IDENTIFICACIÓN DE RIESGOS DEL ALZHEIMER: UN ANÁLISIS PROBABILÍSTICO DE DATOS CLÍNICOS Y MEDICIONES CLAVE

B. Itzelt Gómez Catzín[1], M. Fernanda Gamboa Martinez[1], Gabriela Lujan[1], Rebeca Koch[1] and María P. Rodríguez[1]

[1]Campus Guadalajara, Tecnológico de Monterrey

Abstract

Alzheimer's disease is a progressive disorder that gradually impairs memory and cognitive functions, making it the most common cause of dementia. This study presents a probabilistic approach using Gaussian Bayesian Networks to predict the development of Alzheimer's by analyzing key diagnostic variables. Data were categorized into four subsets: biomarkers, cognitive assessments, brain imaging, and demographic factors. Three directed acyclic graphs (DAGs) were constructed to model the relationships between these variables and estimate conditional probabilities related to Alzheimer's diagnosis. The results provide valuable insights into the significance of different variables and their interrelations, contributing to a more nuanced understanding of Alzheimer's progression.

Keywords: Alzheimer's disease, Gaussian Bayesian Networks, biomarkers, cognitive assessments, brain imaging, directed acyclic graphs (DAGs), probabilistic modeling.

1 INTRODUCCIÓN

La enfermedad de Alzheimer, es un trastorno que deteriora paulatinamente la memoria del ser humano, así como la capacidad de pensar. Este mal provoca que de forma gradual el cerebro se encoja, afectando así las capacidades de funcionamiento de una persona, siendo la causa más común de demencia.

En función de esto último, se realizó un análisis de información y recolección de datos, donde con el uso de herramientas de programación y estadística, se desarrolló una aplicación del método de redes Bayesianas Gaussianas (método lineal), para generar proyecciones probabilísticas sobre las distintas variables que presenta la enfermedad en su diagnóstico, y de esta forma, predecir el comportamiento y desarrollo de la enfermedad en las personas.

2 MÉTODOLOGÍA

2.1 Investigación Previa

Antes de reunirnos con los especialistas médicos, decidimos investigar sobre el Alzheimer y cómo las variables presentes en nuestra base de datos son relevantes para su detección y seguimiento. Nuestra

investigación nos permitió comprender mejor algunos conceptos que antes nos eran desconocidos y, a partir de ello, logramos dividir nuestras variables en cuatro subconjuntos clave: ***biomarcadores, evaluaciones cognitivas, imágenes cerebrales y factores demográficos.***

2.1.1 Biomarcadores

Los biomarcadores son indicadores biológicos medibles que reflejan procesos patológicos en el cuerpo, como la acumulación de proteínas específicas o alteraciones metabólicas en el cerebro, que son cruciales en el diagnóstico del Alzheimer.

- APOE4: El número de alelos APOE4. Este genotipo es uno de los factores de riesgo genéticos más fuertes asociados con el Alzheimer. Estos alelos heredados son aquellos que posiblemente acarren esta mutación de la enfermedad en la ABOE4 que codifica para esta proteína, o en dado caso que no se haya heredado por factores ambientales, se induzca la expresión de esto.
- ABETA, PTAU: Niveles de estas proteínas en el líquido cefalorraquídeo (LCR). La disminución de ABETA y la acumulación de la proteína TAU son marcadores clave en la progresión del Alzheimer.

2.1.2 Evaluaciones Cognitivas

Estas pruebas miden las funciones cognitivas y ayudan a evaluar el grado de deterioro mental, proporcionando información valiosa para el diagnóstico y seguimiento del Alzheimer.

- MMSE: Mini-Mental State Examination, una prueba comúnmente utilizada para evaluar el deterioro cognitivo global.
- ADAS11: Subescala del ADAS-Cog que evalúan la memoria y otras funciones cognitivas, útil para seguir la progresión de la enfermedad.
- CDRSB: Mide la severidad del deterioro cognitivo mediante la suma de puntuaciones de diferentes áreas cognitivas.
- RAVLT: Prueba de aprendizaje verbal que mide la memoria inmediata, el aprendizaje y el olvido. En este caso, se utilizó para el análisis del estudio los resultados del olvido

2.1.3 Imágenes Cerebrales

Las imágenes cerebrales permiten observar cambios estructurales en el cerebro que son indicativos del Alzheimer, como la atrofia en el hipocampo o en la corteza cerebral.

- Hippocampus: Representa el tamaño del hipocampo, una región crítica para la memoria que suele atrofiarse en etapas tempranas de la enfermedad.
- Wholebrain: Es la evaluación del volumen cerebral total y de la corteza, que puede reducirse a medida que avanza la enfermedad.

2.1.4 Factores Demográficos

Estos factores incluyen información básica sobre el paciente que puede influir en la prevalencia y progresión del Alzheimer.

- AGE: La edad es uno de los principales factores de riesgo para el Alzheimer, con la incidencia aumentando significativamente con la edad.
- PTEDUCAT: Es el nivel de educación en años, asociado con la reserva cognitiva, que puede retrasar la aparición de los síntomas.

