EXPLICACIONES ROBUSTAS MEDIANTE FUSIÓN DE EVIDENCIA DEMPSTER–SHAFER

MARCELO FUENZALIDA MAYNARD

Tesis para optar al título de Ingeniero Civil en Informática y Telecomunicaciones y Magister en Ciencias de la Ingeniería

Profesor guía: Jonathan Frez Zachary

FACULTAD DE INGENIERÍA
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES

Santiago, Chile
2025

EXPLICACIONES ROBUSTAS MEDIANTE FUSIÓN DE EVIDENCIA DEMPSTER–SHAFER

MARCELO FUENZALIDA MAYNARD

Tesis para optar al título de Ingeniero Civil en Informática y Telecomunicaciones y Magister en Ciencias de la Ingeniería

Profesor guía: Jonathan Frez Zachary

FACULTAD DE INGENIERÍA
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES

Santiago, Chile
2025

Marcelo Fuenzalida Maynard
✉ marcelo.fuenzalida@mail.udp.cl

Contenido

Capítulo 1

Introducción

1.1. Motivación

En las redes neuronales surge de su capacidad para abordar problemas complejos mediante el aprendizaje automatizado, ofreciendo soluciones eficientes en áreas como la clasificación, el reconocimiento de patrones y la toma de decisiones. Sin embargo, en escenarios donde los datos son inciertos, incompletos o contradictorios, estas técnicas enfrentan desafíos importantes. Este tipo de situaciones es común en aplicaciones críticas como la detección de anomalías, la predicción en tiempo real y los sistemas de apoyo a la decisión.

La teoría de Dempster-Shafer, es conocida por su enfoque en la representación y combinación de evidencias bajo la incertidumbre de datos, ofrece un marco sólido para mejorar la capacidad de las redes neuronales al manejar esta información incierta de manera explícita y coherente. Este enfoque permite integrar múltiples fuentes de datos con diferentes grados de confiabilidad, lo que aumenta la robustez de los modelos predictivos y mejora la calidad de las decisiones automatizadas.

El motivo principal de esta tesis radica en la necesidad de explorar la sinergia entre las redes neuronales y la teoría de Dempster-Shafer, dado su potencial para abordar problemas complejos en dominios como la inteligencia artificial explicable, los sistemas de diagnóstico médico y la ciberseguridad. Además, la motivación personal se centra en contribuir al avance del conocimiento en la intersección de estas dos áreas, proporcionando un marco

que pueda ser aplicado en problemas prácticos donde la incertidumbre es un factor determinante.

Esperando que los resultados de esta investigación no solo refuercen la capacidad de las redes neuronales para gestionar datos inciertos, sino que también sirvan como base para futuras innovaciones en inteligencia artificial confiable y sistemas de toma de decisiones robustos. Este trabajo representa un paso hacia la creación de modelos más flexibles, interpretables y útiles en situaciones reales.

1.2. Contexto

1.2.1. Definiciones

En el ámbito de la teoría de evidencia, inteligencia artificial y aprendizaje automático, se presentan diversos conceptos fundamentales que sustentan el desarrollo y la aplicación de estas disciplinas. Algunos de estos conceptos iniciales provienen del libro *A Mathematical Theory of Evidence* [1].

Uno de estos conceptos es la **Función de Masa Básica (BPA)** (*Basic Probability Assignment, BPA*), que se define como una función $m : 2^{\Theta} \to [0, 1]$, donde 2^{Θ} representa el conjunto de partes del marco de discernimiento Θ. Esta función cumple las propiedades $m(\emptyset) = 0$ y $\sum_{A \subseteq \Theta} m(A) = 1$. El valor $m(A)$ representa la fracción de evidencia que respalda exactamente al subconjunto $A \subseteq \Theta$, sin apoyar subconjuntos más pequeños ni A mismo.

Otro concepto importante es la **Función de Creencia (Bel)**, que acumula la evidencia asociada a un conjunto A y todos sus subconjuntos. Matemáticamente, se define como:

$$Bel(A) = \sum_{B \subseteq A} m(B),$$

donde $A \subseteq \Theta$. El valor $Bel(A)$ mide el grado de confianza en que la hipótesis verdadera pertenece al conjunto A.

En contraste, la **Función de Plausibilidad (Pl)** evalúa hasta qué punto la evidencia no contradice un conjunto A. Su definición es:

$$Pl(A) = 1 - Bel(\neg A),$$

donde $\neg A = \Theta \setminus A$ representa el complemento de A en el marco de discernimiento.

Además, se introduce la **Regla de Combinación de Dempster**, que permite combinar dos funciones de masa independientes, denotadas como m_1 y m_2, para obtener una nueva función de masa combinada m. Esta se expresa como:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K},$$

donde K representa el grado de conflicto entre las evidencias:

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C).$$

El valor $m(A)$ resultante describe la evidencia combinada asignada al subconjunto A.

En el contexto del aprendizaje automático, destacan las **Redes Neuronales Artificiales (ANN)** (*Artificial Neural Networks, ANN*), que son modelos inspirados en las redes neuronales biológicas. Estas redes están formadas por nodos interconectados, conocidos como neuronas artificiales, que emplean un enfoque de conexiones para procesar información. Las ANN son capaces de aprender tanto relaciones lineales como no lineales directamente de los datos. Uno de los modelos más utilizados es el perceptrón multicapa (MLP), que se basa en un esquema supervisado y requiere un conjunto de datos históricos para aprender. Otros modelos incluyen los mapas autoorganizados (*SOM*) para análisis de patrones y las redes recurrentes (*RNN*) para manejar datos secuenciales. [2]

Por otro lado, el **Random Forest (RF)** es una técnica de aprendizaje supervisado que se utiliza en tareas de clasificación y regresión. Este modelo se basa en la construcción de múltiples árboles de decisión, cuya combinación mejora la precisión y robustez. Durante su construcción, se emplean muestras *bootstrap* para entrenar cada árbol, lo que introduce diversidad al modelo. Además, en cada división de los árboles, se selecciona aleatoriamente un subconjunto de características, reduciendo así el riesgo de sobreajuste. En tareas de clasificación, el resultado final se obtiene mediante el voto mayoritario de los árboles, mientras que en regresión se calcula el promedio de sus predicciones. [3]

La **Regresión** se describe como el proceso de encontrar una expresión simbólica que relacione variables de entrada y salida basándose en datos observados [4]. Este enfoque permite identificar patrones significativos que expliquen las dependencias entre las variables de un sistema.

1.2.2. Regla de Dempster

La teoría de Dempster también conocida como "probabilidades superiores e inferiores inducidas por un trazado multivaluado" [5]. Considerando un par de espacios X y S unidos con un trazado multivaluado en Γ el cual se asgina un subconjunto $\Gamma x \subset S$ para todo $x \in X$. Suponiendo que μ es una probabilidad medida que asigna probabilidades a los miembros de una clase $\mathfrak{F}$ con subconjuntos de X.
Si μ es aceptable para juicios de probabilidad sobre un resultado incierto $x \in X$, y para este resultado incierto x, se sabe que corresponde a un resultado incierto $s \in \Gamma x$.
Para Γ multivaluado, sin embargo, uno es guiado a considerar las probabilidades superiores e inferiores definidas como se ve bajo los subconjuntos de S.
Para cada $T \subset S$ se define como:

$$T^* = \{x \subset X, \Gamma x \cap T \neq \oslash\} \tag{1.1}$$

y

$$T_* = \{x \subset X, \Gamma x \neq \oslash \Gamma x \cap T\} \tag{1.2}$$

Dónde T^* es el subconjunto superior y T_* es el subconjunto inferior.

En particular, cuando $S^* = S_*$ es el dominio de Γ. Se define ε como la clase de subconjuntos T de S semejante a T^* y T_* perteneciendo a $\mathfrak{F}$. Suponiendo que $S \in \varepsilon$. Finalmente, se define como la probabilidad superior de $T \in \varepsilon$ como:

$$P^*(T) = \mu(T^*)/\mu(S^*) \tag{1.3}$$

y la probabilidad inferior de $T \in \varepsilon$ se define como:

$$P_*(T) = \mu(T_*)/\mu(S^*) \tag{1.4}$$

$P^*(T)$ y $P_*(T)$ son definidos solo si $\mu(S^*) \neq 0$.

Ya que T^* consiste en aquellos $x \in X$ el cual la posibilidad corresponde bajo Γ a una $s \in T$, uno puede naturalmente respecto $\mu(T^*)$ siendo el mayor valor posible de la probabilidad para la medición μ el cual puede transferir resultados $s \in T$. Similarmente $\mu(T_*)$ consiste en aquellas $x \in X$ el cual puede dirigir a un $s \in T$, entonces que $\mu(T_*)$ representa el mínimo valor de la probabilidad que puede ser transferido al resultado $s \in T$.

El denominador $\mu(S_*)$ en (1.3) y (1.4) es un factor renormalizado necesitado por el factor que el modelo permita, en general, el resultado en X cuando no es trazado dentro del subconjunto significativo de S. El subconjunto $\{x \subset X, \Gamma x \neq \oslash\}$ puede ser eliminado de X y la medición restante del conjunto S_* renormalizado a la unidad.

1.2.3. Teoría de Dempster-Shafer

La teoría de Dempster-Shafer [1] (DST) se desarrolló como una alternativa y generalización de la probabilidad bayesiana. Mientras que la probabilidad bayesiana requiere una asignación explícita de probabilidades a cada evento, DST permite manejar incertidumbres asignando grados de creencia basados en evidencias independientes. Dos pilares fundamentales de esta teoría son en primer lugar, la **función de creencia (Bel)** es un mecanismo que representa los grados de confianza o creencia asignados a un conjunto de hipótesis, basándose en la evidencia disponible. Este enfoque permite modelar situaciones de incertidumbre al proporcionar una medida cuantitativa de la credibilidad que se otorga a diferentes conjuntos de hipótesis.

Por otro lado, la **regla de Dempster** es una herramienta que permite combinar dos funciones de creencia, denotadas como Bel_1 y Bel_2, definidas sobre un mismo marco de discernimiento, Θ. Esta combinación se basa en la suposición de que las evidencias que sustentan cada función son independientes entre sí. El resultado de esta combinación es una nueva función de creencia, Bel_{comb}, que se define de la siguiente manera:

$$Bel_{\text{comb}}(A) = \frac{1}{1-K} \sum_{B \cap C = A} Bel_1(B) \cdot Bel_2(C),$$

donde A representa un subconjunto del marco de discernimiento, y el término K se calcula como:

$$K = \sum_{B \cap C = \emptyset} Bel_1(B) \cdot Bel_2(C),$$

el cual refleja el grado de conflicto entre las dos fuentes de evidencia. Es importante destacar que el factor $\frac{1}{1-K}$ sirve para normalizar la función combinada, asegurando que las creencias asignadas sumen exactamente 1, siempre y cuando el valor de K sea menor que 1.

Es especialmente atractivo en inteligencia artificial y simulaciones, ya que ofrece una forma flexible de gestionar incertidumbres. Sin embargo, requiere

cuidadosa gestión de evidencias conflictivas y una estructura bien definida para representar incertidumbres.

Glenn Shafer, en su libro "A Mathematical Theory of Evidence" [1], establece los principios formales de DST, proporcionando un función de soporte que es una representación matemática que asigna grados de creencia a subconjuntos de hipótesis. También el peso de evidencia como métrica que cuantifica la confianza en la evidencia. Además de la regla de combinación, el cual es un procedimiento para fusionar evidencias independientes, derivado de la intuición de probabilidad condicional. Shafer también destaca limitaciones propias de DST, como la dificultad de manejar conjuntos de evidencia conflictiva, e introduce conceptos como funciones de soporte separables y cuasi-funciones de soporte. Estas bases teóricas han inspirado aplicaciones prácticas y avances en áreas como razonamiento probabilístico, inteligencia artificial y estadística.

Para entenderlo mejor, se mostrará el siguiente ejemplo:

Supongamos un marco de discernimiento $\Theta = \{A, B\}$, donde:

- A representa la proposición "El objeto es Rojo."
- B representa la proposición "El objeto es Azul."

Dos fuentes de evidencia proporcionan funciones de creencia definidas como:

$$\begin{aligned} &\text{Fuente 1:} \quad Bel_1(A) = 0.6, \quad Bel_1(B) = 0.2, \quad Bel_1(\Theta) = 0.8. \\ &\text{Fuente 2:} \quad Bel_2(A) = 0.5, \quad Bel_2(B) = 0.3, \quad Bel_2(\Theta) = 0.8. \end{aligned}$$

El conflicto K se calcula sumando las intersecciones donde las evidencias son incompatibles:

$$K = \sum_{B \cap C = \emptyset} Bel_1(B) \cdot Bel_2(C).$$

Dado que las fuentes no asignan creencias fuera de Θ, $K = 0$. Esto indica que no hay conflicto total.

La creencia combinada para cada proposición se calcula mediante:

$$\begin{aligned} Bel_{\text{comb}}(A) &= \frac{1}{1-K} \sum_{B \cap C = A} Bel_1(B) \cdot Bel_2(C), \\ Bel_{\text{comb}}(B) &= \frac{1}{1-K} \sum_{B \cap C = B} Bel_1(B) \cdot Bel_2(C). \end{aligned}$$

Aplicando esta regla:

$$\begin{aligned} Bel_{\text{comb}}(A) &= \frac{1}{1-0} \cdot \big(Bel_1(A) \cdot Bel_2(A)\big) \\ &= 0.6 \cdot 0.5 = 0.3, \\ Bel_{\text{comb}}(B) &= \frac{1}{1-0} \cdot \big(Bel_1(B) \cdot Bel_2(B)\big) \\ &= 0.2 \cdot 0.3 = 0.06. \end{aligned}$$

El resultado se normaliza para asegurar que las creencias sumen 1:

$$Bel_{\text{final}}(A) = \frac{Bel_{\text{comb}}(A)}{Bel_{\text{comb}}(\Theta)}, \quad Bel_{\text{final}}(B) = \frac{Bel_{\text{comb}}(B)}{Bel_{\text{comb}}(\Theta)}.$$

Dado que $Bel_{\text{comb}}(\Theta) = 0.3 + 0.06 = 0.36$:

$$Bel_{\text{final}}(A) = \frac{0.3}{0.36} \approx 0.833, \quad Bel_{\text{final}}(B) = \frac{0.06}{0.36} \approx 0.167.$$

Es importante mencionar que la creencia mínima sería la fórmula (1.4) mientras que la creencia máxima sería la fórmula (1.3)

1.3. Estado del Arte

Este estado del arte tiene como objetivo analizar las principales aplicaciones y avances relacionados con la Teoría de Dempster-Shafer en inteligencia artificial. Se presentan los enfoques teóricos más relevantes, los casos de uso destacados, y se examinan las ventajas, limitaciones y tendencias actuales en la investigación de esta metodología. Y por ende, se busca proporcionar un panorama completo del impacto y las posibilidades de esta teoría en la evolución de sistemas inteligentes.

1.3.1. Clasificadores basados en DST y Redes Neuronales Profundas

El trabajo de Tong [6] presenta un enfoque innovador que integra la teoría de Dempster-Shafer (DST) con redes neuronales profundas (DNN) para abordar problemas de clasificación con alta incertidumbre. Este sistema combina tres componentes principales que trabajan en conjunto para mejorar tanto la precisión como la robustez del modelo.

Con la extracción de características se realiza a través de una red convolucional (CNN), que procesa datos de entrada como imágenes o señales. La CNN genera representaciones de alto nivel que capturan las características esenciales de los datos, aprovechando su capacidad para modelar patrones complejos.

Estas características son transformadas en funciones de masa mediante una capa especialmente diseñada basada en DST. Esta capa introduce una representación explícita de las incertidumbres inherentes al proceso de clasificación, permitiendo modelar situaciones donde las clases no están claramente definidas debido a ambigüedades en los datos. Así, se logra una interpretación más cautelosa y precisa de la información disponible.

Finalmente el sistema realiza la clasificación mediante una capa de utilidad que permite asignaciones set-valued. En lugar de forzar una decisión en casos de alta incertidumbre, esta capa clasifica una instancia en más de una clase cuando sea necesario, reduciendo el riesgo de errores significativos. Este enfoque es particularmente útil en contextos donde los datos presentan ruidos o las clases tienen solapamientos considerables.

Este modelo híbrido no solo mejora la precisión en tareas de clasificación, sino que también presenta otras ventajas importantes. Por un lado, reduce significativamente los errores al manejar eficazmente patrones ambiguos

y datos fuera del conjunto de entrenamiento, lo que incrementa su capacidad para operar en escenarios del mundo real. Por otro lado, amplía su aplicabilidad a diversos dominios, como el análisis de imágenes médicas y el procesamiento de señales, donde las incertidumbres y anomalías son desafíos comunes. Además, la integración de DNN con DST permite combinar el poder del aprendizaje profundo para extraer características complejas con un marco formal capaz de gestionar incertidumbres.

El enfoque propuesto por Tong demuestra cómo la combinación de redes neuronales profundas y la teoría de Dempster-Shafer puede ofrecer soluciones más precisas y adaptables en problemas de clasificación, abriendo nuevas posibilidades en campos críticos donde es esencial tomar decisiones confiables.

