presentan los resultados como observamos a través de un diagrama que muestra la estructura jerárquica del modelo. Esta visualización gráfica, facilita de gran manera la clasificación de los pacientes, ya que permite ver las rutas de decisión para cada variable que presenten los pacientes.

5. Conclusiones

El análisis realizado mediante el uso de árbol de decisiones para analizar el desarrollo de EMCD, ha permitido identificar de manera clara y estructurada las variables y reglas más importantes para predecir la progresión hacia EMCD, destacando especialmente las resonancias magnéticas periventriculares e infratentoriales, así como los exámenes de bandas oligoclonales, como los factores determinantes en la clasificación de los pacientes.

El desarrollo de este laboratorio demostró el gran valor de los árboles de decisión en la clasificación y organización jerárquica de la información, lo que permite no solo identificar las variables clave, sino también entender cómo interactúan secuencialmente para influir en los resultados. Esto es especialmente útil en contextos clínicos, donde la capacidad para visualizar caminos de decisión claros sobre las variables presentes en los pacientes puede mejorar el entender el comportamiento y llegar a tomar medidas contra el posible desarrollo de EMCD de los pacientes.

En comparación con el laboratorio anterior, donde se utilizó el algoritmo Apriori para generar reglas de asociación, ambos enfoques coinciden en la identificación de variables clave para la progresión a EMCD. Sin embargo, mientras que las reglas de asociación ofrecen combinaciones específicas de factores, el árbol de decisión proporciona una visión más estructurada y jerárquica de la importancia relativa de cada variable, lo que permite entender mejor cómo interactúan en conjunto para influir en el desarrollo de la enfermedad, facilitando así su interpretación.

Entre los aspectos positivos del enfoque utilizado, se destaca la capacidad del modelo para generar reglas precisas y aplicables, incluso en subconjuntos específicos de pacientes, lo que sugiere que el modelo está bien ajustado para identificar patrones relevantes en los datos. La representación gráfica del árbol de decisión también facilita la interpretación de los resultados, haciendo que la información sea más accesible para su aplicación clínica.

No obstante, el estudio presenta oportunidades de mejora. La inclusión de un conjunto de datos más amplio y diverso podría permitir evaluar la generalización del modelo en diferentes poblaciones y así proporcionar una visión global de los factores que influyen en la progresión a EMCD. Por otro lado, la validación del modelo en un conjunto de datos

externos sería fundamental para confirmar la robustez de los hallazgos y garantizar que el modelo sea aplicable en diferentes contextos clínicos.

Referencias

Barrios, J. (2024). La Matriz de Confusión y sus Métricas [Consultado el 11 de agosto de 2024]. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

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/

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

Contreras, M. F. (2024). Laboratorio 3 - Reglas de Asociación [Realizado el 9 de junio de 2024].

Datascientest. (2024). Cross-Validation: Definición e Importancia [Consultado el 11 de agosto de 2024]. https://datascientest.com/es/cross-validation-definicion-e-importancia

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

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.

IBM. (2024a). Árboles de Decisión [Consultado el 11 de agosto de 2024]. https://www.ibm.com/es-es/topics/decision-trees

IBM. (2024b). C5.0 Node en Modelado de Árboles de Decisión [Consultado el 11 de agosto de 2024]. https://www.ibm.com/docs/en/cloud-paks/cp-data/5.0.x?topic=modeling-c50-node

Kili Technology. (2024). Training, Validation, and Test Sets: How to Split Machine Learning Data [Consultado el 11 de agosto de 2024]. https://kili-technology.com/training-data/training-validation-and-test-sets-how-to-split-machine-learning-data

Michelli, J. (2024). The Importance of Accuracy in Our Daily Lives [Consultado el 11 de agosto de 2024]. https://www.michelli.com/the-importance-of-accuracy-in-our-daily-lives/

Numerentur. (2024). Métodos de Poda [Consultado el 11 de agosto de 2024]. https://numerentur.org/metodos-de-poda/

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.

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). *sciencedirect*, 916-953.e19. https://doi.org/https://doi.org/10.1016/B978-0-323-62539-5.00043-6