Al comprender y categorizar estas variables, damos un primer paso crucial hacia el entendimiento de nuestros datos, la clasificación de nuestras variables y la identificación de sus relaciones en la detección de la enfermedad. Este proceso será fundamental para construir un modelo más preciso para la detección y el monitoreo del Alzheimer.

2.2 Consultas con Especialistas Médicos

Después de realizar esta investigación inicial, que nos permitió un primer acercamiento al Alzheimer, tuvimos una reunión con especialistas que nos ayudaron a comprender en mayor profundidad cómo nuestras variables podrían estar interrelacionadas y cuáles de ellas tienen un mayor impacto en la enfermedad. Gracias a su orientación, llegamos a las siguientes conclusiones sobre nuestros subconjuntos de datos. A continuación, presentamos los subconjuntos ordenados por relevancia, desde el más importante hasta el menos relevante:

2.2.1 Evaluaciones cognitivas

Nuestros especialistas médicos nos mencionaron que algunos de los síntomas más característicos del Alzheimer incluyen la pérdida de memoria reciente, problemas con el lenguaje, desorientación en el tiempo y el espacio, dificultad para reconocer a familiares y amigos, e incluso la dependencia completa. Debido a que estos síntomas están estrechamente relacionados con el deterioro cognitivo, las evaluaciones cognitivas son fundamentales en el diagnóstico del Alzheimer.

Las evaluaciones cognitivas consisten en una serie de pruebas diseñadas para medir el funcionamiento de diferentes áreas del cerebro, con el objetivo de detectar déficits clínicos característicos de la enfermedad. Entre las pruebas estandarizadas que utilizamos, se encuentra el Mini-Mental State Examination (MMSE), una prueba de suma importancia para nuestro análisis, ya que es una de las herramientas más conocidas y utilizadas para evaluar la memoria a corto plazo, la orientación, la atención, el lenguaje y las habilidades visuoespaciales. Estos aspectos son cruciales para el diagnóstico del Alzheimer.

Es importante tener en cuenta que, en varias ocasiones, el Alzheimer ha sido confundido con otras formas de demencia debido a que ambas condiciones generan déficits en los dominios cognitivos. Sin embargo, mientras exista una alteración significativa y discapacitante en estos dominios, la enfermedad puede ser considerada Alzheimer. Como mencionó uno de nuestros especialistas: “El Alzheimer es lentamente progresivo, y sus síntomas incluyen dificultades para recordar, comprender y comunicarse. Pero, si me preguntan, el Alzheimer es completamente clínico; no necesita imágenes

ni biomarcadores. Con simples pruebas cognitivas podemos detectar la existencia de la enfermedad".

Por todo esto, consideramos que las evaluaciones cognitivas en nuestra base de datos tienen una gran relevancia para nuestro análisis, ya que están estrechamente relacionadas con el diagnóstico del Alzheimer.

2.2.2 Factores Demográficos

En segundo nivel de importancia, consideramos incluir el subconjunto de factores demográficos, los cuales describen las características básicas de la población en estudio. En este caso, hemos decidido tener en cuenta las variables, edad y género por las siguientes razones:

Al dialogar con nuestros especialistas médicos, nos señalaron que la edad es un factor que se relaciona estrechamente con el Alzheimer, y que las mujeres tienen una mayor predisposición a desarrollar la enfermedad. Para corroborar esta afirmación, investigamos en la plataforma oficial del National Institute on Aging, donde verificamos que "La edad es el factor de riesgo más conocido para la enfermedad de Alzheimer. La mayoría de las personas con la enfermedad la desarrollan a los 65 años o después. Solo un 10% de los casos ocurren antes de esta edad. El riesgo de padecer Alzheimer aumenta significativamente después de los 65 años."

Por lo tanto, consideramos que la edad está fuertemente relacionada con el diagnóstico de Alzheimer, y es fundamental darle prioridad en nuestro análisis.

2.2.3 Biomarcadores

En tercer nivel de importancia, consideramos los biomarcadores, que son indicadores biológicos medibles en el cuerpo y pueden proporcionar información sobre la presencia o progresión de la enfermedad. Existen ciertas proteínas clave, como la beta-amiloide ($\alpha\ \beta$) y p-Tau. Niveles anormales de estas proteínas en el líquido cefalorraquídeo (LCR), como la disminución de $\alpha\ \beta$ y la acumulación de p-Tau, son características patológicas del Alzheimer. Sin embargo, la presencia de anomalías en estas proteínas no está directamente relacionada con el desarrollo de la enfermedad; es posible tener estas alteraciones sin manifestar síntomas clínicos de Alzheimer.

Aunque los biomarcadores son fundamentales para el monitoreo de la enfermedad, el ajuste de tratamientos, y la investigación y desarrollo de nuevas terapias, no son esenciales ni suficientes para el diagnóstico de la enfermedad. El diagnóstico inicial de Alzheimer se confirma principalmente a través de pruebas cognitivas. Una vez que estas pruebas indican la posible presencia de la enfermedad, los biomarcadores pueden usarse como un paso confirmatorio y para guiar el tratamiento adecuado. Por esta razón, no relacionamos directamente los biomarcadores con el diagnóstico inicial de Alzheimer.