1.3.2. Clasificación Flexible e Interpretabilidad

Peñafiel [7] aborda uno de los principales desafíos en inteligencia artificial; el balance entre precisión e interpretabilidad. Propusieron un método que aprende funciones de masa utilizando el descenso de gradiente para optimizar las funciones de masa a partir de datos, asegurando que las decisiones se basen en evidencia sólida derivada del conjunto de entrenamiento. Dando la posibilidad de una integración híbrida combinando reglas de expertos con datos históricos, creando un sistema que aprovecha tanto el conocimiento humano como la evidencia empírica. Este enfoque permite que el modelo mantenga un equilibrio entre flexibilidad y precisión. Dando explicaciones basadas en estas reglas, facilitando la transparencia en decisiones críticas, particularmente en áreas como medicina y finanzas, donde es necesario justificar cada resultado para cumplir con regulaciones y estándares éticos.

Un ejemplo relevante incluye el cumplimiento de regulaciones legales en el sector financiero chileno, donde los modelos deben justificar decisiones como el rechazo de créditos hipotecarios sin recurrir a variables discriminatorias. Este enfoque demostró cómo las explicaciones claras y basadas en reglas pueden fortalecer la confianza en los sistemas de IA, especialmente en dominios donde la equidad y la transparencia son esenciales.

1.3.3. Monitoreo de Condición y Diagnóstico

En el ámbito industrial, el monitoreo de condición (Condition Monitoring, CM) se ha beneficiado de la integración de DST con inteligencia artificial.

Según Rosli [8] Con la evidencia combinada, en lugar de analizar parámetros individuales (como vibración o presión), se combinan múltiples fuentes de datos usando las reglas de DST. Esto permite una evaluación integral del estado de los equipos, proporcionando un análisis más robusto y confiable. Y con el diagnóstico de fallos usando técnicas avanzadas como redes neuronales y máquinas de soporte vectorial (SVM) generan funciones de masa que se combinan para determinar la probabilidad de fallos en componentes clave, como motores, compresores o turbinas. Este enfoque mejora la detección temprana de problemas, minimizando tiempos de inactividad y costos asociados.

Además, la incorporación de algoritmos avanzados, como redes neuronales de retropropagación y lógica difusa, ha mejorado la precisión de los diagnósticos al reducir la incertidumbre en escenarios complejos. Estos avances son particularmente valiosos en industrias donde las fallas mecánicas pueden tener consecuencias catastróficas, como en la generación de energía, la aviación y la manufactura de alta precisión.

Un caso de estudio destacado incluye el uso de DST en sistemas de monitoreo predictivo para plantas de energía, donde se analizaron múltiples fuentes de datos para predecir fallos en turbinas. Este enfoque permitió no solo identificar fallos potenciales con mayor precisión, sino también optimizar programas de mantenimiento basados en las probabilidades de falla combinadas.

1.3.4. Algoritmo de red neuronal Dempster-Shafer para navegación de vehículos terrestres

Aggarwal [9] propone un algoritmo basado en la integración de la teoría de Dempster-Shafer y redes neuronales artificiales (ANN) para mejorar la navegación de vehículos terrestres utilizando datos de sistemas de navegación inercial (INS) y de posicionamiento global (GPS). Los componentes principales de su propuesta incluyen, las ventajas de DST a diferencia de los métodos tradicionales que dependen de distribuciones de probabilidad conocidas, DST asigna masas de probabilidad a conjuntos de posibilidades, lo que permite manejar incertidumbres de manera más efectiva. Como propuesta híbrida de combinar DST y ANN para gestionar incertidumbres en sistemas de bajo costo, especialmente en situaciones de pérdida de señal GPS o entornos con interferencias. Este enfoque aprovecha la capacidad de aprendizaje

de las ANN junto con la robustez de DST en la fusión de evidencias. Y con los resultados obtenidos la implementación mostró mejoras significativas en la precisión de la navegación en comparación con técnicas basadas únicamente en ANN o filtros de Kalman extendidos. Esto se traduce en trayectorias más confiables incluso en entornos desafiantes.

El enfoque híbrido de Aggarwal resulta particularmente relevante en aplicaciones donde los sensores pueden proporcionar datos incompletos o contradictorios, como en sistemas de navegación autónoma.

1.3.5. Diagnóstico automatizado de fallos en rodamientos

Hui [10] desarrolla un algoritmo híbrido que combina redes neuronales artificiales (ANN) y la teoría de Dempster-Shafer (DST) para diagnosticar fallos en rodamientos. Este enfoque se basa en datos de vibraciones procesados por herramientas estadísticas como análisis de onda y transformaciones empíricas, y presenta los siguientes elementos clave, una estrategia que es la primera capa del sistema clasifica las señales mediante ANN; sin embargo, los resultados ambiguos o inciertos se refinan utilizando DST en una segunda capa, lo que mejora la precisión de las decisiones. Mostrando los beneficios cuando DST elimina resultados conflictivos y mejora significativamente la clasificación, especialmente en escenarios donde las señales están contaminadas con ruido o provienen de entornos de operación variables. Y con los resultados obtenidos el método híbrido demuestra ser más preciso y confiable que el uso independiente de ANN o herramientas de aprendizaje automático tradicionales. Esto se traduce en diagnósticos más efectivos en sistemas críticos.

Este enfoque es especialmente relevante en aplicaciones industriales donde el diagnóstico preciso de fallos en componentes rotativos, como rodamientos, es esencial para evitar fallas catastróficas. Los experimentos realizados por Hui validaron la eficacia del sistema en entornos industriales reales, destacando su capacidad para manejar señales complejas y mejorar la confiabilidad operativa.

1.3.6. Diagnóstico de severidad de fallos en rodamientos usando DBN

Yu [11] propone un marco jerárquico basado en redes de creencias profundas (DBN) y DST para diagnosticar tanto la presencia de fallos como su

gravedad. Este enfoque introduce los siguientes aportes clave, que son la optimización híbrida con el uso de algoritmos de optimización como Particle Swarm Optimization (PSO) y Algoritmos Genéticos (GA) para ajustar los parámetros de las DBN. Esto mejora la eficiencia y precisión del modelo, permitiendo un entrenamiento más robusto. Usando la fusión de información con la teoría de DST que se emplea para integrar los resultados de múltiples clasificadores, lo que facilita el manejo de señales complejas y datos provenientes de diversas orientaciones. Esta capacidad de fusión permite resolver conflictos y gestionar incertidumbres inherentes en sistemas reales.
La metodología propuesta mejora significativamente la precisión y confiabilidad en comparación con enfoques previos, especialmente en escenarios donde los datos son ruidosos o presentan alta dimensionalidad.

El trabajo de Yu tiene aplicaciones destacadas en áreas como la industria manufacturera y la monitorización de maquinaria crítica, donde la identificación precisa de fallos y su severidad es esencial para garantizar la continuidad operativa y reducir los costos de mantenimiento no planificado.

1.3.7. Clasificación de calidad del agua usando ANN, SVM y DST

El trabajo realizado por Ladjal [12] propone un enfoque híbrido para la clasificación de la calidad del agua en el embalse Tilesdit, ubicado en Argelia, mediante la integración de redes neuronales artificiales (ANN), máquinas de soporte vectorial (SVM) y la teoría de Dempster-Shafer (DST). Este estudio se centra en el monitoreo de los recursos hídricos y realiza aportes significativos en diversos aspectos.

Primero, se introduce un método multicategoría que clasifica los datos en tres niveles de calidad del agua, basándose en parámetros fisicoquímicos como el pH, la turbidez y la temperatura. Esta metodología permite realizar evaluaciones detalladas y precisas de la calidad del agua en el embalse.

Además, el estudio destaca la fusión de decisiones como una de sus principales contribuciones. Para ello, se utiliza la teoría de Dempster-Shafer como herramienta para combinar los resultados obtenidos de las ANN y las SVM, resolviendo conflictos y mejorando la precisión en la clasificación. Este enfoque resulta especialmente efectivo en contextos donde los datos son ruidosos o ambiguos, ya que permite manejar la incertidumbre de manera más eficiente.

Por último, el impacto de este enfoque híbrido se refleja en su capacidad para superar el rendimiento de los clasificadores individuales en términos de precisión y robustez. Esto es especialmente relevante para el monitoreo de recursos hídricos en regiones caracterizadas por datos complejos y ruidosos, contribuyendo significativamente a la gestión sostenible de estos recursos.

Los resultados del estudio validan la eficacia del enfoque híbrido para ofrecer clasificaciones más confiables, destacando su potencial aplicación en ámbitos ambientales y en la gestión de recursos naturales.

1.3.8. Clasificador neuronal basado en DST para patrones adaptativos

Denoeux [13] presenta un clasificador neuronal basado en la teoría de Dempster-Shafer (DST), diseñado para gestionar incertidumbres y fusionar evidencia en la clasificación de patrones. Este enfoque introduce varios elementos destacados que lo hacen innovador y efectivo.

Se propone una arquitectura novedosa, donde una red neuronal incluye capas especializadas para calcular asignaciones de probabilidad y combinar evidencia utilizando la regla de DST. Esta estructura permite manejar de manera eficaz escenarios con datos conflictivos o incompletos, optimizando la precisión en la clasificación.

Además, el clasificador cuenta con capacidades avanzadas que lo distinguen de otros enfoques. Puede rechazar patrones ambiguos y detectar novedades, adaptándose de forma dinámica a cambios en el entorno y a datos no vistos previamente. Esto le confiere una gran flexibilidad y utilidad en contextos variables.

El método también destaca por sus aplicaciones prácticas, especialmente en sistemas de diagnóstico industrial y en la clasificación en tiempo real en entornos complejos. Su robustez frente a fallos de sensores y cambios ambientales supera las limitaciones de técnicas estadísticas y neuronales tradicionales.

Finalmente, los experimentos realizados por Denoeux validan este enfoque, demostrando mejoras significativas en términos de precisión y robustez, incluso cuando los datos son ruidosos o inconsistentes. Este trabajo subraya el potencial del clasificador neuronal en diversas aplicaciones prácticas, consolidándolo como una herramienta efectiva en el campo de la inteligencia artificial y el análisis de patrones.

1.3.9. Regresión Logística y Redes Neuronales desde la Perspectiva de la Teoría de Dempster-Shafer

Thierry Denoeux [14] explora cómo los algoritmos tradicionales, como la regresión logística y las redes neuronales, pueden ser interpretados dentro del marco de la teoría de Dempster-Shafer (DST). Este enfoque permite transformar las salidas de estos modelos en funciones de masa DST, facilitando la cuantificación tanto de la evidencia a favor de una clase como de la falta de información discriminativa.

Una de las contribuciones principales del trabajo es la introducción de decisiones basadas en intervalos de dominancia, una técnica que mejora la capacidad de tomar decisiones en situaciones donde la evidencia no es concluyente. Este método resulta especialmente útil en aplicaciones que requieren un manejo explícito de la incertidumbre, como la detección de novedades y la calibración evidencial.

El estudio también destaca por reinterpretar algoritmos supervisados clásicos bajo el marco de DST, proporcionando un enfoque más informativo para representar la incertidumbre. Además, permite identificar casos en los que la evidencia es insuficiente para tomar decisiones concluyentes, reduciendo así el riesgo de errores en entornos complejos. Entre las aplicaciones prácticas más relevantes se encuentran la detección de anomalías, el diagnóstico en entornos industriales y los sistemas de decisión médica, donde la explicitación de la incertidumbre contribuye a mejorar tanto la confianza como la interpretación de los resultados.

En conjunto, este enfoque representa un avance significativo en la integración de la teoría de Dempster-Shafer con modelos tradicionales de aprendizaje supervisado, ofreciendo herramientas más robustas para la toma de decisiones en escenarios marcados por la incertidumbre.

1.3.10. Predicción de Sentimientos Basada en la Teoría de Dempster-Shafer

Basiri [15] y su equipo desarrollaron un enfoque innovador que aplica la teoría de Dempster-Shafer (DST) al análisis de sentimientos en textos libres. Este método jerárquico combina las puntuaciones de polaridad de las oraciones para calcular una polaridad general del texto, logrando mejoras significativas en comparación con los métodos tradicionales.

El proceso comienza con el análisis de polaridad de cada oración del texto,

determinando si es positiva, negativa o neutral. Esta información se considera como evidencia individual. Posteriormente, las puntuaciones de polaridad obtenidas de las oraciones se fusionan utilizando la regla de combinación de Dempster, lo que permite integrar múltiples fuentes de evidencia al mismo tiempo que se maneja de manera efectiva la incertidumbre y los conflictos. Finalmente, la polaridad combinada proporciona una visión global del texto, lo que mejora la precisión de la clasificación.

Los experimentos realizados en conjuntos de datos como CitySearch y TripAdvisor demostraron que este enfoque supera a métodos convencionales como la votación mayoritaria o el promedio simple. La utilización de DST permite capturar mejor las variabilidades y contradicciones presentes en los textos, resultando en un análisis de sentimientos más robusto y confiable.

Este método tiene un impacto notable en diversas aplicaciones. Por ejemplo, en los sistemas de recomendación, mejora la interpretación de las reseñas de usuarios en plataformas de comercio electrónico y turismo. En el análisis de redes sociales, ayuda a comprender mejor las opiniones y tendencias en plataformas sociales. Además, en el ámbito de los asistentes virtuales, permite generar respuestas más precisas y contextualmente relevantes en aplicaciones de atención al cliente.

Este trabajo basado en DST representa un avance significativo en el análisis de sentimientos. Su capacidad para manejar la incertidumbre y resolver conflictos de evidencia lo convierte en una herramienta más informativa y precisa, destacando su potencial en múltiples aplicaciones prácticas.

1.3.11. Predicción de Hipertensión Usando Imágenes Faciales

Ang [16] propone un método no invasivo para predecir la hipertensión mediante el análisis de características faciales y técnicas de procesamiento de imágenes. Este enfoque innovador destaca por su aplicabilidad en dispositivos portátiles y aplicaciones móviles, ofreciendo una solución accesible y práctica para el monitoreo de la salud.

El estudio se basa en imágenes faciales de 1,099 sujetos, a partir de las cuales se identifican características asociadas con la hipertensión. El proceso comienza con la extracción de características utilizando técnicas avanzadas de procesamiento de imágenes. Se analizan variables como la forma y el color de áreas específicas del rostro, incluyendo las mejillas, la nariz y la frente.

Posteriormente, estas características son sometidas a análisis estadísticos mediante métodos como el análisis de covarianza (ANCOVA) y la técnica de selección LASSO, para identificar correlaciones significativas con la hipertensión. Los resultados del estudio validan que ciertos atributos faciales, como la forma de la nariz y los colores en las mejillas y la frente, presentan una relación notable con esta condición médica.

Este enfoque tiene un impacto potencial significativo en diversos ámbitos. En primer lugar, podría integrarse en dispositivos portátiles como relojes inteligentes o aplicaciones móviles para el monitoreo diario de la salud. También puede servir como herramienta de diagnóstico preventivo, facilitando la detección temprana de hipertensión tanto en entornos clínicos como no clínicos. Además, puede contribuir a la promoción de la salud al proporcionar una herramienta accesible para concienciar y manejar de forma activa los factores de riesgo asociados.

El trabajo de Ang demuestra cómo la combinación de tecnologías de procesamiento de imágenes y análisis estadístico puede transformar la forma en que se detectan y monitorean las condiciones médicas, abriendo nuevas posibilidades para sistemas de salud más accesibles, personalizados y eficaces.

1.3.12. El Problema del Newsvendor: Revisión y Extensiones

Moutaz Khouja [17] presenta una revisión integral del problema clásico de un solo período (SPP), también conocido como el problema del "newsvendor". Este modelo se centra en determinar la cantidad óptima de inventario que maximiza las ganancias esperadas cuando la demanda es incierta.

El autor organiza las extensiones del SPP en once categorías principales. Entre estas, destacan los modelos que adoptan objetivos alternativos, como la minimización de costos o la maximización de la satisfacción del cliente. También se incluyen las políticas de precios dinámicos, que permiten ajustar los precios en función de factores como la demanda y las condiciones del mercado. Otra categoría aborda los modelos multi-producto, que optimizan inventarios para múltiples artículos con interdependencias, y los sistemas multi-localización, que gestionan inventarios en múltiples ubicaciones o centros de distribución.

El estudio también señala posibles direcciones para futuras investigaciones. Entre ellas se encuentra la integración del aprendizaje automático para mejorar las estimaciones de demanda y optimizar decisiones. Asimismo, se sugiere incorporar incertidumbres adicionales en los modelos, como cambios

en la oferta o eventos disruptivos. Finalmente, se destaca la importancia de explorar nuevas aplicaciones del SPP en industrias emergentes como el comercio electrónico, la salud y la energía renovable.

Este artículo constituye una referencia esencial para investigadores y profesionales interesados en la gestión de inventarios y la optimización industrial. Al combinar enfoques tradicionales con técnicas modernas, proporciona un marco versátil para abordar problemas más complejos en un amplio espectro de aplicaciones, subrayando la relevancia del SPP en la toma de decisiones estratégicas.

1.3.13. Una Vista Simple de la Teoría de Dempster-Shafer

Lotfi Zadeh [18] ofrece una explicación accesible de la teoría de Dempster-Shafer (DST) utilizando un enfoque intuitivo basado en bases de datos relacionales. Este trabajo destaca cómo los conceptos de DST pueden extenderse para mejorar la gestión de incertidumbre en sistemas expertos y aplicaciones prácticas.

En este contexto, Zadeh introduce los conceptos de creencia y plausibilidad, los cuales representan intervalos de probabilidad en situaciones de incertidumbre. Además, el artículo explora cómo las relaciones de segundo orden, es decir, atributos no determinísticos, pueden modelarse dentro del marco de DST para responder a consultas complejas en bases de datos. Esto permite manejar datos incompletos de una manera más robusta y flexible.