2.2.4 Imágenes Cerebrales

Por último, incluimos las imágenes cerebrales, que permiten observar cambios importantes en la estructura del cerebro y pueden ser indicativos de Alzheimer. Entre estos cambios, se incluyen la reducción del volumen cerebral total y la atrofia de los ventrículos y el hipocampo. Aunque estos cambios estructurales están asociados con el Alzheimer y pueden señalar un proceso patológico en

curso, no siempre implican que la persona desarrollará la enfermedad.

Ver alteraciones en la estructura cerebral no está directamente relacionado con el diagnóstico de Alzheimer ni garantiza el desarrollo de la enfermedad. Estos hallazgos deben ser interpretados en el contexto de una evaluación clínica completa. Por lo tanto, concluimos que este subconjunto de datos tiene una relevancia menor para el diagnóstico de Alzheimer, y por eso lo ubicamos en el último lugar de nuestra lista.

Este orden refleja la opinión de los especialistas sobre la importancia relativa de cada subconjunto de datos en el estudio del Alzheimer, y está respaldado por nuestras conclusiones y la investigación realizada sobre la enfermedad.

2.3 Querying

En este proyecto se busca dar respuesta a preguntas relacionadas al diagnóstico del Alzheimer. El proceso de responder preguntas se conoce como *querying*, y por lo tanto, a las preguntas planteadas las llamaremos *queries*. Estos queries generalmente preguntan acerca de la probabilidad de que suceda un evento, dada la evidencia de otro.

2.4 Construcción de Variables

Para dar respuesta a las queries planteadas, se comienza estableciendo las variables asociadas. Las variables presentes en la base de datos son contínuas lo que significa que contienen datos numéricos que no son estrictamente enteros.

2.5 Creación de DAGs (Gráficos Acíclicos Dirigidos)

Las relaciones entre las variables establecidas pueden ser explicadas mediante grafos acíclicos dirigidos, también conocidos como DAG's. Los DAG's están compuestas por nodos, donde cada nodo del grafo representa una variable, y arcos entre los nodos, los cuales indican una relación de dependencia condicional entre las variables correspondientes. Es decir, si existe un arco de un nodo A a un nodo E, se interpreta de la siguiente forma: E depende de A, dado el resto de las variables de las cuales depende el nodo E.

Por ejemplo, decimos que el nivel de proteína TAU depende directamente de la edad si $A \rightarrow PTAU$. Donde el nodo independiente en la relación, en este caso el nodo A, se conoce como *nodo padre*, y el nodo dependiente en la relación, en este caso el nodo PTAU, se conoce como *nodo hijo*.

De esta forma, con base en la investigación y entrevistas realizadas, se propusieron tres grafos acíclicos dirigidos con distintas relaciones de dependencia entre las variables de interés, los cuales podrán estimar una respuesta a las queries planteados.

2.6 Red Bayesiana

Una red bayesiana es un modelo probabilístico que representa un conjunto de variables y sus dependencias condicionadas mediante un grafo acíclico dirigido. Lo que hace especial a una red bayesiana es su capacidad para manejar incertidumbre y hacer inferencias probabilísticas, permitiendo calcular la probabilidad de ciertos eventos dados otros eventos observados. Esto es útil en situaciones donde

necesitamos modelar relaciones complejas entre variables en un entorno de incertidumbre. Ajustar una DAG a una red bayesiana implica asignar una función de distribución a los nodos en la DAG.

2.7 Red Bayesiana Gausiana

Una red bayesiana gaussiana es un tipo de modelo probabilístico que se utiliza cuando todas las variables de interés son continuas y siguen una distribución normal. En este modelo, cada variable (o nodo) puede estar influenciada por otras variables a través de relaciones lineales.

Si una variable no tiene ”padres” (es decir, no depende de ninguna otra variable en la red), se asume que sigue una distribución normal simple. Sin embargo, si una variable tiene ”padres”, su valor se modela como una combinación lineal de esas variables ”padre”, más un término de error que también sigue una distribución normal. Este término de error tiene una varianza específica para cada variable que no cambia, sin importar las otras variables que la influyan. Este enfoque permite modelar de manera efectiva las relaciones entre variables continuas, como lo es este caso para distintos parámetros determinantes en la diagnosticación de Alzheimer.

2.8 Estimación de Parámetros

Las funciones de distribución en una red bayesiana gaussiana cuentan con parámetros asociados, que son números fijos pero desconocidos y representan características de la población. En este caso, los parámetros incluyen los coeficientes que definen las relaciones lineales entre los nodos ”padre” y ”hijo” y las varianzas del término de error de cada nodo. Dado que estos parámetros son desconocidos, es necesario estimarlos para poder calcular las probabilidades y las relaciones entre las variables.

Este proceso implica que, dado un DAG propuesto y un conjunto de datos, se estima tanto la media condicional como la varianza de cada nodo utilizando el método de máxima verosimilitud (MLE). Este método busca los valores de los parámetros que maximizan la probabilidad de observar los datos dados, permitiendo obtener una red bayesiana que refleje de manera óptima las relaciones lineales y la distribución normal de los datos observados.