El trabajo realiza varias contribuciones significativas. Primero, demuestra cómo la representación de creencia y plausibilidad puede modelar incertidumbres en datos incompletos, proporcionando un marco más preciso para el análisis. Luego, extiende las bases de datos relacionales al incluir atributos no determinísticos, mejorando su capacidad para gestionar incertidumbre. Y por último, aplica estas ideas en sistemas expertos, integrando datos incompletos y mejorando la toma de decisiones en entornos de inteligencia artificial.

Este enfoque tiene aplicaciones prácticas notables. Permite la integración de datos incompletos, mejorando la calidad de las consultas en bases de datos relacionales con información parcial. También amplía las capacidades de los sistemas expertos para tomar decisiones en entornos inciertos, como el diagnóstico médico y el análisis financiero. Además, resalta la relevancia de DST en inteligencia artificial moderna, especialmente en el manejo de incertidumbre y conflictos en sistemas avanzados.

El artículo de Zadeh enfoca la importancia de DST en la inteligencia artificial y en el manejo de datos relacionales complejos.

1.3.14. Apoyo a la preparación colaborativa de planes de emergencia

La planificación de emergencias es una tarea compleja que requiere la integración de diversas fuentes de información, enfoques metodológicos avanzados y la colaboración interdisciplinaria. En este marco, el artículo "Supporting Collaborative Preparation of Emergency Plans" [19] presenta un análisis detallado y una propuesta innovadora para afrontar estos retos.

La teoría de Dempster-Shafer (DST) se posiciona como una herramienta eficaz para manejar datos incompletos y escenarios inciertos en la planificación de emergencias. Esta teoría permite asignar grados de plausibilidad, certeza e incertidumbre a hipótesis relacionadas con escenarios específicos, proporcionando una base sólida para la toma de decisiones.

En el artículo la DST se aplica en tres áreas clave. Primero, se utiliza para generar mapas de adecuación que visualizan áreas con mayor probabilidad de cumplir criterios específicos, como seguridad en casos de tsunamis. Después, facilita el manejo de incertidumbre al incorporar y procesar información incompleta o conflictiva proveniente de múltiples fuentes, mejorando la precisión de las decisiones. Y finalmente, la DST se emplea como un marco común que permite la colaboración efectiva entre expertos de distintas disciplinas en la elaboración de planes de emergencia.

Este enfoque tiene un impacto significativo en varios aspectos de la gestión de emergencias. En la planificación territorial, permite identificar zonas seguras y vulnerables para diseñar planes de evacuación efectivos. En la respuesta a emergencias, optimiza el uso de recursos y estrategias basándose en evaluaciones probabilísticas robustas. Asimismo, en el entrenamiento y la simulación, facilita la creación de escenarios más realistas que reflejan las incertidumbres y conflictos inherentes a situaciones de emergencia.

El trabajo demuestra cómo la DST puede integrarse en herramientas prácticas para abordar desafíos críticos en la gestión de emergencias, fortaleciendo la capacidad de respuesta ante desastres y minimizando el impacto en las comunidades afectadas. Este enfoque destaca la importancia de combinar métodos avanzados y colaboración interdisciplinaria para mejorar la preparación y respuesta ante situaciones de crisis.

1.3.15. Agregación no jerárquica de estructuras de creencias

Bhattacharya [20] presenta modelos de agregación no jerárquica diseñados para priorizar estructuras de creencias de manera dinámica, una característica especialmente útil en sistemas donde las prioridades cambian frecuentemente. Estos modelos destacan por su flexibilidad, ya que permiten ajustar las prioridades de las estructuras de creencias en tiempo real, adaptándose a cambios contextuales.

En aplicaciones como el reconocimiento automático de objetivos, los modelos de Bhattacharya permiten que los grados de importancia de ciertos objetivos varíen dinámicamente según las estrategias tácticas. Por otro lado, en la gestión de carteras de inversión, las prioridades de las acciones pueden ajustarse en función de fluctuaciones de precios o cambios en el mercado. Esta capacidad de adaptación resulta particularmente valiosa en escenarios complejos y dinámicos donde las condiciones pueden cambiar rápidamente. Además, los modelos sobresalen en su capacidad para manejar múltiples niveles de incertidumbre, aumentando su relevancia en estos contextos.

No obstante, estos avances vienen acompañados de desafíos importantes. En primer lugar, la alta complejidad computacional de la regla de combinación de Dempster representa una barrera significativa. En escenarios con múltiples fuentes de evidencia y alta dimensionalidad, el costo computacional se incrementa exponencialmente debido al número de subconjuntos que deben evaluarse. En segundo lugar, el manejo de conflictos extremos entre las evidencias plantea otro reto. Cuando las fuentes son altamente contradictorias, los resultados generados por la regla de combinación pueden ser poco intuitivos o inestables. Este problema ha impulsado la investigación de métodos alternativos o aproximados que permitan gestionar los conflictos de manera más efectiva.

En resumen, los modelos de Bhattacharya abren nuevas posibilidades en la gestión dinámica de prioridades y en el manejo de incertidumbre en sistemas complejos. Sin embargo, superar los desafíos de complejidad computacional y manejo de conflictos extremos será crucial para maximizar su aplicabilidad y efectividad en escenarios prácticos.

1.3.16. Integración de Dempster-Shafer y Naive Bayes

Mulyani [21] propone un modelo híbrido que combina la teoría de Dempster-Shafer (DST) con el método de Naive Bayes para mejorar el diagnóstico de

enfermedades febriles. Este enfoque busca superar las limitaciones de cada técnica, aprovechando sus fortalezas complementarias. Por un lado, DST es útil para manejar incertidumbres y combinar evidencias de múltiples fuentes, pero puede generar resultados ambiguos cuando enfrenta altos niveles de conflicto entre las evidencias. Por otro lado, Naive Bayes, basado en probabilidades condicionales, produce una única solución clara, lo que facilita la interpretación de los resultados, aunque depende en gran medida de la calidad de los datos de entrenamiento.

El modelo propuesto opera en dos etapas. En la primera, se aplica DST para integrar evidencias proporcionadas por expertos médicos, capturando las incertidumbres asociadas con los síntomas reportados por los pacientes. Si DST genera resultados ambiguos debido a conflictos en las evidencias, se recurre a Naive Bayes en una segunda etapa. Este método utiliza datos históricos para identificar la enfermedad más probable, resolviendo las ambigüedades presentes en los resultados de DST.

La validación del enfoque se realizó mediante experimentos con datos médicos reales provenientes de hospitales en Indonesia. El modelo demostró ser eficaz en el diagnóstico de cinco enfermedades febriles comunes: dengue, malaria, meningitis, infecciones respiratorias y fiebre tifoidea. Aunque la precisión del modelo estuvo limitada por la cantidad y calidad de los datos de entrenamiento disponibles, el enfoque híbrido presentó ventajas significativas en comparación con el uso independiente de DST o Naive Bayes. Este trabajo destaca el potencial de combinar métodos probabilísticos y basados en evidencia para mejorar el diagnóstico médico en entornos donde los datos pueden ser incompletos o contradictorios.

1.3.17. Marco de aprendizaje en conjunto basado en Dempster-Shafer para el nowcasting de la contaminación del aire

Tung [22] para el contexto del nowcasting, la predicción de fenómenos meteorológicos en horizontes temporales muy cortos ha evidenciado un interés creciente en la aplicación de técnicas de aprendizaje profundo para mejorar la precisión de las estimaciones de calidad del aire. Este tipo de predicción resulta crucial tanto para la sostenibilidad urbana como para la gestión de eventos extremos, debido a la volatilidad y complejidad del ambiente atmosférico urbano.

Diversas investigaciones han abordado esta problemática mediante el uso de modelos de aprendizaje profundo individuales y conjuntos (ensembles). Por ejemplo, la Administración Nacional de Meteorología de Rumanía desarrolló NowDeepN, un sistema que combina redes neuronales profundas para predecir con precisión precipitaciones intensas y granizo utilizando datos de radar. De forma similar, NowcastNet, desarrollado a partir de datos de la NOAA y de la Administración Meteorológica China, integra modelos físicos con redes neuronales. En España, un enfoque de ensamble fue utilizado para predecir eventos de niebla, logrando mejores resultados que modelos individuales.

Frente a la necesidad de métodos robustos y adaptables, especialmente para variables como el material particulado (PM2.5), que presentan una alta variabilidad espacial y temporal, algunos trabajos recientes han recurrido a la teoría de evidencia de Dempster-Shafer (DSET). Esta teoría ha demostrado ser eficaz en la fusión de datos y modelos, y ha sido empleada, por ejemplo, en Argelia para mejorar la clasificación de precipitaciones, o en estudios de susceptibilidad a inundaciones mediante la combinación de Random Forest y SVM. DSET se presenta así como una herramienta poderosa para integrar estimaciones diversas y manejar la incertidumbre inherente a los entornos urbanos.

1.4. Problemas del uso de la inteligencia artificial bayesiana

Las redes neuronales en inteligencia artificial enfrentan diversos desafíos al integrarse con el modelo bayesiano, especialmente en comparación con la teoría de Dempster-Shafer (DST), que ofrece un enfoque más flexible y adaptativo en contextos de incertidumbre y datos complejos.

Uno de los principales problemas del modelo bayesiano es su dependencia de distribuciones de probabilidad a priori, que pueden ser difíciles de especificar cuando no se cuenta con información previa confiable. Esto introduce el riesgo de sesgos significativos en los resultados. Por el contrario, DST no depende de estas distribuciones, lo que le permite manejar de forma más efectiva situaciones donde los datos iniciales son escasos o poco fiables.

En escenarios con datos insuficientes o ruidosos, los modelos bayesianos pueden generar estimaciones poco confiables debido a su alta dependencia de la calidad y cantidad de datos. Además, suelen sobreajustarse a los da-

tos ruidosos, comprometiendo su capacidad de generalización. DST, por su parte, maneja la incertidumbre de manera explícita, integrando evidencias de múltiples fuentes con mayor robustez y tolerando inconsistencias en los datos.

La complejidad computacional también representa un desafío significativo para las redes neuronales bayesianas, ya que procesos como la inferencia y la estimación de parámetros suelen ser costosos en términos computacionales, especialmente al emplear métodos como Monte Carlo Markov Chains (MCMC). Aunque DST también puede ser computacionalmente exigente en escenarios con muchas hipótesis, su representación compacta de creencias y plausibilidades reduce los cálculos necesarios, haciéndolo más eficiente en ciertas aplicaciones.

Otro aspecto relevante es el manejo de conflictos en la evidencia. El modelo bayesiano no aborda de forma explícita las inconsistencias entre fuentes de datos, limitándose a suavizarlas mediante ajustes probabilísticos. Esto puede ocultar discrepancias importantes que podrían ser cruciales para la toma de decisiones. DST, en cambio, representa y cuantifica el conflicto de manera directa, permitiendo identificar y analizar contradicciones en la evidencia para mejorar las decisiones.

Además, el modelo bayesiano requiere asignar probabilidades a todas las hipótesis posibles, incluso cuando algunas carecen de evidencia suficiente, lo que puede llevar a una representación artificialmente precisa. DST ofrece una ventaja al permitir asignar creencias a subconjuntos de hipótesis, utilizando intervalos de certeza y plausibilidad que reflejan la incertidumbre de manera más realista y adaptable.

La interpretabilidad también es un área donde DST supera al modelo bayesiano. En problemas complejos y de alta dimensionalidad, los resultados bayesianos pueden ser difíciles de entender, lo que limita su aplicación en áreas críticas como la medicina. DST facilita la comprensión al separar conceptos de certeza y plausibilidad, lo que lo hace más accesible incluso para audiencias no técnicas.

Por último, la escalabilidad en sistemas con múltiples fuentes de evidencia es otro desafío del modelo bayesiano, especialmente cuando las fuentes son interdependientes o tienen niveles de confianza variables. DST maneja eficientemente estas situaciones, combinando evidencia heterogénea y gestionando conflictos de manera efectiva, lo que lo hace ideal para aplicaciones en sistemas distribuidos, redes de sensores y análisis de big data.

Aunque las redes neuronales bayesianas son útiles en muchos contextos, la teoría de Dempster-Shafer ofrece una alternativa poderosa y más versátil para manejar incertidumbre, conflictos en la evidencia y problemas de interpretación, consolidándose como una herramienta clave en inteligencia artificial y análisis de datos complejos.

Capítulo 2

Analisis de resultados de modelos de inteligencia artificial usando Teoría de Demspter-Shafer

2.1. Respuestas al momento de utilizar inteligencias artificiales

En los últimos años, la inteligencia artificial ha adquirido un papel relevante en el desarrollo de programas y aplicaciones orientadas al análisis de datos y a la automatización de procesos. Estas herramientas, diseñadas por especialistas, tienen el potencial de aportar información valiosa para la toma de decisiones en distintos contextos. Sin embargo, en muchos casos, los resultados que presentan dichos programas no son comprensibles para los usuarios finales.

Esta dificultad no solo se origina en la complejidad propia de los modelos de inteligencia artificial, sino también en la forma en que los desarrolladores de programas diseñan la presentación de los resultados. Con frecuencia, la información se comunica en formatos técnicos, poco intuitivos o sin un adecuado proceso de simplificación, lo que limita su accesibilidad para quienes no poseen un conocimiento especializado.

Como consecuencia, los usuarios encuentran obstáculos para interpretar la información, aplicarla correctamente o confiar en ella, lo cual reduce la efectividad de las herramientas basadas en inteligencia artificial. Este pro-

blema evidencia la necesidad de que los desarrolladores no solo se concentren en la precisión técnica de los modelos, sino también en la forma en que los resultados son comunicados, con el fin de garantizar que sean claros, útiles y adaptados a las necesidades reales de las personas.

2.2. Respuestas utilizando la teoría de Demspter-Shafer

Al analizar los datos generados por redes neuronales basadas en la teoría de Dempster-Shafer, es posible identificar información altamente relevante para el proceso de interpretación, particularmente en lo que respecta a los features más influyentes en la obtención de un resultado. Comprender qué atributos tienen mayor peso y por qué desempeñan un papel fundamental en la generación de las predicciones constituye un aspecto esencial, ya que, como se observó en la sección 1.3, los resultados obtenidos mediante este enfoque muestran una mayor precisión en comparación con los derivados de redes neuronales bayesianas tradicionales.

Sin embargo, aunque el modelo ofrece medidas como el valor de masa, la certeza y la plausibilidad asociadas a cada feature, estos indicadores suelen expresarse en forma numérica, lo que dificulta su interpretación directa por parte de los usuarios. En otras palabras, disponer únicamente de cifras no permite comprender fácilmente el grado de importancia o la influencia real que cada variable ejerce sobre el resultado final.

Ante esta limitación, surge la necesidad de desarrollar mecanismos que permitan traducir dichas métricas a un lenguaje más de alto nivel. Un sistema de interpretación en lenguaje natural facilitaría no solo la comprensión de los valores obtenidos, sino también la utilización práctica de los resultados en investigaciones posteriores. De esta manera, los analistas podrían extraer conclusiones más claras, apoyar la toma de decisiones con mayor confianza y reducir la brecha existente entre el análisis matemático del modelo y su aplicabilidad en escenarios reales.

2.3. DSExplainer: Marco para explicabilidad con intervalos de creencia y plausibilidad

El framework DSExplainer constituye un avance en la explicabilidad de modelos de inteligencia artificial al integrar valores SHAP con la teoría de la evidencia de Dempster–Shafer (DST), con el objetivo de enriquecer la interpretación de atribuciones mediante una caracterización explícita de su incertidumbre epistémica. En particular, el enfoque preserva la descomposición aditiva propia de SHAP, pero la complementa con una representación intervalar (creencia–plausibilidad) que permite evaluar la confiabilidad de cada explicación. Esta capacidad resulta especialmente pertinente en dominios de alto impacto, tales como la salud, las finanzas y la toma de decisiones ambientales, donde la estabilidad de una explicación es un requisito metodológico adicional a su magnitud.

2.3.1. Fundamentos teóricos

La teoría de Dempster–Shafer extiende el formalismo probabilístico clásico al permitir representar de forma separada la evidencia disponible y la ignorancia residual. Sea Θ un marco de discernimiento y 2^Θ su conjunto potencia; una asignación básica de probabilidad (basic probability assignment, BPA) se define como una función $m : 2^\Theta \to [0, 1]$ que satisface:

$$m(\emptyset) = 0 \quad \text{y} \quad \sum_{A \subseteq \Theta} m(A) = 1. \tag{2.1}$$

Para cualquier hipótesis $H \subseteq \Theta$, se definen las funciones de creencia y plausibilidad como:

$$Bel(H) = \sum_{A \subseteq H} m(A), \tag{2.2}$$

$$Pl(H) = \sum_{A \cap H \neq \emptyset} m(A). \tag{2.3}$$

El intervalo $[Bel(H), Pl(H)]$ entrega una representación intervalar del soporte evidencial asociado a H; en particular, la diferencia $Pl(H) - Bel(H)$ se interpreta como una medida de incertidumbre (o ignorancia) respecto de dicha hipótesis.

2.3.2. Mapeo de SHAP a DST

El mapeo propuesto por DSExplainer puede describirse como un procedimiento en etapas que transforma contribuciones SHAP en masas de evidencia definidas sobre un conjunto de hipótesis explicativas. Primero, para cada instancia se calculan los valores SHAP ϕ_i, que cuantifican la contribución de cada característica al resultado del modelo. A continuación, las magnitudes de las contribuciones (en valor absoluto) se transforman en masas de evidencia mediante una función de mapeo g:

$$m'(H) = g(|\phi_H|), \quad \text{donde } H \in \mathcal{F}, \tag{2.4}$$

donde $\mathcal{F}$ denota el conjunto de hipótesis explicativas consideradas, incluyendo características individuales y, cuando corresponda, interacciones seleccionadas.

Posteriormente, se realiza una normalización para obtener una BPA válida, reservando además una masa explícita de ignorancia para capturar evidencia no modelada:

$$m(H) = \frac{m'(H)}{\sum_{A \in \mathcal{F}} m'(A)} \cdot (1 - m(\Theta)), \qquad m(\Theta) = \epsilon. \tag{2.5}$$

Finalmente, el signo de la contribución se maneja de manera separada del proceso de asignación de masas, con el fin de preservar la dirección del efecto:

$$s(H) = \operatorname{sgn}(\phi_H) \in \{-1, 0, 1\}. \tag{2.6}$$

2.3.3. Fusión de evidencia

Con el fin de estabilizar las asignaciones de masa frente a la variabilidad muestral, DSExplainer estima BPAs sobre B réplicas bootstrap y fusiona la evidencia mediante la regla de Dempster:

$$m = m_1 \oplus m_2 \oplus \cdots \oplus m_B. \tag{2.7}$$

Para dos fuentes m_1 y m_2, la combinación se define como:

$$(m_1 \oplus m_2)(H) = \frac{\sum_{A \cap B = H} m_1(A)\, m_2(B)}{1 - K}, \qquad K = \sum_{A \cap B = \emptyset} m_1(A)\, m_2(B), \tag{2.8}$$

donde K cuantifica el conflicto entre las fuentes de evidencia. Como resultado, se obtienen intervalos $[Bel(H), Pl(H)]$ que integran tanto la magnitud de la atribución como su estabilidad epistémica bajo perturbaciones inducidas por remuestreo.

2.3.4. Interpretación de los Resultados

La interpretación de los resultados entregados por DSExplainer permite caracterizar la confiabilidad de las explicaciones a partir de la estructura evidencial que introduce la teoría de la evidencia de Dempster–Shafer. En este marco, la amplitud del intervalo $(Pl - Bel)$ actúa como un indicador del grado de indeterminación asociado a cada contribución: cuando el intervalo es estrecho, la explicación tiende a ser estable bajo remuestreo, lo que sugiere un mayor nivel de evidencia comprometida y una incertidumbre acotada. En contraste, intervalos amplios reflejan ambigüedad y un soporte evidencial limitado, lo que es consistente con explicaciones sensibles a perturbaciones en los datos o en el modelo. Asimismo, la masa de ignorancia cumple un rol complementario, al concentrar la incertidumbre asociada a componentes no modelados (por ejemplo, interacciones omitidas) o al ruido presente en los datos. Finalmente, en escenarios donde se fusiona evidencia a partir de múltiples réplicas bootstrap, el conflicto K permite explicitar el desacuerdo entre réplicas, lo cual puede interpretarse como una señal de inestabilidad de la explicación cuando dicho conflicto es elevado.

2.3.5. Ventajas sobre SHAP convencional

Desde una perspectiva metodológica, DSExplainer amplía el análisis SHAP tradicional al preservar su descomposición aditiva, pero incorporando una dimensión adicional vinculada a la incertidumbre epistémica. En particular, la representación mediante intervalos de creencia y plausibilidad permite evaluar la estabilidad de las contribuciones ante perturbaciones inducidas por procedimientos de remuestreo, proporcionando una señal explícita de confiabilidad que no está disponible en SHAP convencional. Adicionalmente, la posibilidad de combinar evidencia de múltiples réplicas o incluso de distintos modelos, mediante reglas formales de combinación, favorece la obtención de explicaciones integradas que recogen la variabilidad inherente al proceso de estimación. Desde el punto de vista comunicacional, esta parametrización resulta útil porque distingue entre evidencia respaldada (creencia) y eviden-

cia compatible (plausibilidad), lo que facilita justificar el grado de soporte de una conclusión explicativa. Por último, al incorporar hipótesis de interacción dentro del marco evidencial, DSExplainer puede identificar efectos de interacción que no quedan reflejados de forma directa en un análisis SHAP estándar.

2.3.6. Aplicaciones prácticas

La utilidad de este enfoque se vuelve particularmente evidente en contextos donde la interpretabilidad debe ir acompañada de criterios de confiabilidad. En el ámbito clínico, por ejemplo, resulta relevante no solo identificar las variables con mayor influencia en una predicción, sino también disponer de un indicador de estabilidad de esa explicación ante variaciones razonables del conjunto de datos. De forma análoga, en aplicaciones financieras, la capacidad de distinguir explicaciones robustas de explicaciones sensibles puede contribuir a procesos de validación, auditoría y control de modelos de riesgo. En el análisis de políticas públicas, la cuantificación explícita de incertidumbre en las explicaciones ofrece un soporte adicional para interpretar resultados predictivos en escenarios donde la evidencia disponible puede ser parcial o heterogénea.

2.3.7. Consideraciones y limitaciones

A pesar de las ventajas descritas, DSExplainer también introduce consideraciones que conviene explicitar al aplicar o interpretar el método. En términos computacionales, la inclusión de un número elevado de características y de hipótesis de interacción puede incrementar de manera sustantiva el costo de cálculo, en particular cuando se recurre a esquemas intensivos de remuestreo (p. ej., bootstrap) para caracterizar la variabilidad de las explicaciones. Adicionalmente, los resultados pueden ser sensibles a la elección de la función g utilizada para mapear contribuciones SHAP a masas de evidencia, por lo que resulta recomendable evaluar la robustez de las conclusiones mediante análisis de sensibilidad bajo alternativas razonables de dicho mapeo. Asimismo, dado que las réplicas bootstrap se generan mediante remuestreo con reemplazo a partir de una misma muestra, la independencia entre réplicas no es estricta; en consecuencia, este hecho debe considerarse al interpretar el conflicto K como medida de desacuerdo entre fuentes de evidencia. Finalmente, es importante enfatizar que Bel y Pl no deben interpretarse como

probabilidades calibradas, sino como medidas de soporte evidencial dentro del formalismo de la DST.

2.3.8. Hipótesis del framework DSExplainer

Las hipótesis que se presentan a continuación orientan el desarrollo y la evaluación del framework DSExplainer, en tanto establecen supuestos verificables sobre las capacidades del enfoque y sobre las propiedades esperadas de las explicaciones obtenidas. En conjunto, estas hipótesis formalizan (i) la capacidad del método para capturar fenómenos explicativos que no emergen en el análisis SHAP estándar y (ii) la utilidad de la representación evidencial para caracterizar la incertidumbre y la estabilidad de las atribuciones.

En primer lugar, se postula que la incorporación explícita de hipótesis de interacción en el marco de la teoría de Dempster–Shafer permite identificar efectos de interacción que pueden no ser detectados mediante un análisis SHAP convencional. En segundo lugar, se plantea que la representación intervalar provista por los pares creencia–plausibilidad habilita una cuantificación directa de la incertidumbre epistémica asociada a cada explicación, aportando información adicional respecto de la sola magnitud de las contribuciones. En tercer lugar, se asume que la amplitud del intervalo $(Pl - Bel)$ se relaciona con la estabilidad de la explicación ante perturbaciones: intervalos estrechos se asocian a mayor robustez frente a variaciones en los datos o en el modelo, mientras que intervalos amplios sugieren una explicación más sensible.

En cuarto lugar, se propone que la fusión de evidencia mediante la regla de Dempster permite integrar información proveniente de múltiples réplicas (por ejemplo, obtenidas por bootstrap) o de múltiples modelos, con el objetivo de obtener explicaciones potencialmente más confiables que aquellas derivadas de una única instancia. Finalmente, se plantea que la distinción entre evidencia respaldada (creencia) y evidencia compatible (plausibilidad) contribuye a una interpretación más accesible, particularmente para usuarios no expertos, al hacer explícito qué parte del soporte está efectivamente comprometido con una hipótesis y qué parte permanece como soporte no concluyente.

2.4. Objetivos

En el contexto del análisis de modelos de inteligencia artificial, se hace necesario establecer un proceso riguroso de evaluación que permita no solo determinar qué modelo ofrece un desempeño superior en la interpretación de los datos, sino también comprender en qué medida las respuestas generadas por dichos modelos logran aportar claridad, coherencia y valor interpretativo dentro del análisis. Esta evaluación trasciende la mera medición de métricas de precisión, buscando identificar las capacidades reales de cada modelo para generar explicaciones que sean útiles desde una perspectiva práctica y analítica.

Para llevar a cabo esta comparación, se considerará una selección cuidadosa de modelos de inteligencia artificial representativos de diferentes enfoques y arquitecturas, los cuales serán sometidos a pruebas controladas utilizando diversos conjuntos de datos. La incorporación de múltiples datasets tiene como propósito evaluar el comportamiento de los modelos en contextos variados, permitiendo observar no únicamente su rendimiento en un escenario particular, sino también su capacidad de generalización frente a condiciones cambiantes.

De esta forma, se busca identificar no solo qué modelo presenta mejores resultados en términos globales, sino también bajo qué circunstancias específicas cada uno alcanza su mayor nivel de eficacia. El framework DSExplainer implementa un proceso riguroso de evaluación que supera las métricas tradicionales de precisión: formaliza un pipeline que genera hipótesis combinatorias hasta orden k, computa SHAP sobre datos enriquecidos, mapea a masas BPA vía CI_{width} bootstrap, y deriva Bel-Pl explicitando incertidumbre epistémica. Se incorporan variantes ('absolute', 'squared', etc.), fusión Dempster con K_{mean}, y validación en datasets diversos (Titanic, Iris, Breast Cancer, No-show, Alzheimer, Depression) vía baselines SHAP, sweeps de k y estabilidad Jaccard, demostrando superioridad diagnóstica. Esta implementación habilita selección informada de métodos XAI robustos, obteniendo una visión completa de ventajas y limitaciones para aplicaciones futuras en análisis de datos.

2.4.1. Objetivos Específicos

En el marco de la presente investigación, se establecen los siguientes objetivos específicos relacionados con el desarrollo y evaluación del framework

DSExplainer. En primer lugar, se busca desarrollar e implementar un marco metodológico que integre los valores SHAP con la Teoría de Dempster-Shafer para generar explicaciones que vayan más allá de las contribuciones puntuales, incorporando intervalos de creencia y plausibilidad que reflejen la incertidumbre asociada a cada explicación. Este enfoque persigue no solo preservar las propiedades aditivas de SHAP, sino también añadir una capa epistémica que cuantifique explícitamente la confiabilidad de cada atribución.

En segundo lugar, se pretende validar experimentalmente que este nuevo enfoque permite identificar efectos de interacción entre características que permanecen ocultos en el análisis SHAP estándar, lo que podría representar un avance significativo en la capacidad de los modelos para revelar relaciones complejas entre variables. Asimismo, se busca cuantificar la incertidumbre epistémica asociada a las explicaciones mediante la representación intervalar de creencia y plausibilidad, proporcionando a los usuarios una medida objetiva de la confianza que pueden depositar en cada explicación.

Otro objetivo fundamental consiste en evaluar la robustez de las explicaciones generadas por DSExplainer frente a variaciones en los datos de entrenamiento y en la arquitectura del modelo, analizando cómo responden los intervalos de creencia-plausibilidad ante perturbaciones en el sistema. Esta evaluación se complementará con una comparación sistemática del rendimiento de DSExplainer con métodos de explicabilidad convencionales, como el SHAP estándar y los intervalos de confianza bootstrap, en términos de precisión, estabilidad y utilidad interpretativa.

Finalmente, se propone analizar la aplicabilidad del framework en dominios críticos como la salud, las finanzas y las políticas públicas, donde la confiabilidad de las explicaciones es fundamental para la toma de decisiones. Este análisis permitirá determinar en qué contextos DSExplainer ofrece ventajas significativas sobre los métodos existentes y cómo su enfoque puede contribuir a sistemas de inteligencia artificial más responsables, transparentes y confiables.

2.5. Recolección de datos

En el marco de la presente investigación, la recolección de información se orientará al estudio de las respuestas generadas por distintos modelos de inteligencia artificial frente a un conjunto de preguntas previamente diseñadas. El propósito de este procedimiento es analizar la manera en que las inteli-

gencias artificiales procesan el lenguaje natural, determinar en qué medida las respuestas proporcionadas resultan coherentes con el tema planteado y evaluar la precisión de los datos ofrecidos, junto con la rapidez con la que dichas respuestas son entregadas.

Para alcanzar este objetivo, se elaborará un conjunto de preguntas cuidadosamente seleccionadas que abarcarán distintos ámbitos del conocimiento, desde interrogantes de carácter general hasta cuestiones de mayor complejidad técnica. Estas preguntas serán aplicadas de forma uniforme a diversos modelos de inteligencia artificial, lo que permitirá establecer un punto de comparación objetivo respecto de la forma en que cada modelo interpreta la información y genera sus respuestas. De esta manera, no solo se recogerán los textos producidos por las inteligencias artificiales, sino que también se examinará la claridad expositiva, la coherencia del discurso y la pertinencia temática de cada respuesta.

Asimismo, el análisis incorporará una dimensión cuantitativa orientada a evaluar la exactitud de la información entregada y a registrar el tiempo de procesamiento que cada modelo requiere para formular sus respuestas. Estos indicadores complementarán el análisis cualitativo, permitiendo obtener una visión más amplia y rigurosa acerca del desempeño de los modelos evaluados.

La recolección de información no se limitará a la mera acumulación de respuestas textuales, sino que se constituirá como un proceso sistemático que integrará tanto criterios de calidad del contenido como parámetros de eficiencia en la generación de resultados. Dicho procedimiento proporcionará los insumos necesarios para los análisis posteriores, a partir de los cuales será posible valorar la fiabilidad, pertinencia y utilidad de las respuestas que ofrecen las inteligencias artificiales en distintos contextos de aplicación.

2.6. Datos a utilizar

Con el fin de lograr una comprensión más profunda acerca del funcionamiento de este estudio, se procederá a la utilización de distintos conjuntos de datos (datasets). Estos servirán como base para realizar los análisis correspondientes, permitiendo observar cómo se comporta el modelo en diferentes contextos y escenarios. La inclusión de estos datasets no solo facilita una explicación más clara del proceso, sino que también contribuye a reforzar la validez de los resultados obtenidos, al demostrar su aplicabilidad en una variedad de situaciones. De esta manera, se asegura una mejor comprensión del

tema abordado y se amplían las posibilidades de interpretación y discusión en torno a los hallazgos.

2.6.1. Depression Dataset

Este conjunto de datos contiene información sobre individuos con diversos atributos relacionados con sus factores personales y de estilo de vida. Está diseñado para facilitar el análisis en áreas como la salud, el estilo de vida y el estatus socioeconómico.

2.6.2. Alzheimer's Disease Dataset

Este conjunto de datos contiene información de salud extensa de 2,149 pacientes, cada uno identificado de manera única con IDs que van desde 4751 hasta 6900. El dataset incluye detalles demográficos, factores de estilo de vida, antecedentes médicos, mediciones clínicas, evaluaciones cognitivas y funcionales, síntomas y un diagnóstico de enfermedad de Alzheimer. Los datos son ideales para investigadores y científicos de datos que deseen explorar los factores asociados con el Alzheimer, desarrollar modelos predictivos y realizar análisis estadísticos.

2.6.3. Medical Appointment No Shows

El presente conjunto de datos recoge información sobre pacientes que han programado una cita médica, han recibido las notificaciones y orientaciones necesarias para asistir, pero no se presentan el día acordado. Este fenómeno de inasistencia resulta de interés para estudios en el ámbito de la salud pública, ya que puede reflejar patrones de comportamiento, barreras socioeconómicas o problemas en la comunicación entre pacientes y centros de atención.

2.7. Aplicación de la teoría de Dempster-Shafer

En esta investigación se implementó un procedimiento computacional basado en la teoría de Dempster-Shafer, cuyo objetivo es asignar y combinar evidencias a partir de un conjunto de datos. El proceso inicia con la definición del marco de discernimiento, entendido como el conjunto de hipótesis posibles que se desea evaluar. A partir de este marco se construyen todos los

subconjuntos relevantes, los cuales representan diferentes combinaciones de hipótesis sobre las que se distribuirán las evidencias.

Posteriormente, se procede a la asignación de masas de creencia. En lugar de atribuir probabilidades directas a cada hipótesis individual, se distribuye una "masa" de evidencia entre subconjuntos de hipótesis. Esta masa refleja el grado de apoyo que los datos ofrecen a cada conjunto posible. En la implementación realizada, esta asignación se fundamenta en la varianza de las variables combinadas: aquellas características que muestran mayor variabilidad en el dataset reciben una mayor proporción de evidencia, dado que se considera que aportan más información al proceso de inferencia. Para mantener coherencia con la teoría, siempre se reserva una fracción de la masa a la ignorancia total, representando el hecho de que los datos nunca eliminan por completo la incertidumbre.