2.9 Medidas de Bondad de Ajuste

En las redes bayesianas en general, las medidas de bondad de ajuste, como el BIC (Criterio de Información de Bayes) y el AIC (Criterio de Información de Akaike), se utilizan para evaluar qué tan bien una red propuesta explica los datos observados. Estas medidas ayudan a comparar diferentes estructuras de redes y seleccionar la que mejor se ajuste a los datos, teniendo en cuenta tanto la precisión del ajuste como la complejidad del modelo.

En las redes bayesianas gausianas, el **AIC** penaliza la falta de ajuste de la red a los datos, pero también castiga la complejidad de la red. Un AIC más bajo indica un mejor equilibrio entre el ajuste y la simplicidad del modelo. El **BIC**, por otro lado, es similar al AIC, pero penaliza la complejidad del modelo de manera más fuerte, favoreciendo redes más simples cuando hay un gran número de datos. Un BIC más bajo sugiere que la red tiene un buen ajuste a los datos sin ser innecesariamente compleja.

Ambos criterios son fundamentales para evitar el sobre ajuste, es decir, para garantizar que la red bayesiana no se ajuste demasiado a los datos de la muestra a costa de ser menos generalizable a nuevos datos.

3 APLICACIÓN

El objetivo principal de este estudio, es dar respuesta a 3 queries que buscan resolver ciertas incógnitas del área de neurología, para ser más concretas, de la enfermedad del Alzheimer a través de probabilidades condicionales:

1. ¿Qué tan probable es que una persona desarrolle Alzheimer si su nivel educativo es menor o igual a 12 años?
2. ¿Cuál es la probabilidad de que un paciente en edad temprana de detección desarrolle Alzheimer dada la presencia de una anomalía en el biomarcador APOE4 o que presente reducción en la corteza y volumen cerebral (WB)?
3. ¿Cuál es la probabilidad de que una persona muestre atrofia en el hipocampo dado que muestra niveles altos de proteína Tau?

3.1 Comprensión de los Datos

Antes de plantear las consultas y construir las redes bayesianas para este estudio, se realizó una ardúa investigación y una serie de entrevistas con especialistas y estudiantes del área médica. Dada la complejidad de los datos y las limitaciones del diccionario de términos, su conocimiento fue fundamental para el desarrollo y la comprensión de nuestras variables. Esta colaboración, junto con la exploración detallada de los datos, nos permitió establecer las relaciones entre cada una de las variables.

3.2 Exploración de la Base de Datos

Al comenzar con el análisis de la base de datos proporcionada por ADNI, nos encontramos con varios obstáculos. Primeramente, se descubrió una gran cantidad de columnas con bastantes datos faltantes. Esto resultó en información incompleta sobre estudios médicos importantes. Encontramos 63 variables de diferentes estudios y mediciones médicas a pacientes con posibles diagnósticos de esta enfermedad. Al realizar las entrevistas e investigaciones con los diferentes especialistas y fuentes, se logró reducir a las once variables que se describieron en 2.1. Igualmente, es importante destacar, que dado que el ajuste de los datos fue con una red Bayesiana Gaussiana, estos datos deben ser numéricos. Ese fue otro criterio a considerar para la filtración de los datos, así como la exclusión de variables que pudieron haber aportado gran valor, como lo son el sexo y el diagnóstico. Más adelante, se comentará cómo se podrían incorporar este tipo de datos, combinando datos categóricos y numéricos en la creación del modelo.

3.3 Manejo de los Valores Nulos

Anteriormente, se mencionó que se contaba con una gran cantidad de valores nulos. Hubo variables importantes para el estudio, de las cuales solo el 20% de los datos representaban valores no nulos. En otros casos, hubo variables donde casi el 50% de los datos resultaron ser no nulos. Ante esta dificultad, nos enfrentamos con dos opciones: la filtración de datos faltantes, o imputación con inferencia

posterior de Monte Carlo.

La primera opción consiste en eliminar todas las filas del conjunto de datos, cuyas variables contienen valores faltantes. Esta opción es sencilla de implementar y garantiza que solo se utilicen observaciones completas en el análisis. No obstante, al implementar este método, pasamos de contar con 16,421 registros, a contar con aproximadamente 1700 registros. Este método conllevó a una pérdida significativa de datos, sin embargo, se realizó un análisis entre los datos completos y sin imputar con respecto a la variable de diagnóstico, y a través de dos gráficos de barras (Figuras 1 y 2) encontramos que las distribuciones son bastante parecidas siguiendo una misma proporción, por lo tanto, considerando que los diagnósticos realizados por los expertos de la salud, hayan sido en su totalidad acertados, podemos reducir nuestros datos sin riesgo de sesgo en los modelos.

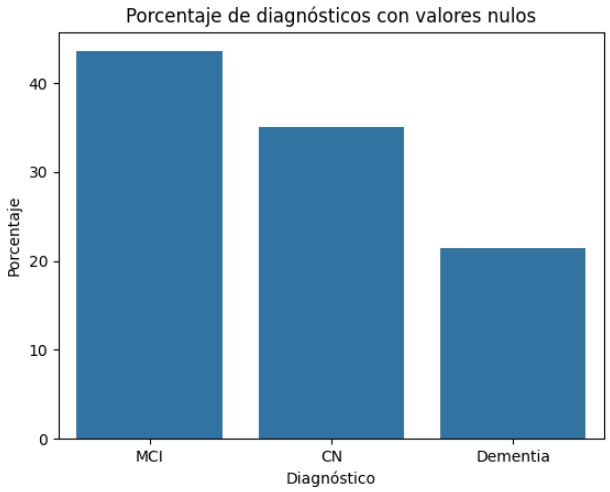

Figure 1. Diagnóstico con valores nulos

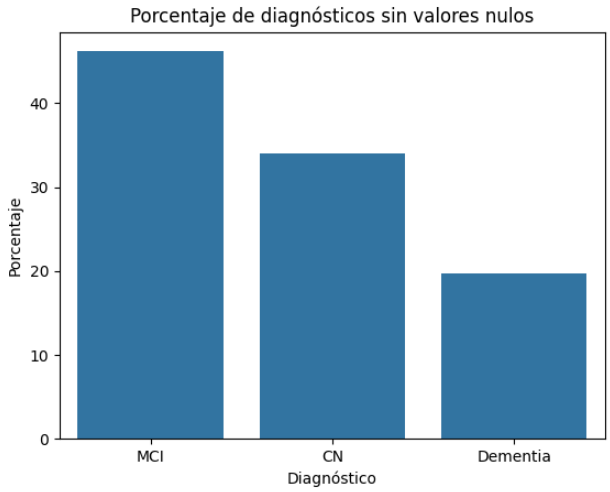

Figure 2. Diagnóstico sin valores nulos

La segunda opción, consiste en utilizar un modelo bayesiano para imputar los valores faltantes en cada observación a partir de su distribución posterior conjunta, condicionada a las variables que sí están observadas. Esta distribución se estima de manera empírica utilizando el método likelihood weighting, y estos valores imputados son la media o moda de esa distribución, por lo tanto, los valores imputados pueden variar entre diferentes ejecuciones en la función. Este método permite utilizar todo el conjunto de datos, lo que mejoraría la precisión y potencial del análisis.

Con base en lo anterior, la opción más atractiva sería la segunda, sin embargo, al momento de realizar las medidas de bondad de ajuste, los grafos realizados con estos datos presentaron scores abrumadoramente bajos (de los cuales hablaremos más adelante en la sección 3.6). Comparándolo con los resultados encontrados con el otro método, resultó de mayor agrado elegir la primera opción para el manejo de nuestros datos faltantes.

3.4 Creación de DAGs (Gráficos Acíclicos Dirigidos)

Con base en las variables definidas en 2.1 creamos tres diferentes DAGs que darán respuesta a nuestras tres incógnitas a responder