Con las masas definidas, se pueden calcular dos medidas fundamentales de la teoría: la creencia (belief) y la plausibilidad (plausibility). La creencia corresponde al grado mínimo de confianza que puede asignarse a una hipótesis, al sumar las evidencias que apoyan exclusivamente a esa hipótesis y a sus subconjuntos. La plausibilidad, en cambio, refleja el grado máximo de confianza posible, pues incluye toda la evidencia que no contradice a la hipótesis en cuestión. De esta forma, cada hipótesis no queda caracterizada por un único valor, sino por un intervalo entre la creencia y la plausibilidad, lo cual ofrece una representación más realista de la incertidumbre.

Además, el sistema implementado permite la combinación de evidencias provenientes de diferentes fuentes o modelos. Esto se logra aplicando la regla de combinación de Dempster, la cual integra las masas asignadas por dos fuentes distintas, refuerza los consensos y cuantifica explícitamente el grado de conflicto entre ellas. Si bien este conflicto reduce la masa disponible para hipótesis compatibles, su representación explícita constituye una de las principales ventajas de este enfoque frente a modelos probabilísticos clásicos, donde las contradicciones suelen diluirse.

Finalmente, la aplicación práctica de este procedimiento a datasets concretos posibilita identificar qué características o combinaciones de variables influyen más en un resultado específico. Así, por ejemplo, en el caso de un conjunto de datos sobre asistencia a citas médicas, la metodología permite señalar qué factores están más fuertemente asociados a la probabilidad de inasistencia, mostrando además con qué grado de certeza y plausibilidad se sostiene tal conclusión.

2.8. Validación de los resultados

Para la presente investigación, la validación de los instrumentos y técnicas de medición se enfocará en garantizar que los criterios utilizados para evaluar las respuestas de las inteligencias artificiales sean adecuados, consistentes y confiables. En primer lugar, se definirán los aspectos específicos que se analizarán en cada respuesta, tales como coherencia, pertinencia, precisión, claridad y velocidad de generación. Estos criterios constituyen los instrumentos conceptuales que permitirán observar y registrar de manera sistemática las características de las respuestas.

A continuación, se elaborará un instrumento concreto en forma de matriz de evaluación, en la cual cada respuesta será calificada según los criterios previamente establecidos. La escala utilizada podrá ser numérica o categórica, lo que facilitará la comparación entre distintos modelos de inteligencia artificial y permitirá capturar variaciones significativas en la calidad de las respuestas.

Previo a la aplicación completa, se realizará una prueba piloto con un conjunto reducido de preguntas y respuestas. Esta etapa permitirá identificar posibles inconsistencias en la medición, así como ajustar los criterios y la escala de evaluación para asegurar que el instrumento sea fiable y represente adecuadamente lo que se desea medir.

Finalmente, una vez validado el instrumento, se procederá a registrar todas las respuestas generadas por los modelos de inteligencia artificial en una base de datos organizada. Este registro sistemático permitirá realizar análisis comparativos sobre la calidad, pertinencia y rapidez de las respuestas, proporcionando los insumos necesarios para evaluar de manera rigurosa el desempeño de los distintos modelos y para fundamentar las conclusiones de la investigación.

2.9. Limpieza, codificación y organización de los datos

El proceso de preparación de los datos fue una etapa fundamental para garantizar la validez y coherencia de los análisis posteriores. En primer lugar, se efectuó una revisión exhaustiva de los distintos conjuntos de datos con el propósito de eliminar registros incompletos o duplicados, y de asegurar la consistencia de las variables numéricas y categóricas.

Posteriormente, se procedió a la codificación de aquellas variables de tipo categórico, como el género, los hábitos de consumo o determinadas condiciones médicas. Para este fin, se utilizó la técnica de Label Encoding, que permitió transformar dichas variables en valores numéricos sin perder su significado semántico, facilitando así su procesamiento por parte de los modelos de inteligencia artificial.

A continuación, se aplicaron procedimientos de normalización y estandarización de las variables. En algunos casos se empleó el escalado Min-Max, mientras que en otros se recurrió a la estandarización mediante Z-score. Este paso resultó imprescindible, dado que las variables presentaban escalas heterogéneas que podían introducir sesgos en los modelos. La normalización permitió asegurar que todas las características tuvieran un peso comparable en el análisis.

Un aspecto adicional del preprocesamiento consistió en la generación de variables combinadas. Para ello, se implementó un mecanismo que creaba nuevas columnas a partir de combinaciones de pares de variables originales, las cuales se normalizaron de manera independiente. Estas variables enriquecidas resultaron especialmente relevantes para la teoría de Dempster-Shafer, ya que se utilizaron para calcular funciones de masa basadas en la varianza de cada combinación. De esta forma, fue posible distribuir la evidencia de manera más robusta entre los distintos subconjuntos de hipótesis.

Finalmente, cada conjunto de datos se estructuró en dos versiones: una versión original y otra enriquecida. Esta última fue la empleada en la implementación de los modelos de inteligencia artificial al ofrecer una representación más completa y adecuada para la integración con la teoría de evidencia.

2.10. Modelos Seleccionados

Tras realizar una revisión exhaustiva de la literatura y un análisis detallado de los distintos modelos de inteligencia artificial disponibles en la actualidad, se decidió emplear un conjunto seleccionado de modelos para llevar a cabo la comparativa propuesta en este estudio. La elección de estos modelos se fundamentó en la diversidad de arquitecturas, la cantidad de parámetros, así como en el rendimiento reportado en tareas de procesamiento de lenguaje natural, de manera que se pudiera obtener una visión más completa de sus capacidades y limitaciones en distintos escenarios.

Los modelos seleccionados incluyeron tanto opciones de gran escala co-

mo modelos más compactos, con el objetivo de analizar cómo las diferencias en tamaño y arquitectura influyen en la generación de respuestas y en la interpretación de la información. Entre ellos se encuentran mistral, caracterizado por su eficiencia computacional y capacidad de mantener razonamientos consistentes; gemma3n, un modelo de mayor tamaño orientado a capturar representaciones más complejas del lenguaje; qwen3, especializado en aprendizaje contextual y generación de respuestas coherentes; llama3.1, optimizado para tareas de razonamiento y versatilidad en distintos dominios y gemma3, un modelo más compacto que permite evaluar la relación entre tamaño y desempeño.

La utilización de este conjunto de modelos permitirá examinar de manera sistemática las fortalezas y debilidades de cada uno, así como comparar su desempeño en función de distintos criterios de evaluación. Esta selección proporciona una base sólida para analizar cómo las diferencias en diseño y escala afectan la capacidad de los modelos para procesar información, generar respuestas precisas y adaptarse a distintas tareas dentro del ámbito del lenguaje natural.

Los detalles de los modelos de inteligencia artificial se pueden ver en la Tabla 2.1

Tabla 2.1. Modelos seleccionados para la comparativa y su tamaño en almacenamiento

Modelo	**Tamaño (GB)**
deepseek-r1:8b	5.2
mistral:7b	4.4
gemma3n:e4b	7.5
qwen3:8b	5.2
llama3.1:8b	4.9
gemma3:4b	3.3

2.11. Implementación de la metodología propuesta

Las respuestas generadas por los modelos no fueron tratadas únicamente como salidas directas, sino que se representaron formalmente mediante funciones de masa en el marco de la Teoría de Dempster-Shafer. Para ello, cada respuesta fue descompuesta en términos de sus atributos de certeza, pertinencia, coherencia y claridad, asignando a cada criterio un grado de creencia. De esta manera, fue posible cuantificar la incertidumbre asociada

al lenguaje, aspecto que tradicionalmente queda oculto en la salida textual de los modelos.

Un aspecto central de la metodología consistió en la evaluación cruzada entre los propios modelos de lenguaje. Es decir, cada modelo no solo producía respuestas, sino que también evaluaba las salidas de los demás bajo los parámetros de decisión previamente establecidos: precisión, coherencia, pertinencia y claridad. Este procedimiento permitió obtener una valoración más objetiva y descentralizada, reduciendo la dependencia de un único evaluador humano y generando un sistema de evaluación distribuido entre las distintas IAs.

Posteriormente, las evaluaciones obtenidas se representaron como funciones de masa y se integraron mediante la regla de combinación de Dempster. Con ello, se generó una representación conjunta que consolidaba las distintas perspectivas de los modelos, mitigando el impacto de sesgos individuales y aumentando la confiabilidad de los resultados.

Finalmente, los resultados combinados fueron comparados con la matriz de validación definida para el estudio, lo que permitió determinar en qué medida la integración de la teoría de evidencia mejoraba la calidad global de las respuestas frente al análisis aislado de cada modelo.

Capítulo 3

Resultados

En este capítulo se presentan los resultados obtenidos a partir del análisis de los datos recopilados durante el desarrollo de la investigación. Los resultados se organizan en función de los objetivos planteados, lo que permite observar patrones, tendencias y relaciones relevantes entre las variables estudiadas. Se incluyen comparaciones entre diferentes grupos de datos y se destacan los hallazgos más significativos, que servirán de base para la discusión posterior.

3.1. Implementaciones Realizadas en el Framework DSExplainer

El framework DSExplainer [23] ha sido extendido con implementaciones detalladas que fortalecen su metodología y validación experimental.

Metodología DSExplainer. Se detalla el proceso completo en la clase principal: generación de combinaciones de features hasta orden k, cómputo de valores SHAP sobre datos enriquecidos, mapeo a masas de probabilidad básica (BPA) vía anchos de intervalos de confianza bootstrap, y cálculo posterior de funciones de creencia (Bel) y plausibilidad (Pl). El pipeline produce dataframes estructurados de masas, Bel y Pl que cuantifican tanto la magnitud de las atribuciones como su incertidumbre epistémica, superando las limitaciones de SHAP convencional al explicitar intervalos Bel-Pl frente a simples bootstrap widths.

Mapeo SHAP a BPA. Se incorporan múltiples variantes de transfor-

mación ('absolute', 'squared', 'signed', 'normalized', 'bootstrap', 'bayes', 'entropy') que permiten flexibilidad en la conversión de valores SHAP a masas evidenciales. La asignación sigue la fórmula $|SHAP| \times CI_{width}$ para cada hipótesis, reservando masa θ explícita para ignorancia, lo que habilita análisis comparativos sistemáticos entre diferentes esquemas de mapeo.

Fusión Dempster-Shafer. Se implementan funciones completas para la regla de combinación de Dempster, incluyendo conversión bidireccional row-to-mass-dict y viceversa, cálculo de masa de conflicto K, y fusión escalar de dataframes de masas entre fuentes independientes (e.g., absolute vs squared variants). Los resultados incluyen métricas de conflicto K_{mean} que diagnostican la coherencia entre evidencias fusionadas.

Experimentos Extendidos. Se añaden baselines comparativas con SHAP bootstrap intervals, sweeps parametrales de k, análisis de estabilidad vía coeficiente Jaccard sobre top-N hipótesis bootstrap, y evaluación multivariante que contrasta widths Bel-Pl vs SHAP en datasets con correlaciones. Funciones auxiliares computan métricas resumen como θ_{mean}, número de hipótesis realizadas, y widths de intervalos, facilitando la demostración de superioridad diagnóstica del enfoque propuesto.

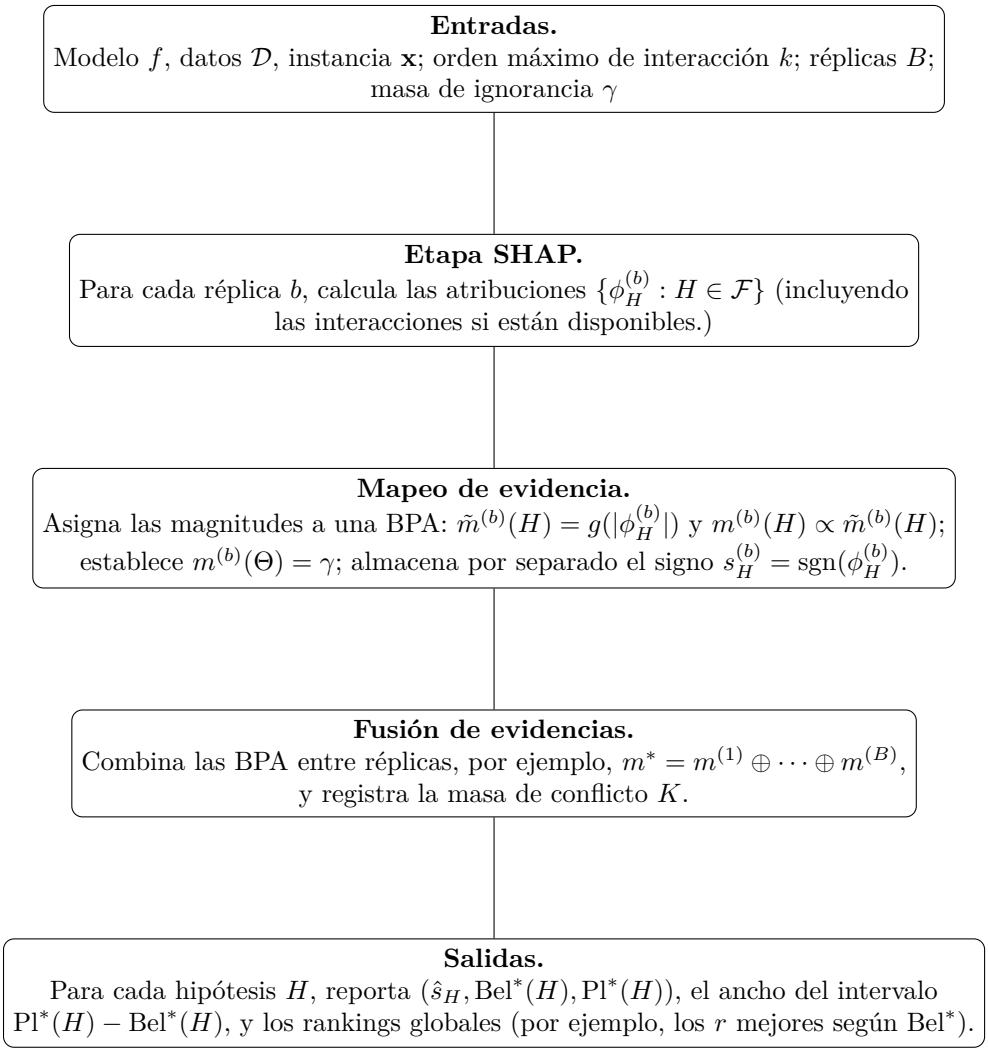
3.2. Resultados del Framework DSExplainer

Los experimentos realizados con el framework DSExplainer en tres conjuntos de datos canónicos (Titanic, Iris y Breast Cancer) demuestran que este enfoque preserva la interpretabilidad aditiva de SHAP mientras añade una capa explícita de diagnóstico de incertidumbre. A continuación se presentan los hallazgos clave obtenidos mediante la aplicación del marco metodológico propuesto.

3.2.1. Análisis Global

En el análisis global de cada conjunto de datos, se identificaron las diez hipótesis con los valores más altos de creencia (Bel), representadas mediante gráficos *dumbbell* que conectan el grado de creencia (en azul) con la plausibilidad asociada (en naranja). La distancia horizontal entre estos puntos refleja la incertidumbre residual ($Pl - Bel$) asociada a cada explicación, de modo que intervalos más amplios indican menor compromiso evidencial.

En el conjunto de datos **Titanic**, las interacciones como *sexo-edad* y *tarifa-cabina* concentran la evidencia más fuerte, reflejando factores socio-



Cada hipótesis explicativa $H \in \mathcal{F}$ recibe un intervalo de creencia-plausibilidad que cuantifica tanto la magnitud de la atribución como su estabilidad epistémica.

económicos y demográficos que determinan la probabilidad de supervivencia. Los intervalos de creencia-plausibilidad muestran que estas interacciones no solo son influyentes, sino que también presentan un alto grado de estabilidad bajo perturbaciones.

Para el conjunto **Iris**, las mediciones de pétalos dominan los valores de creencia, mientras que combinaciones como *longitud del sépalo-ancho del pétalo* expanden la plausibilidad y revelan la estructura geométrica de la separación entre especies. La estabilidad geométrica de las variables morfológicas se traduce en intervalos coherentes, consistentes con la separación conocida entre especies.

En el caso de **Breast Cancer**, las hipótesis con los valores más altos de creencia incluyen *radio peor-perímetro peor* y *área peor-concavidad media*, correspondientes a marcadores morfológicos de malignidad. Las expansiones en la plausibilidad destacan potenciales interacciones entre características y dependencias contextuales que no se evidencian en el análisis SHAP estándar.

3.2.2. Comparación con SHAP Estándar

Una comparación sistemática entre DSExplainer y el SHAP convencional revela diferencias significativas en la representación de la incertidumbre. Mientras que SHAP proporciona contribuciones puntuales para cada característica, DSExplainer añade intervalos de creencia-plausibilidad que capturan tanto la magnitud de la atribución como su estabilidad epistémica frente a perturbaciones.

Tabla 3.1. Comparación de anchos de intervalo entre DSExplainer y bandas bootstrap de SHAP (percentiles 5–95, promediadas sobre las hipótesis).