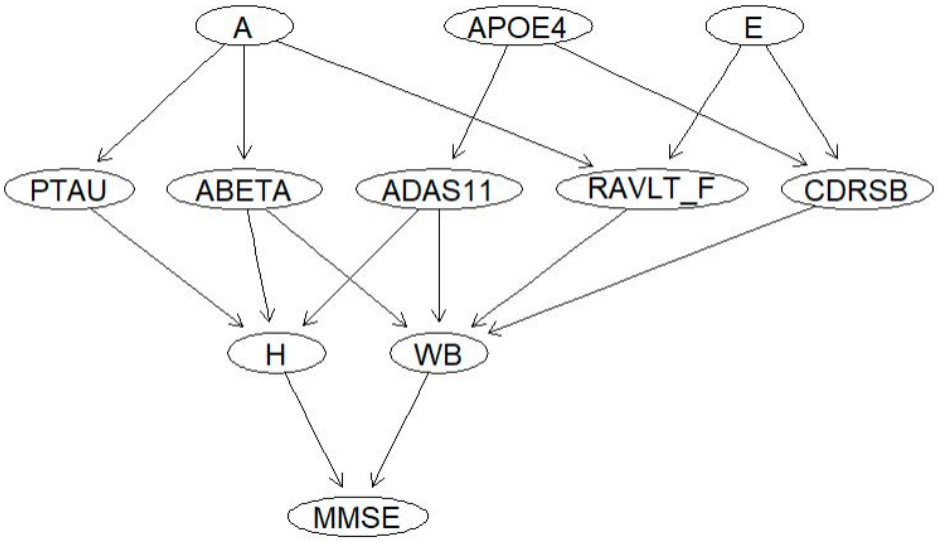

Figure 3. Primera DAG

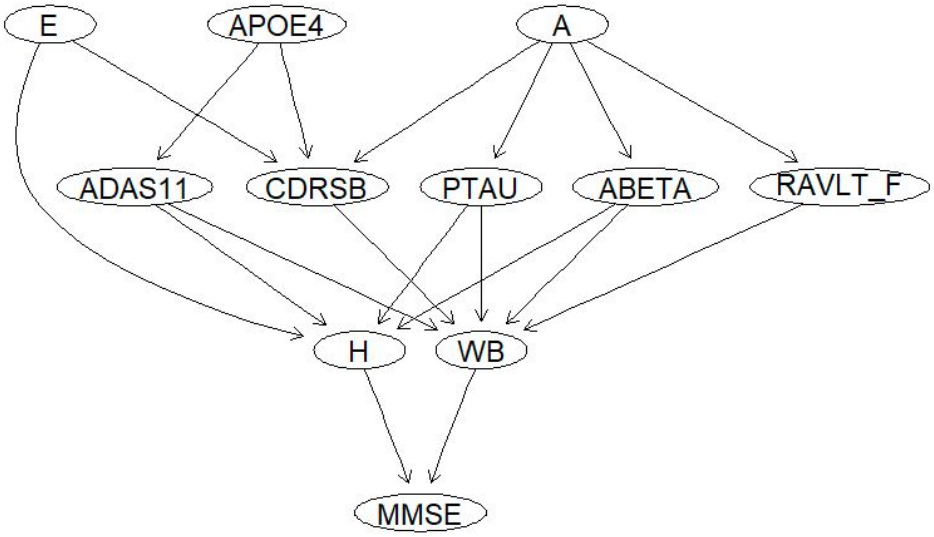

Figure 4. Segunda DAG

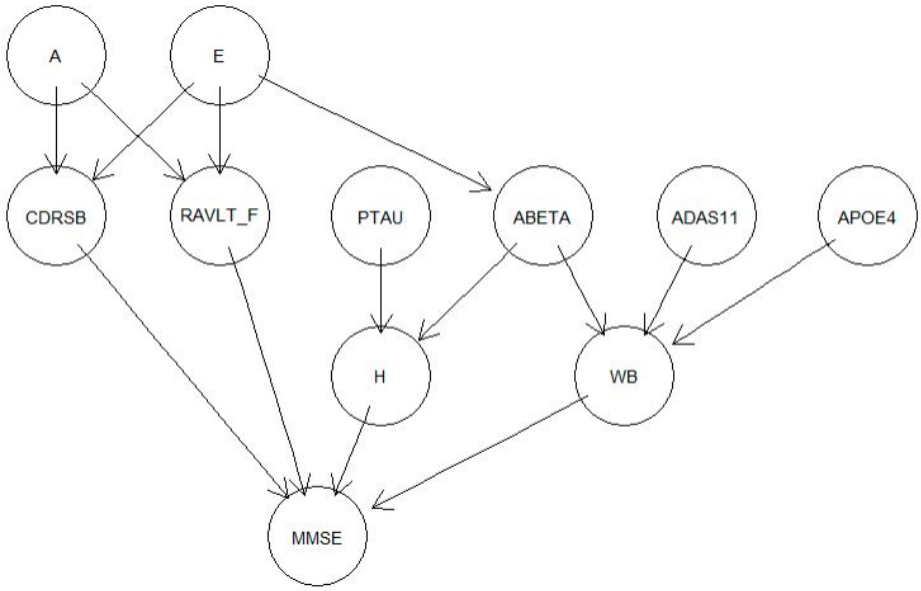

Figure 5. Tercera DAG

3.5 Ajuste de la Red Bayesiana Gausiana para las DAGs propuestas

3.6 Medidas de Bondad de Ajuste

DAGs	BIC-g Score
1	-78349.57
2	-78340.58
3	-78138.3

Table 1. Resultados BIC

DAGs	AKAIKE Score
1	-78247.84
2	-78233.49
3	-78041.92

Table 2. Resultados AKAIKE

Podemos observar de las dos pruebas realizadas, obtuvimos un mejor resultado en ambos casos para la tercera red bayesiana. En particular, consideramos que la primera tuvo una mejor implementación de la teoría al momento de establecer los arcos entre nodos, además de haberse realizado junto a una experta del área.

3.7 Inclusión de Variables Categóricas

Sabemos que la suposición fundamental de las redes bayesianas gaussianas es que todas las variables incorporadas (tanto padres como hijos) siguen una distribución normal. Por lo tanto, al intentar incorporar variables categóricas como el sexo y el diagnóstico en una GBN, que no siguen dicha distribución, modelarlas como tales introduce un error fundamental en nuestro modelo, ya que las relaciones de probabilidad entre las variables no se representarían correctamente.

Para afrontar este problema, existen alternativas como codificar las variables categóricas en valores numéricos. Sin embargo, esto puede introducir relaciones erróneas en nuestra red y no siempre es apropiado. Este método, en este caso, realmente no nos sirve porque, aunque se transformen en valores de 0 y 1, las variables continúan comportándose como categóricas y no nos proporcionan la

información necesaria.

Otra opción, para poder trabajar, por ejemplo, con el sexo, podría ser crear una variable latente continua que represente el efecto del sexo en las variables observables. Esta distribución latente interactúa con las demás variables del modelo. Por ejemplo, la probabilidad de desarrollar la enfermedad (P(Alzheimer)) podría depender de la variable latente sexo, que a su vez influiría en otras variables continuas dentro de la red.