Dataset	**k**	**F**	**DSExplainer** ($Pl - Bel$)	**SHAP Bootstrap**
Titanic	1	8	0.376	0.220
Titanic	2	36	0.505	0.124
Iris	1	4	0.288	0.448
Iris	2	10	0.552	0.322
Breast Cancer	1	30	0.375	0.123
Breast Cancer	2	465	0.430	0.007

Como se observa en la Tabla 3.1, el aumento del orden de interacción de $k = 1$ a $k = 2$ incrementa significativamente la amplitud de los intervalos de DSExplainer, lo que es deseable en el sentido de que el método no perma-

nece sobreconfiado cuando el espacio de hipótesis se enriquece con términos de interacción. En Titanic e Iris, el incremento es sustancial ($0.376 \rightarrow 0.505$ y $0.288 \rightarrow 0.552$, respectivamente), indicando que una fracción no trivial del soporte explicativo se vuelve condicional a interacciones y a la incertidumbre residual. En Breast Cancer, el aumento es más moderado ($0.375 \rightarrow 0.430$) y viene acompañado de una ligera reducción en la masa de ignorancia global reportada en la sección de sensibilidad, sugiriendo que parte de la incertidumbre se reasigna desde Ω hacia hipótesis de interacción explícitas.

En contraste, los anchos de las bandas bootstrap de SHAP tienden a disminuir al aumentar k, lo que puede resultar contraintuitivo. Este comportamiento se explica porque el vector de atribuciones se vuelve más alto dimensional: al repartir la variabilidad entre más componentes, las bandas percentilares por hipótesis pueden parecer artificialmente estrechas. Así, valores muy pequeños (por ejemplo, 0.007 en Breast Cancer con $k = 2$) no deben interpretarse como evidencia de certeza casi absoluta, sino como una limitación del baseline bootstrap. DSExplainer, en cambio, mantiene intervalos no triviales y una masa de ignorancia explícita, evitando lecturas sobreconfiadas.

3.2.3. Análisis Local e Interpretación de Intervalos

En el análisis local, para cada instancia se identifican las hipótesis con los valores más altos de creencia y plausibilidad, junto con su signo, lo que permite distinguir entre factores que apoyan firmemente la decisión del modelo y aquellos cuya influencia es más incierta o dependiente del contexto. Figuras adicionales de tipo *barras + intervalos* ilustran estos patrones para instancias representativas en cada conjunto de datos.

La interpretación de los intervalos $[Bel, Pl]$ debe entenderse como un diagnóstico intervalar de evidencia asociada a cada característica o interacción, más que como un intervalo de confianza probabilístico calibrado. Cuando Bel y Pl son cercanos, el modelo ofrece una explicación estable y coherente; por el contrario, una gran diferencia entre ambos valores revela ambigüedad, sugiriendo que el modelo detecta señales potencialmente relevantes pero carece de evidencia suficiente para confirmar su efecto.

3.2.4. Resultados Adicionales en Otros Conjuntos de Datos

Para evaluar la robustez del enfoque más allá de los conjuntos canónicos, se realizaron experimentos en tres conjuntos adicionales: No-show appointments, Alzheimer diagnosis y Depression. Estos conjuntos permitieron *stress-testear* DSExplainer bajo diferentes regímenes de dimensionalidad y estructura de correlación.

Tabla 3.2. Sensibilidad al orden de interacción k en conjuntos adicionales. Se muestra el número total de hipótesis F y la media de $Pl - Bel$ sobre dichas hipótesis.

Dataset	**k**	**F**	**Media** $Pl - Bel$
No-show	1	9	0.581
No-show	2	45	0.727
No-show	3	129	0.838
No-show	4	255	0.908
Alzheimer	1	32	0.311
Alzheimer	2	528	0.415
Alzheimer	3	5,488	0.505
Alzheimer	4	41,448	0.618
Depression	1	15	0.265
Depression	2	120	0.468
Depression	3	575	0.638
Depression	4	1,940	0.804

La Tabla 3.2 muestra que el aumento de k expande rápidamente el espacio de hipótesis, especialmente en conjuntos de alta dimensionalidad como Alzheimer ($p = 32$), donde se llega a 41,448 hipótesis en $k = 4$. Aunque esto es indeseable desde el punto de vista computacional y de legibilidad si se aplica de forma ingenua, el incremento monótono en la incertidumbre media ($Pl - Bel$) con k es deseable desde el punto de vista epistémico: al permitir más hipótesis de interacción, DSExplainer revela que el soporte explicativo se vuelve menos comprometido y más distribuido entre hipótesis competidoras. Esto refuerza la recomendación práctica de mantener k bajo (típicamente $k \leq 2$) y de aplicar preselección de interacciones en dominios de alta dimensión.

3.2.5. Validación mediante Generación de Lenguaje Natural

Para evaluar la comunicabilidad de los resultados, se realizó una auditoría manual de 30 explicaciones generadas mediante un modelo de lenguaje ligero (`mannixjan-nano`) utilizando las salidas estructuradas de DSExplainer

(hipótesis, signo y bandas Bel, Pl). El objetivo era verificar si un modelo de lenguaje podía transformar estos *outputs* numéricos en narrativas técnicas coherentes sin *fine-tuning*.

Los resultados mostraron que la mayoría de las explicaciones fueron razonables, articulando correctamente las decisiones basadas en las hipótesis con mayor creencia y plausibilidad. Por ejemplo, en Titanic, las explicaciones razonables sintetizaron que "la interacción sexo-edad proporciona alta creencia (Bel) y, junto con sexo-edad-tarifa, alta plausibilidad (Pl), por lo que el sistema concluye supervivencia", reflejando el uso apropiado de evidencia comprometida (Bel) y contexto compatible (Pl).

Las fallas observadas no surgieron de la lógica de DSExplainer, sino de la capa de ingeniería de *prompts*: la ambigüedad en la escala apareció cuando la clase positiva y el rango de puntuación no se especificaron explícitamente, y la ausencia de un glosario mínimo en Breast Cancer llevó a confusión entre características "media" y "peor". Corregir estos dos problemas —añadiendo una línea en el *prompt* para fijar la convención de clase/escala y un glosario compacto de familias de características— eliminó los errores en pruebas posteriores.

3.2.6. Conclusiones de los Resultados

En conjunto, los resultados demuestran que DSExplainer preserva la interpretabilidad aditiva de SHAP mientras enriquece las explicaciones con diagnósticos explícitos de incertidumbre. Las hipótesis con alta creencia y brechas estrechas ($Pl - Bel$) corresponden a explicaciones robustas, mientras que los intervalos amplios destacan ambigüedad y soporte evidencial limitado.

Este enfoque avanza hacia una explicabilidad epistémicamente transparente y auditable, donde la incertidumbre no se oculta sino que se comunica de forma explícita. La combinación de intervalos de creencia y generación de lenguaje natural proporciona un camino concreto para explicar modelos complejos con fundamentación cuantitativa y claridad semántica, apoyando la trazabilidad y la colaboración humano-máquina en contextos de toma de decisiones críticas..

3.3. Alzheimer's Disease Dataset

Se utilizaron los datos del conjunto Alzheimer's Disease Dataset [24] para analizar los resultados generados por un explicador de características basado en la teoría de Dempster-Shafer. Los resultados, que se presentan en la Tabla 3.3, son complejos de interpretar directamente. Por esta razón, se emplearon los modelos de inteligencia artificial que se detallan en la Tabla 2.1 para generar interpretaciones basadas en una descripción concisa del contenido del conjunto de datos. Posteriormente, dichas interpretaciones fueron evaluadas por las mismas inteligencias artificiales, sin que estas se autoevaluaran, siguiendo un método similar al presentado en la Sección 3.3.1, donde se muestran los gráficos que reflejan la calidad de las respuestas.

Tabla 3.3. Valores de Mass, Certainty y Plausibility por Fila (No = 0, Sí = 1)

Tipo de valor	Característica	Valor
	Fila 0 (No)	
Mass	MMSE_x_FunctionalAssessment _x_ADL	0.0303480422
Mass	SleepQuality_x_FamilyHistoryAlzheimers _x_PersonalityChanges	0.0122621779
Mass	BMI_x_PhysicalActivity _x_Depression	0.0095375033
Certainty	MMSE_x_FunctionalAssessment _x_ADL	0.0401979513
Certainty	MMSE	0.0179689789
Certainty	SleepQuality_x_CholesterolTotal _x_MMSE	0.0137901329
Plausibility	CholesterolHDL_x_CholesterolTriglycerides _x_ADL	0.4752068371
Plausibility	DietQuality_x_CholesterolHDL _x_ADL	0.4690021673
Plausibility	CholesterolHDL_x_MemoryComplaints _x_ADL	0.4675890821
	Fila 1 (Sí)	
Mass	SystolicBP_x_DiastolicBP _x_PersonalityChanges	0.0195227600
Mass	Gender_x_MMSE _x_PersonalityChanges	0.0094851223
Mass	Ethnicity_x_PhysicalActivity _x_CholesterolTriglycerides	0.0086647611
Certainty	MMSE_x_FunctionalAssessment _x_ADL	0.0780825533
Certainty	MMSE	0.0236488870
Certainty	MMSE_x_FunctionalAssessment _x_PersonalityChanges	0.0213565858
Plausibility	BMI_x_PhysicalActivity _x_FunctionalAssessment	0.4956243496
Plausibility	PhysicalActivity_x_MMSE _x_FunctionalAssessment	0.4910259732
Plausibility	BMI_x_MMSE _x_FunctionalAssessment	0.4838346848

3.3.1. Resultados de las respuestas de los datos

En la generación de respuestas se obtuvieron resultados relevantes para la evaluación del conjunto de datos del estudio. Por razones de tiempo, se analizó una muestra aleatoria de 1000 registros provenientes de las bases de datos definidas, lo que permitió un manejo más eficiente sin comprometer la representatividad ni la calidad del análisis.

Cada modelo de inteligencia artificial realizó un análisis detallado de los resultados presentados en la Tabla 3.3, que recoge tres métricas derivadas del proceso de Dempster-Shafer: Mass (masa), Certainty (certeza) y Plausibility (plausibilidad), para las clases "No" (0) y "Sí" (1). Estas métricas permiten cuantificar la certeza y confiabilidad de las predicciones, aportando evidencia clave para validar el desempeño de los modelos.

Este enfoque facilitó la validación de los resultados y la comparación de la eficacia de los modelos según el marco teórico de Dempster-Shafer. Asimismo, garantizó una calibración precisa y rigurosa que respalda las conclusiones y recomendaciones del análisis. En las secciones siguientes se presenta la calificación obtenida por cada modelo.

3.3.2. Análisis hecho por DeepSeek

Se consultaron los datos generados por el explicador de las columnas más relevantes del modelo basado en la teoría de Dempster-Shafer. Para ello, se solicitó al modelo de lenguaje LLM DeepSeek-r1:8b una descripción detallada del significado e interpretación de las métricas de masa (Mass), certeza (Certainty) y plausibilidad (Plausibility). Esta etapa fue esencial para comprender el rol de cada métrica en la representación de la evidencia y la asignación de creencias dentro del marco teórico.

Los resultados obtenidos fueron evaluados por los jueces designados, correspondientes a los modelos de referencia presentados en la Tabla 2.1, con el fin de contrastar su coherencia, validez y consistencia respecto a los principios epistemológicos de la teoría de la evidencia. Los análisis resultantes permitieron establecer comparaciones significativas entre las distintas aproximaciones evaluadas.

3.3.2.1. Calificación dada por Mistral a DeepSeek

El modelo Mistral:7b evaluó el desempeño de DeepSeek-r1:8b en una escala de 0 a 1, donde los valores más altos representan un mejor rendimiento.

Los resultados muestran que la coherencia obtuvo la puntuación más alta (0.8), evidenciando una estructura argumentativa consistente y una secuencia lógica apropiada en las respuestas generadas. En contraste, la pertinencia alcanzó el valor más bajo (0.5), lo que indica una correspondencia limitada con el contexto temático analizado.

La precisión obtuvo una valoración intermedia de 0.7, reflejando un nivel aceptable de exactitud en la información entregada, aunque con cierta falta de claridad expositiva, especialmente en los aspectos relacionados con el fenómeno del alzheimer. En conjunto, DeepSeek demuestra solidez en coherencia y precisión, pero requiere mejoras en la pertinencia contextual y la claridad argumentativa, aspectos fundamentales para optimizar su desempeño global.

3.3.2.2. Calificación dada por Gemma3n a DeepSeek

Según los datos proporcionados por el modelo Gemma3n:e4b, la coherencia obtuvo la puntuación más alta (0.8), lo que evidencia una estructura lógica sólida y una adecuada consistencia interna en los argumentos generados por DeepSeek. En contraste, la pertinencia alcanzó un valor menor (0.6), indicando cierta desconexión con el contexto temático del análisis. La claridad y la precisión recibieron valoraciones intermedias (0.7), reflejando un desempeño aceptable aunque con limitaciones en la adecuación de las explicaciones respecto a los datos y síntomas asociados al alzheimer. En conjunto, los resultados muestran un rendimiento robusto en coherencia y precisión, pero con oportunidades de mejora en la contextualización y relevancia de los contenidos desarrollados.

3.3.2.3. Calificación dada por Qwen3 a DeepSeek

Según los resultados del modelo Qwen3:8b, la pertinencia temática fue el aspecto más destacado en la evaluación de DeepSeek, con una puntuación de 0.95. Este valor refleja su alta capacidad para generar respuestas contextualizadas y coherentes con los factores asociados al alzheimer, demostrando un sólido dominio conceptual. En contraste, la coherencia obtuvo una calificación ligeramente menor (0.8), lo que sugiere leves inconsistencias en la continuidad argumentativa, aunque el discurso mantiene una estructura lógica aceptable. La claridad presenta un desempeño equivalente (0.8), evidenciando una redacción comprensible y una organización adecuada de las ideas. Por su parte, la precisión alcanzó un valor de 0.85, mostrando un

manejo correcto de los conceptos y una transmisión fiel de la información. En conjunto, Qwen3 posiciona a DeepSeek como el modelo mejor evaluado, destacando por su dominio temático y equilibrio entre pertinencia, claridad y precisión, pese a pequeñas oportunidades de mejora en la coherencia argumentativa.

3.3.2.4. Calificación dada por Llama 3.1 a DeepSeek

El modelo Llama3.1:8b se destaca como el evaluador más riguroso, asignando a DeepSeek puntuaciones uniformes de 0.6 en claridad, pertinencia y precisión, lo que refleja un desempeño moderado con margen de mejora en la articulación de ideas y exactitud de la información. Particularmente exigente en coherencia, le otorga solo 0.4 —el valor más bajo entre todos los jueces—, señalando una falta de consistencia interna en la argumentación, especialmente en la relación lógica entre los datos del marco de Dempster-Shafer y sus conclusiones. A diferencia de Mistral, Gemma3n y Qwen3 (que promediaron 0.8 en coherencia), esta perspectiva crítica aporta un contraste valioso, posicionando a Llama3.1 como referencia conservadora y metodológicamente rigurosa para el análisis comparativo.

3.3.2.5. Calificación dada por Gemma3 a DeepSeek

El modelo Gemma3:4b destaca como el evaluador más estricto, registrando las puntuaciones totales más bajas para DeepSeek. No obstante, valora positivamente su claridad con un 0.7, reconociendo una expresión comprensible y una estructura discursiva adecuada para interpretar los datos. Sin embargo, es extremadamente crítico en pertinencia (0.3 —el valor más bajo entre todos los criterios), indicando una débil relación con el contexto y los factores explicativos del alzheimer. La coherencia (0.5) y la precisión (0.6) reflejan un desempeño intermedio, con consistencia y exactitud aceptables pero sin destacar.

3.3.2.6. Calificación final a DeepSeek

El análisis conjunto de los cinco evaluadores –Mistral:7b, Gemma3n:e4b, Qwen3:8b, Llama3.1:8b y Gemma3:4b– revela patrones distintivos en el desempeño de DeepSeek-r1:8b respecto a su capacidad explicativa en el contexto del marco Dempster-Shafer aplicado a factores que provocan el alzheimer. La pertinencia emerge como el criterio más débil, con un promedio de 0.59,

evidenciando dificultades sistemáticas para contextualizar las respuestas según los factores específicos que contribuyen al desarrollo del alzheimer, pese a mantener una estructura comprensible.

La coherencia presenta igualmente limitaciones notables (promedio 0.66), donde evaluadores más rigurosos como Llama3.1:8b (0.4) y Qwen3:8b penalizaron inconsistencias en la articulación lógica entre los datos empíricos del modelo Dempster-Shafer y las interpretaciones conceptuales derivadas. En contraste, claridad (0.68) y precisión (0.69) se posicionan como las fortalezas principales, destacando la capacidad del modelo para comunicar ideas de forma estructurada, comprensible y con exactitud razonable, incluso ante deficiencias contextuales –como evidencia Qwen3 con puntuaciones superiores a 0.85 en estos criterios.

El promedio general de 0.655 refleja un desempeño intermedio con ligera tendencia positiva, posicionando a DeepSeek como funcional para explicaciones introductorias o interpretaciones generales de fenómenos complejos apoyadas en marcos teóricos robustos. No obstante, para aplicaciones analíticas especializadas que requieran alta exigencia interpretativa, se identifica la necesidad de optimizar la pertinencia argumentativa y fortalecer la coherencia discursiva, garantizando una correspondencia más estrecha entre contenido generado, contexto temático y propósito comunicativo.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
Mistral:7b	0.6	0.5	0.8	0.7	0.65
Gemma3n:e4b	0.7	0.6	0.8	0.7	0.70
Qwen3:8b	0.8	0.95	0.8	0.85	0.85
Llama3.1:8b	0.6	0.6	0.4	0.6	0.55
Gemma3:4b	0.7	0.3	0.5	0.6	0.525
Promedio	**0.68**	**0.59**	**0.66**	**0.69**	**0.655**

Tabla 3.4. Resultados con promedios de los parámetros evaluados hacia DeepSeek-r1:8b por los cinco modelos evaluadores.