El modelo aprende las probabilidades condicionales a partir de los datos. Utilizas técnicas bayesianas para actualizar las distribuciones a priori con la evidencia proporcionada por los datos observados, obteniendo así las distribuciones a posteriori. A cada categoría de la variable sexo (como hombre y mujer) se le asigna una distribución a priori. Esto influiría en el modelo, ya que si P(Alzheimer) depende del sexo, entonces, con una probabilidad a priori mayor de que la observación sea de una mujer, podría existir una mayor probabilidad de que una mujer desarrolle Alzheimer en comparación con un hombre, según cómo interactúan las demás variables en la red.

3.8 Inclusión del sexo de las personas

Al intentar implementar este método en nuestro estudio, nos dimos cuenta de que resultaba mucho más complejo de lo que habíamos previsto inicialmente. El método consiste en incorporar variables categóricas, como el sexo, en nuestro DAG (Directed Acyclic Graph) mediante la introducción de variables latentes. Esta técnica permite representar el efecto de las variables categóricas en el modelo de manera indirecta, mediante la introducción de una variable continua latente que capta su influencia.

El proceso de integrar variables categóricas en una red bayesiana gaussiana (GBN) implica definir adecuadamente cómo estas variables afectan a otras variables observables en el modelo. En nuestro caso, esto significó ajustar el DAG para incluir la variable latente ”Sexo” y configurar su impacto en variables continuas como el volumen del hipocampo y el volumen cerebral. Este enfoque permite capturar efectos que no se observan directamente, pero que son fundamentales para una comprensión completa del fenómeno estudiado.

Sin embargo, hemos aprendido que, aunque esta herramienta es extremadamente poderosa y puede ofrecer insights valiosos, su aplicación requiere una comprensión profunda. Debemos asegurarnos de ajustar adecuadamente los parámetros del modelo y definir correctamente las relaciones entre las variables latentes y observables. Esto no solo implica un conocimiento detallado del método, sino también una validación rigurosa de los resultados obtenidos para garantizar su precisión y relevancia.

Ahora, entendemos que este método es una herramienta poderosa que, si se aplica correctamente, puede proporcionar valiosos insights en nuestros estudios. Sin embargo, es fundamental comprender bien cómo funciona y cuándo es adecuado usarlo. La complejidad de la implementación y la necesidad de adaptar los datos adecuadamente subraya la importancia de una preparación meticulosa y un conocimiento profundo del modelo para aplicarlo de manera efectiva en el análisis de datos.

4 RESULTADOS

4.1 Query 1

$$P(\text{MMSE} \leq 15 \mid E \leq 12) = 0.0002113746 \tag{1}$$

El resultado obtenido por el algoritmo nos arroja una probabilidad de 0.0002113746. Esta probabilidad es extremadamente baja, lo que sugiere que el nivel educativo no es un factor determinante en la detección del Alzheimer. Sin embargo, al modificar la consulta y considerar un nivel educativo superior a 12 años, la probabilidad disminuye aún más a 0.000126029. Esto sugiere que, a mayor nivel educativo y mayor actividad mental, es menos probable que una persona desarrolle la enfermedad de Alzheimer.

4.2 Query 2

$$P(\text{MMSE} <= 20 \cup 54 \leq A \leq 67 \mid \text{APOE4} = 1 \cup \text{WB} < 947663) = 0.002961571 \tag{2}$$

La probabilidad resultante de 0.002927978, aunque mayor que en el caso del nivel educativo, sigue siendo baja. Esto sugiere que, aunque la presencia del biomarcador APOE4 o la reducción en la corteza y volumen cerebral son factores de riesgo, por sí solos no son determinantes absolutos para el desarrollo de la enfermedad en el modelo considerado.

Las edades que estaban presente en los datos rondaban entre los 54 y 85 años, consideramos elegir el rango de edad de 54 y 67 años porque es una edad en la que normalmente la enfermedad comienza a presentar síntomas. Con respecto al gen APOE4, si una persona ya presenta una mutación en el alelo, ya tiene indicios de la enfermedad.

4.3 Query 3

$$P(H \leq 5956 \mid \text{PTAU} \geq 60) = 0.03360433 \tag{3}$$

El resultado de probabilidad obtenido, 0.03341027, es el más alto entre las tres consultas realizadas. Sin embargo, sigue siendo relativamente bajo, lo que indica que los niveles altos de la proteína Tau no tienen una relación sumamente significativa con la atrofia en los hipocampos. Esto sugiere que no todos los individuos con niveles elevados de Tau desarrollarán atrofia en esta región cerebral.

Elegimos estos valores, porque al realizar las entrevistas médicas, se nos informó que entre mayores valores de proteína Tau, se tiene una mayor acumulación. Para el valor del hipocampo, entre mayor atrofia presente (es decir, entre mayor reducción de volumen sufra) tendrá un menor valor registrado, al realizar las estadísticas de esta variable, encontramos que los valores en el cuartil 25 eran menor o iguales a 5956, por lo que consideramos que era un valor adecuado para realizar nuestra consulta.

5 CONCLUSIONES

En el presente proyecto, nuestro objetivo fue analizar las relaciones entre la enfermedad de Alzheimer y diversos factores como biomarcadores, imágenes cerebrales, factores demográficos y pruebas cognitivas. El propósito era identificar cuál de estas variables podría tener la relación más significativa con la enfermedad, con el fin de orientar mejor los esfuerzos hacia una detección temprana y un

seguimiento adecuado del Alzheimer.