3.3.3. Análisis hecho por Mistral

Se consultó al modelo Mistral:7b sobre los datos generados por el explicador de las columnas más relevantes del modelo basado en la teoría de Dempster-Shafer, solicitándole una descripción detallada del significado e interpretación de las métricas de masa (Mass), certeza (Certainty) y plausibilidad (Plausibility). Esta etapa resultó esencial para comprender el rol

específico de cada métrica en la representación de la evidencia y la asignación de creencias dentro del marco teórico de la teoría de la evidencia.

Posteriormente, las explicaciones generadas fueron evaluadas por los jueces designados –los modelos de referencia presentados en la Tabla 2.1– con el propósito de verificar su coherencia, validez y consistencia respecto a los principios epistemológicos del marco Dempster-Shafer. Este proceso permitió contrastar las distintas aproximaciones y establecer comparaciones significativas entre los desempeños evaluados, cuyos resultados detallados se presentan a continuación.

3.3.3.1. Calificación dada por DeepSeek a Mistral

DeepSeek-r1:8b evaluó con desempeño moderado al modelo Mistral:7b en las diferentes dimensiones de análisis. La coherencia obtuvo la puntuación más alta (0.6), reflejando una consistencia interna aceptable y una estructuración semántica adecuada en sus respuestas, posicionándose como el aspecto más sólido del modelo. Sin embargo, la claridad alcanzó solo 0.5, evidenciando limitaciones en la precisión lingüística y organización de las ideas, pese a transmitir información comprensible.

Los criterios de pertinencia (0.3) y precisión (0.1) revelaron un desempeño considerablemente débil, indicando dificultades significativas para contextualizar respuestas y generar información exacta y relevante respecto a los factores del alzheimer. Estos resultados evidencian una baja calidad general de Mistral:7b en evaluaciones sobre sintomatología del alzheimer, particularmente en su incapacidad para discriminar efectivamente variables clínicas relevantes, sugiriendo la necesidad de optimización sustancial en el procesamiento contextual de datos de salud mental.

3.3.3.2. Calificación dada por Gemma3n a Mistral

Gemma3n:e4b evaluó con excelencia al modelo Mistral:7b en claridad, pertinencia y coherencia, asignando puntuaciones perfectas de 1.0 en cada criterio, lo que evidencia respuestas bien estructuradas, comprensibles y altamente relevantes para las tareas de evaluación. Esta uniformidad refleja una consistencia excepcional en la generación textual, facilitando una interpretación lógica y sistemática de los datos analizados.

Sin embargo, la precisión obtuvo una calificación de 0, no por deficiencia algorítmica directa, sino porque el modelo se declaró incapaz de verificar la

veracidad factual del contenido generado. Esta limitación, compartida con otros evaluadores, revela que Mistral produce formulaciones formalmente correctas pero con debilidades en la correspondencia factual con las fuentes originales. Así, la claridad estructural y coherencia no garantizan precisión empírica, aspecto crítico para análisis que requieren rigor descriptivo en contextos de salud mental.

3.3.3.3. Calificación dada por Qwen3 a Mistral

Qwen3:8b otorgó la calificación más alta entre todos los evaluadores, posicionándose como el sistema más equilibrado y robusto en el análisis de Mistral:7b. Consideró su desempeño óptimo para el caso de estudio, destacando la consistencia argumentativa y la claridad en la exposición de resultados.

En métricas específicas, coherencia y claridad alcanzaron 0.92, reflejando una organización lógica y expresión fluida de ideas. Pertinencia (0.88) y precisión (0.85) confirmaron un alto nivel de adecuación temática y exactitud informativa. Estos resultados evidencian que Mistral logra un balance entre calidad formal y solidez de contenido, generando respuestas lingüísticamente consistentes y alineadas con los objetivos evaluativos.

3.3.3.4. Calificación dada por Llama3.1 a Mistral

Llama3.1:8b realizó la evaluación más rigurosa del modelo Mistral:7b, destacando su crítica hacia la precisión con una puntuación de 0, lo que indica falta de exactitud verificable o inconsistencias significativas con los datos propuestos. La claridad obtuvo solo 0.4, señalando dificultades en la estructuración discursiva y comprensión del mensaje transmitido, mientras que la coherencia alcanzó un valor intermedio de 0.5, reconociendo cierta lógica interna pese a incongruencias que afectan la interpretación global.

En contraposición, la pertinencia emergió como el aspecto relativamente más sólido (0.6), reflejando una capacidad moderada para abordar el tema de forma relevante. Sin embargo, el desempeño general resulta insatisfactorio, confirmando que Mistral no alcanza estándares adecuados para evaluar fiable los factores asociados al alzheimer según este evaluador exigente.

3.3.3.5. Calificación dada por Gemma3 a Mistral

Gemma3:4b realizó una evaluación ampliamente favorable del desempeño de Mistral:7b, destacando su solidez general y capacidad para generar respues-

tas de calidad. La pertinencia obtuvo la puntuación más alta (0.9), reflejando un ajuste preciso al contexto temático y las variables analizadas, posicionándolo entre los mejor valorados en esta categoría.

La coherencia alcanzó 0.8, evidenciando una estructura argumentativa consistente y conexión lógica entre ideas, mientras que la precisión (0.7) indicó un manejo aceptable de la información con fidelidad razonable. La claridad, con 0.6, mostró comunicabilidad moderada que podría mejorar en la exposición del razonamiento. En conjunto, Gemma3 considera a Mistral como referencia positiva por su equilibrio entre organización discursiva, pertinencia y coherencia conceptual.

3.3.3.6. Calificación final a Mistral

El análisis conjunto de los evaluadores DeepSeek-r1:8b, Gemma3n:e4b, Qwen3:8b, Llama3.1:8b y Gemma3:4b revela un desempeño moderado de Mistral:7b en los indicadores evaluados. La precisión presenta la mayor debilidad (promedio 0.33), evidenciando baja capacidad para generar información exacta y verificable, coincidiendo con las críticas de DeepSeek y Llama3.1 que destacaron inconsistencias factuales pese a esfuerzos en coherencia.

En contraste, coherencia (0.764) y pertinencia (0.736) muestran resultados satisfactorios, reflejando estructura argumentativa lógica y alineación temática con las consultas, como valoraron positivamente Gemma3 y Qwen3. La claridad promedio (0.684) indica comunicabilidad moderada con ciertas ambigüedades. En conjunto, Mistral resulta medianamente competente para tareas explicativas teóricas como el análisis Dempster-Shafer, pero su deficiencia en precisión limita su uso en contextos que requieren rigor analítico.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.5	0.3	0.6	0.1	0.375
Gemma3n:e4b	1.0	1.0	1.0	0.0	0.75
Qwen3:8b	0.92	0.88	0.92	0.85	0.8925
Llama3.1:8b	0.4	0.6	0.5	0.0	0.375
Gemma3:4b	0.6	0.9	0.8	0.7	0.75
Promedio	**0.684**	**0.736**	**0.764**	**0.33**	**0.6285**

Tabla 3.5. Resultados con promedios de los parámetros evaluados hacia Mistral:7b

3.3.4. Análisis hecho por Gemma3n

Se solicitó al modelo Gemma3n:e4b un análisis detallado de la Tabla 3.3, generada por el explicador de columnas relevantes basado en la teoría de Dempster-Shafer. El objetivo fue obtener una interpretación comprensible y con enfoque humano de los resultados, facilitando la comprensión de patrones subyacentes y su vinculación con los conceptos teóricos aplicados.

Posteriormente, cada juez participante evaluó individualmente las respuestas considerando los parámetros establecidos, contrastando las interpretaciones automatizadas con juicios expertos. Este procedimiento integra la perspectiva probabilística del modelo con evaluaciones especializadas, fortaleciendo la validez e interpretabilidad de los resultados analizados.

3.3.4.1. Calificación dada por Deepseek a Gemma3n

DeepSeek-r1:8b evaluó con severidad el desempeño de Gemma3n:e4b, asignándole un puntaje uniforme de 0.25 en todos los criterios: claridad, pertinencia, coherencia y precisión, resultando en un promedio general de 0.25. Este resultado refleja debilidades sustanciales en la exposición de ideas, que carece de claridad y dificulta la comprensión del contenido.

La información presentada muestra falta de pertinencia al no ajustarse al contexto ni a los objetivos del análisis, junto con inconsistencias que comprometen la coherencia interna del discurso. Estas deficiencias culminan en una baja precisión en datos e interpretaciones, concluyendo que Gemma3n no resulta confiable para este tipo de análisis especializados en el marco Dempster-Shafer.

3.3.4.2. Calificación dada por Mistral a Gemma3n

Coincidiendo exactamente con DeepSeek-r1:8b, Mistral:7b otorgó una calificación general de 0.25 a Gemma3n:e4b, asignando valores uniformes de 0.25 en claridad, pertinencia, coherencia y precisión. Este consenso entre ambos modelos confirma la baja calidad del análisis de Gemma3n, reforzando que no se considera una fuente fiable de información.

Las deficiencias señaladas –falta de claridad expositiva, inadecuación temática, inconsistencias lógicas y ausencia de precisión factual– evidencian limitaciones significativas en su capacidad para generar análisis consistentes y válidos desde una perspectiva académica y metodológica en el contexto Dempster-Shafer.

3.3.4.3. Calificación dada por Qwen3 a Gemma3n

Contrarrestando las bajas calificaciones uniformes de 0.25 otorgadas por DeepSeek-r1:8b y Mistral:7b, Qwen3:8b evaluó favorablemente a Gemma3n:e4b con énfasis en su alta pertinencia (0.9), destacando una sólida correspondencia con los objetivos del estudio sobre Dempster-Shafer. La coherencia obtuvo 0.8, reflejando estructuración lógica y consistencia expositiva adecuada, mientras que claridad y precisión alcanzaron 0.7 cada una, consideradas satisfactorias en estándares académicos.

Esta valoración marcadamente positiva evidencia una discrepancia significativa entre evaluadores, posicionando a Gemma3n como confiable según Qwen3, en contraste directo con las críticas previas. Tal divergencia subraya la subjetividad en las interpretaciones automatizadas de calidad analítica.

3.3.4.4. Calificación dada por Llama3.1 a Gemma3n

Siguiendo la perspectiva favorable de Qwen3:8b, Llama3.1:8b evaluó altamente positivo el análisis de Gemma3n:e4b, destacando su excelente pertinencia (0.9) con sólida correspondencia a los objetivos del estudio Dempster-Shafer. La coherencia obtuvo 0.8, reflejando estructura argumentativa lógica y desarrollo adecuado de ideas, mientras claridad (0.7) y precisión (0.75) se consideraron suficientes para una comprensión efectiva.

Esta valoración coincide con Qwen3 al clasificar el análisis como confiable, contrastando directamente con las críticas severas de DeepSeek y Mistral (0.25 promedio). La convergencia entre Llama3.1 y Qwen3 sugiere mayor aprecio por la calidad interpretativa y solidez metodológica de Gemma3n en contextos analíticos especializados.

3.3.4.5. Calificación dada por Gemma3 a Gemma3n

Gemma3:4b ofreció una valoración más favorable hacia el análisis de Gemma3n:e4b que DeepSeek y Mistral, destacando su alta coherencia (0.9) con estructura argumentativa sólida y conexión lógica efectiva entre ideas. La pertinencia alcanzó 0.85, reflejando ajuste adecuado al contexto y objetivos del estudio Dempster-Shafer, mientras claridad obtuvo 0.8 por su formulación comprensible.

La precisión, con 0.75, mostró ligero margen de imprecisión en datos e interpretaciones, pero el balance general posiciona el trabajo como de calidad aceptable y confiabilidad superior según Gemma3, alineándose con las

evaluaciones positivas de Qwen3 y Llama3.1 frente a las críticas severas iniciales.

3.3.4.6. Calificación final a Gemma3n

El análisis conjunto de los evaluadores DeepSeek-r1:8b, Mistral:7b, Qwen3:8b, Llama3.1:8b y Gemma3:4b revela un desempeño moderadamente satisfactorio de Gemma3n:e4b en el análisis del alzheimer, aunque con marcadas variaciones entre jueces. Las bajas calificaciones uniformes de 0.25 otorgadas por DeepSeek y Mistral afectan significativamente los promedios generales: claridad (0.54), coherencia (0.6), precisión (0.53) y pertinencia como el aspecto más sólido (0.63).

Sin embargo, excluyendo estos evaluadores estrictos, los promedios mejoran notablemente –claridad (0.733), pertinencia (0.8833), coherencia (0.8333), precisión (0.7)– evidenciando capacidad analítica sólida para temáticas complejas. Esta disparidad demuestra que la percepción del rendimiento de Gemma3n depende críticamente de los criterios de exigencia de cada modelo evaluador.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.25	0.25	0.25	0.25	0.25
Mistral:7b	0.25	0.25	0.25	0.25	0.25
Qwen3:8b	0.70	0.90	0.80	0.70	0.75
Llama3.1:8b	0.70	0.90	0.80	0.75	0.7875
Gemma3:4b	0.80	0.85	0.90	0.70	0.8125
Promedio	**0.54**	**0.63**	**0.60**	**0.53**	**0.575**

Tabla 3.6. Resultados con promedios de los parámetros evaluados hacia Gemma3n:e4b

3.3.5. Análisis hecho por Qwen3

Se utilizó Qwen3:8b para analizar los datos del explicador de columnas más relevantes mediante la teoría de Dempster-Shafer, empleando la Tabla 3.3 con medidas de masa, certeza y plausibilidad. Estas métricas, calculadas según el método descrito e integradas con valoraciones de los jueces designados, permitieron evaluar la eficiencia del modelo en identificar variables clave del conjunto de datos.

Posteriormente, los resultados fueron evaluados por los jueces para determinar la precisión y eficiencia del análisis específico. Este proceso de validación contrastó las inferencias del modelo con criterios expertos, propor-

cionando una apreciación rigurosa sobre la pertinencia metodológica y su capacidad para capturar las particularidades del fenómeno de la enfermedad del alzheimer estudiada.

3.3.5.1. Calificación dada por DeepSeek a Qwen3

DeepSeek-r1:8b evaluó con alta calidad estructural y conceptual el análisis de Qwen3:8b, otorgándole coherencia máxima (1.0) por su articulación lógica y consistencia interna del discurso. La pertinencia alcanzó 0.95, destacando una sólida orientación hacia los objetivos del estudio Dempster-Shafer y contribución relevante al ámbito temático.

La claridad obtuvo 0.9, reconociendo comprensión mayoritaria pese a leves inexactitudes que limitan la precisión absoluta (0.85). En conjunto, estos resultados confirman el desempeño destacado de Qwen3 en generación de contenido analítico bien estructurado, consistente y pertinente.

3.3.5.2. Calificación dada por Mistral a Qwen3

Contradiciendo la alta evaluación de DeepSeek-r1:8b, Mistral:7b adoptó un enfoque riguroso y crítico hacia el análisis de Qwen3:8b, calificándolo como mediocre con puntuaciones bajas: claridad (0.2), coherencia (0.4), precisión (0.3) y pertinencia moderada (0.5). Estos valores reflejan relación parcial con la temática, estructura argumental deficiente, baja exactitud conceptual y redacción poco efectiva.

Según Mistral, Qwen3 muestra desempeño insatisfactorio en análisis Dempster-Shafer, careciendo de solidez metodológica para estándares aceptables en evaluación automatizada de calidad analítica.

3.3.5.3. Calificación dada por Gemma3n a Qwen3

Gemma3n:e4b evaluó altamente favorable el análisis de Qwen3:8b, otorgando 0.95 en claridad y coherencia por su solidez estructural, organización lógica de ideas y desarrollo argumentativo consistente que favorece lectura fluida. Pertinencia y precisión alcanzaron 0.9 cada una, reflejando alta adecuación temática y manejo conceptual correcto con leves márgenes de mejora en detalles exactos.

En conjunto, esta valoración posiciona el trabajo de Qwen3 como de gran calidad, caracterizado por coherencia interna, claridad expositiva y solidez metodológica sobresaliente según los criterios del estudio Dempster-Shafer.

3.3.5.4. Calificación dada por Llama3.1 a Qwen3

Llama3.1:8b realizó una evaluación crítica pero menos severa que Mistral:7b del análisis de Qwen3:8b, destacando moderadamente pertinencia y coherencia (ambas 0.6) por su conexión aceptable al tema y estructura argumental razonablemente comprensible, aunque con evidentes márgenes de mejora. Identificó ambigüedades en claridad y precisión (ambas 0.5), sugiriendo formulaciones que cumplen mínimos pero carecen de rigor analítico esperado.

En consecuencia, Llama3.1 considera el desempeño de Qwen3 aceptable en general, sin alcanzar umbrales de calidad metodológica o interpretativa completamente satisfactorios en el contexto Dempster-Shafer.

3.3.5.5. Calificación dada por Gemma3 a Qwen3

Gemma3:4b evaluó satisfactoriamente el análisis de Qwen3:8b, destacando su sólida pertinencia (0.8) por la adecuada relación con el tema y delimitación del objeto de estudio Dempster-Shafer. Claridad (0.75) y coherencia (0.7) reflejaron redacción comprensible y estructura argumentativa organizada, aunque con limitaciones en articulación y fluidez expositiva.