Tras una extensa investigación, tanto interna como en consulta con especialistas médicos, obtuvimos una mejor comprensión de los factores que se consideran más influyentes en el desarrollo del Alzheimer. Las opiniones de los especialistas variaron, y aunque cada uno destacó diferentes aspectos, al final acordamos un orden de prioridad para las variables que creemos que influyen en la enfermedad y formulamos nuestras consultas en función de ello.

Sin embargo, al obtener las respuestas de nuestro modelo en forma de probabilidades, nos encontramos con valores extremadamente bajos. Estos resultados nos indican que ni las pruebas cognitivas, ni los biomarcadores, ni las imágenes cerebrales por sí solos son suficientes para garantizar un diagnóstico temprano del Alzheimer. Aquí es donde las palabras de la Dra. Fernanda cobran relevancia: ”Es muy complicado darle un orden de prioridad a esos cuatro factores que mencionas, porque todos están relacionados. El diagnóstico del Alzheimer es un proceso multifacético y se basa en un conjunto de herramientas, no en elegir la mejor”.

Teniendo en cuenta su perspectiva, concluimos que el desarrollo del Alzheimer es un proceso altamente complejo y multifactorial. Ninguna variable única parece ser un predictor fuerte de la enfermedad, lo que refleja la interacción de múltiples factores, incluidos los genéticos, ambientales y de estilo de vida.

Aunque observamos algunos patrones, como que un mayor nivel educativo podría estar asociado con una menor probabilidad de desarrollar Alzheimer, las probabilidades obtenidas fueron en general muy bajas. Esto refuerza la necesidad de un enfoque integral y multidimensional para la detección y el manejo del Alzheimer, combinando pruebas cognitivas, análisis de biomarcadores, estudios de neuroimagen, y la consideración de factores socio demográficos y de estilo de vida.

Finalmente, estos resultados sugieren que aún hay un amplio margen de mejora en la investigación del Alzheimer y sus posibles relaciones con diversas variables. Es importante ser cautelosos al interpretar estos resultados, ya que las probabilidades bajas no necesariamente implican que los factores estudiados no tengan relevancia, sino que pueden estar interactuando con otros elementos que no fueron considerados en este análisis.

6 AGRADECIMIENTOS

Agradecemos profundamente la valiosa ayuda y aportación de los doctores Alejandro Zarraga y Pablo Saldaña, así como a la doctora Fernanda Konrad, cuya comprensión de los datos y orientación en la selección de variables, así como en el establecimiento de las consultas, nos fue de gran guía para la creación de las redes bayesianas. Asimismo, extendemos nuestro agradecimiento a la doctora Itzel Quirino Chávez, por guiarnos en la construcción de los arcos entre los nodos, la determinación de los valores correspondientes para los cálculos de probabilidades, e interpretación de cada variable seleccionada. Sin duda, sus aportaciones y conocimientos fueron fundamentales para el éxito de este trabajo.

A MATERIAL SUPLEMENTARIO

El código completo utilizado para el desarrollo del proyecto se encuentra en este repositorio de Drive: **Repositorio.**

REFERENCES

National Institue on Aging (2024) *¿Qué causa la enfermedad de Alzheimer?* Recuperado de: https://www.nia.nih.gov/espanol/enfermedad-alzheimer/causa-enfermedad-alzheimer#:~:text=La%20edad%20es%20el%20factor,despu%C3%A9s%20de%20los%2065%20a%C3%B1os

Manuel H. Janeiro, Carlos G. Ardanaz, Noemí Sola-Sevilla, Jinya Dong, María Cortés-Erice, Maite Solas, Elena Puerta, and María J. Ramírez. (2021) *Biomarcadores en la enfermedad de Alzheimer. Recuperado de: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10197768/*

Bombón-Albán P., Campoverde-Pineda E., Medina-Carrillo M. (2022) Revisión de las pruebas cognitivas breves para pacientes con sospecha de demencia. *Recuperado de: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-87482022000300098#:~:text=Las%20pruebas%20cognitivas%20breves%20com%C3%BAnmente,la%20detecci%C3%B3n%20de%20la%20demencia*

Fundación Pasqual Maragall (2020) ¿A qué áreas del cerebro afecta la enfermedad de Alzheimer?. *Recuperado de: https://blog.fpmaragall.org/areas-del-cerebro#:~:text=El%20hipocampo%2C%20pues%2C%20es%20una,para%20el%20proceso%20de%20aprendizaje*

Wikipedia contributors. (2024, February 23). Latent variable model. Wikipedia. *Recuperado de: https://en.wikipedia.org/wiki/Latent_variable_model*

Scutari, M. (2024) Imputing missing values from a Bayesian network. *Recuperado de: https://www.bnlearn.com/examples/impute/*

ChatGPT. (31 de agosto de 2024). Redacción y discusión sobre la importancia de probabilidades en la detección y seguimiento del Alzheimer [Chat]. OpenAI. https://chat.openai.com/