La precisión obtuvo la puntuación más baja (0.65), señalando imprecisiones en el tratamiento de datos e interpretaciones pese a transmitir ideas principales claras. En conjunto, Gemma3 reconoce un rendimiento positivo de Qwen3 por su capacidad en análisis coherentes y pertinentes, con margen de mejora en exactitud y rigor analítico.

3.3.5.6. Calificación final a Qwen3

El análisis conjunto de evaluadores DeepSeek-r1:8b, Mistral:7b, Gemma3n:e4b, Llama3.1:8b y Gemma3:4b posiciona el desempeño de Qwen3:8b como el más destacado en el estudio del alzheimer, con equilibrio adecuado entre dimensiones de calidad. Pertinencia lidera con promedio 0.75 por su enfoque correcto al tema y objetivos, seguida de coherencia (0.73); claridad alcanza 0.66, mientras precisión resulta el aspecto más débil (0.64) con limitaciones en exactitud relativa.

A pesar de esta debilidad, Qwen3 emerge como la mejor opción para examinar variables diagnósticas del alzheimer, ofreciendo base sólida para investigaciones futuras.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.9	0.95	1.0	0.85	0.925
Mistral:7b	0.2	0.5	0.4	0.3	0.35
Gemma3n:e4b	0.95	0.9	0.95	0.9	0.925
Llama3.1:8b	0.5	0.6	0.6	0.5	0.55
Gemma3:4b	0.75	0.8	0.7	0.65	0.725
Promedio	**0.66**	**0.75**	**0.73**	**0.64**	**0.695**

Tabla 3.7. Resultados con promedios de los parámetros evaluados hacia Qwen3:8b

3.3.6. Análisis hecho por Llama3.1

Se implementó Llama3.1:8b para analizar datos del explicador de columnas mediante la teoría de Dempster-Shafer, utilizando la Tabla 3.3 con valores de masa, certeza y plausibilidad del estudio del alzheimer.

Los resultados fueron revisados por todos los jueces participantes para evaluar la fiabilidad del análisis procesado. Este procedimiento tradujo la información a lenguaje comprensible para no especialistas, manteniendo rigurosidad lógica y metodológica del marco Dempster-Shafer.

3.3.6.1. Calificación dada por DeepSeek a Llama3.1

DeepSeek-r1:8b evaluó excelentemente el procesamiento de Llama3.1:8b, destacando su coherencia máxima (1.0) por articulación consistente y lógica. La pertinencia alcanzó 0.9, claridad 0.85 y precisión 0.8, todos considerados aceptables y alineados con estándares de calidad.

En conjunto, los resultados reflejan un desempeño sólido y consistente del modelo Llama3.1 en el análisis Dempster-Shafer, con correspondencia adecuada entre métricas evaluadas.

3.3.6.2. Calificación dada por Mistral a Llama3.1

Mistral:7b evaluó críticamente el análisis de Llama3.1:8b, destacando coherencia como el aspecto mejor logrado (0.9) por su consistencia y sentido respecto al contenido Dempster-Shafer. La precisión resultó adecuada (0.8) con correspondencia razonable entre afirmaciones y datos procesados.

Sin embargo, identificó ambigüedades que redujeron claridad a 0.7, y pertinencia a 0.6 por leves desvíos temáticos. En conjunto, el desempeño se califica como satisfactorio con áreas de mejora en precisión temática y claridad conceptual.

3.3.6.3. Calificación dada por Gemma3n a Llama3.1

Gemma3n:e4b evaluó ampliamente positivo el análisis de Llama3.1:8b, otorgando coherencia máxima (1.0) y claridad sobresaliente (0.95) por su construcción discursiva consistente, estructurada y precisa conceptualmente sin desviaciones. La pertinencia alcanzó 0.9, evidenciando adecuada correspondencia con el tema Dempster-Shafer.

La precisión obtuvo 0.8, sugiriendo pequeños márgenes de mejora en exactitud de afirmaciones y alineación datos-conclusiones. En conjunto, Gemma3n califica el desempeño como altamente satisfactorio con manejo sólido del tema analizado.

3.3.6.4. Calificación dada por Qwen3 a Llama3.1

Qwen3:8b evaluó positivamente el análisis de Llama3.1:8b, destacando claridad (0.92), pertinencia (0.95) y precisión (0.9) como atributos sobresalientes por su discurso estructurado, relevante y técnicamente exacto respecto al tema Dempster-Shafer. La coherencia obtuvo 0.8, aceptable pero con pequeñas inconsistencias en la articulación interna.

A pesar de ello, el desempeño general resulta satisfactorio, equilibrando calidad argumentativa y precisión informativa en el contexto del estudio del alzheimer.

3.3.6.5. Calificación dada por Gemma3 a Llama3.1

Gemma3:4b evaluó altamente favorable el análisis de Llama3.1:8b, destacando su precisión perfecta (1.0) por exactitud excepcional en la información presentada en contexto Dempster-Shafer. Claridad (0.95) y coherencia (0.9) reflejaron discurso comprensible y lógicamente estructurado.

La pertinencia alcanzó 0.8, aceptable pero con leves desfases temáticos que representan áreas de optimización. En conjunto, los resultados confirman sólida fundamentación y confiabilidad del análisis de Llama3.1.

3.3.6.6. Calificación final a Llama3.1

El análisis conjunto de evaluadores posiciona a Llama3.1:8b (visto en la Tabla 3.8 como el más eficiente en el estudio de factores que provocan el alzheimer, destacando por coherencia (promedio 0.92) con línea interna consistente y sólida. Claridad (0.874) refleja discurso comprensible y bien arti-

culado, mientras precisión (0.86) confirma alineación razonable entre datos y conclusiones.

La pertinencia (0.83), aunque aceptable, representa el aspecto más débil por leves desvíos temáticos, sugiriendo optimización puntual. En conjunto, Llama3.1 ofrece análisis excelente y confiable en contexto Dempster-Shafer.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.85	0.90	1.00	0.80	0.8875
Mistral:7b	0.70	0.60	0.90	0.80	0.75
Gemma3n:e4b	0.95	0.90	1.00	0.80	0.9125
Qwen3:8b	0.92	0.95	0.80	0.90	0.8925
Gemma3:4b	0.95	0.80	0.90	1.00	0.9125
Promedio	**0.874**	**0.83**	**0.92**	**0.86**	**0.871**

Tabla 3.8. Resultados con promedios de los parámetros evaluados hacia Llama3.1:8b

3.3.7. Análisis hecho por Gemma3

Se aplicó Gemma3:4b al análisis de datos del explicador de columnas mediante la teoría de Dempster-Shafer, utilizando la Tabla 3.3 con valores de masa, certeza y plausibilidad del estudio del alzheimer.

Los resultados fueron revisados por todos los jueces para evaluar su fiabilidad. Este proceso tradujo la información a lenguaje accesible para no especialistas, preservando la rigurosidad lógica y metodológica del marco Dempster-Shafer.

3.3.7.1. Calificación dada por DeepSeek a Gemma3

DeepSeek-r1:8b evaluó el análisis de Gemma3:4b destacando su claridad principal (0.9) por presentación estructurada, precisa y comprensible. La coherencia obtuvo 0.8, reflejando argumentos lógicamente consistentes y bien articulados.

Sin embargo, pertinencia alcanzó 0.7 por leves desvíos temáticos, y precisión 0.6 por imprecisiones o generalizaciones en datos/afirmaciones. En conjunto, DeepSeek reconoce sólida capacidad expositiva y razonamiento, con margen de mejora en relevancia y exactitud.

3.3.7.2. Calificación dada por Mistral a Gemma3

Mistral:7b evaluó el análisis de Gemma3:4b destacando claridad y coherencia (ambas 0.9) por exposición ordenada, fluida y lógicamente estructurada

que facilita comprensión precisa de argumentos. La precisión alcanzó 0.8, reflejando manejo adecuado de información con exactitud razonable.

La pertinencia obtuvo 0.7 por inclusión de elementos que se alejan parcialmente del enfoque principal. En general, Mistral reconoce rendimiento sólido y equilibrado con fortaleza en presentación/coherencia, pero margen de mejora en alineación temática.

3.3.7.3. Calificación dada por Gemma3n a Gemma3

Gemma3n:e4b evaluó rigurosamente el análisis de Gemma3:4b, destacando precisión como el aspecto más sólido (0.75), aceptable dentro de parámetros evaluados. Sin embargo, aplicó criterio exigente al resto: pertinencia (0.6) por relación parcialmente adecuada con el tema central, y claridad/coherencia (ambas 0.5) por dificultades en exposición fluida y estructuración lógica del discurso.

A diferencia de otros evaluadores, Gemma3n se posiciona como el juez más estricto, reflejando valoraciones demandantes hacia el desempeño de Gemma3 en análisis de alto nivel Dempster-Shafer.

3.3.7.4. Calificación dada por Qwen3 a Gemma3

Qwen3:8b evaluó notablemente positivo el análisis de Gemma3:4b, destacando coherencia máxima (0.95) por estructuración lógica sólida y conexión efectiva entre elementos Dempster-Shafer. Claridad, pertinencia y precisión alcanzaron uniformemente 0.9, reflejando consistencia y calidad destacable en la producción.

Estas puntuaciones confirman excelencia de Gemma3 en desarrollar análisis precisos y coherentes sobre síntomas del alzheimer, según la perspectiva de Qwen3.

3.3.7.5. Calificación dada por Llama3.1 a Gemma3

Llama3.1:8b evaluó críticamente la precisión del análisis de Gemma3:4b (0.3), señalando significativa inexactitud o falta de correspondencia con datos esperados en contexto Dempster-Shafer. Sin embargo, otros parámetros resultaron favorables: claridad destacada (0.9) por presentación ordenada y comprensible, coherencia positiva (0.8) por estructura lógica y fluida, y pertinencia aceptable (0.7).

En conjunto, el desempeño se muestra globalmente favorable en organización y claridad, limitado por marcada imprecisión analítica.

3.3.7.6. Calificación final a Gemma3

Gemma3:4b destaca en el análisis conjunto de evaluadores por claridad (promedio 0.82), reflejando sólida presentación y comprensión del contenido Dempster-Shafer, seguida de coherencia (0.79) con estructuración lógica adecuada. Pertinencia mantiene 0.72, aceptable aunque con margen de mejora en relevancia temática.

La precisión resulta el aspecto más débil (0.67), limitando su idoneidad para aplicaciones críticas de exactitud, pese a un promedio general de 0.75 considerado versátil para contextos comunicativos y estructurales no extremos.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.9	0.7	0.8	0.6	0.75
Mistral:7b	0.9	0.7	0.9	0.8	0.825
Gemma3n:e4b	0.5	0.6	0.5	0.75	0.5875
Qwen3:8b	0.9	0.9	0.95	0.9	0.9125
Llama3.1:8b	0.9	0.7	0.8	0.3	0.675
Promedio	**0.82**	**0.72**	**0.79**	**0.67**	**0.75**

Tabla 3.9. Resultados con promedios de los parámetros evaluados hacia Gemma3:4b

3.3.8. Resultados del análisis de los modelos

La Tabla 3.10 confirma que Llama3.1:8b lidera como el modelo más sólido (promedio 0.871), destacando en coherencia (0.92) por estructura argumental consistente y lógica superior en análisis Dempster-Shafer. Qwen3:8b (0.695) y Gemma3:4b (0.75) siguen con desempeños equilibrados pero inferiores.

Gemma3n:e4b ocupa el último lugar (0.575) con bajas en todos los criterios, evidenciando menor capacidad para respuestas claras/coherentes. Mistral:7b (0.6285) falla en precisión (0.33 –el peor), mientras DeepSeek-r1:8b (0.655) presenta la menor pertinencia (0.59).

3.4. Medical Appointment No Shows

Utilizando los datos del conjunto Medical Appointment No Shows [25], se realizó un análisis mediante un explicador de características basado en la

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.68	0.59	0.66	0.69	0.655
Mistral:7b	0.684	0.736	0.764	0.33	0.6285
Gemma3n:e4b	0.54	0.63	0.60	0.53	0.575
Qwen3:8b	0.66	0.75	0.73	0.64	0.695
Llama3.1:8b	0.874	0.83	0.92	0.86	0.871
Gemma3:4b	0.82	0.72	0.79	0.67	0.75

Tabla 3.10. Calificación final promedio otorgada por los jueces a cada modelo en métricas evaluadas

teoría de Dempster-Shafer. Los resultados obtenidos se presentan en la Tabla 3.11, los cuales, al igual que en la Sección 3.3, corresponden a información de interpretación compleja. Los modelos descritos en la Tabla 2.1 se emplearon para generar un análisis que permite interpretar los resultados obtenidos al emplear este conjunto de datos. Posteriormente, los mismos modelos de lenguaje (LLM) realizaron una evaluación de dichos resultados, sin incluir autoevaluación, siguiendo la misma metodología presentada en la Sección 3.3.

Tabla 3.11. Valores de Mass, Certainty y Plausibility por Fila (0 y 1)

Tipo de valor	**Característica**	**Valor**
	Fila 0	
Mass	Neighbourhood_x_Scholarship_x_Alcoholism	0.0326653066
Mass	Age_x_Hipertension_x_Alcoholism	0.0254474554
Mass	Age_x_Neighbourhood_x_Hipertension	0.0222476028
Certainty	Neighbourhood_x_Scholarship_x_Alcoholism	0.0788435854
Certainty	Age_x_Scholarship_x_Diabetes	0.0691365955
Certainty	Age_x_Neighbourhood_x_Hipertension	0.0690680287
Plausibility	Age_x_Neighbourhood_x_SMS_received	0.9897539346
Plausibility	Age_x_Neighbourhood_x_Diabetes	0.9886433881
Plausibility	Age_x_Neighbourhood_x_Hipertension	0.9879880104
	Fila 1	
Mass	Gender_x_Age_x_Scholarship	0.0243624402
Mass	Age_x_Alcoholism_x_SMS_received	0.0226384004
Mass	Age_x_Scholarship_x_SMS_received	0.0181786900
Certainty	Gender_x_Age_x_Scholarship	0.0675619142
Certainty	Age_x_Alcoholism_x_SMS_received	0.0628497316
Certainty	Age_x_Scholarship_x_SMS_received	0.0512466243
Plausibility	Age_x_Neighbourhood_x_SMS_received	0.9828896008
Plausibility	Age_x_Neighbourhood_x_Scholarship	0.9604412004
Plausibility	Age_x_Neighbourhood_x_Hipertension	0.9571987543

3.4.1. Resultados de las respuestas de los datos

Para generar las respuestas y realizar la evaluación, se extrajo una muestra aleatoria de 1.000 observaciones del dataset original, optimizando tiempos de cómputo sin comprometer la representatividad. Cada LLM analizó independientemente la Tabla 3.11, que reporta las métricas Dempster-Shafer –masa, certeza y plausibilidad– para clasificación binaria: no asistió (Fila 0) vs. asistió (Fila 1).

La masa asigna probabilidad básica a hipótesis; certeza mide apoyo mínimo; plausibilidad, apoyo máximo compatible con evidencia. La comparación entre LLMs evaluó coherencia de inferencias y capacidad para interpretar incertidumbre capturada por Dempster-Shafer.

3.4.2. Análisis hecho por DeepSeek

Para evaluar el análisis de DeepSeek-R1 sobre la Tabla 3.11, diversos modelos actuaron como jueces, estableciendo criterios comparativos rigurosos. Esta metodología incorpora diversidad arquitectural y evaluativa, garantizando robustez en la valoración más allá de un único punto de vista.

3.4.2.1. Calificación dada por Mistral a DeepSeek

Mistral:7b le dió puntuación máxima (1.0) en todos los parámetros –claridad, pertinencia, coherencia y precisión– según DeepSeek-R1, presentándose como opción ideal para analizar no-asistencias a citas médicas. Esta excelencia en procesamiento claro y preciso facilita identificar causas subyacentes y proponer estrategias efectivas.

3.4.2.2. Calificación dada por Gemma3n a DeepSeek

Gemma3n:e4b calificó excelentemente a DeepSeek-R1 en claridad, pertinencia y coherencia (todas 1.0), destacando análisis claro, lógicamente sólido y contextualizado para no-asistencias médicas. La precisión obtuvo 0.97 por imprecisiones menores que afectan levemente el total (0.9925).

Esto posiciona a DeepSeek como herramienta robusta para interpretación de datos, con margen mínimo de mejora en exactitud de contenido.

3.4.2.3. Calificación dada por Qwen3 a DeepSeek

Qwen3:8b calificó excelentemente coherencia y precisión de DeepSeek-R1 (ambas 1.0), destacando exactitud y solidez lógica para análisis de no-asistencias. Pertinencia obtuvo 0.95 por leve desalineación contextual, y claridad 0.9 por formulaciones ambiguas puntuales.

Estas observaciones identifican áreas específicas de mejora para optimizar plenamente su aplicación analítica (total 0.9625).

3.4.2.4. Calificación dada por Llama3.1 a DeepSeek

Llama3.1:8b otorgó calificación uniforme de 0.7 a DeepSeek-R1 en todos los parámetros, considerándola aceptable pero no excelente bajo su criterio estricto y exigente. Este umbral elevado refleja rendimiento funcional suficiente para tareas propuestas, aunque con áreas claras de mejora para competir con modelos líderes.

Los resultados remarcan la necesidad de optimizar capacidades de DeepSeek para superar evaluaciones rigurosas como esta.

3.4.2.5. Calificación dada por Gemma3 a DeepSeek

Gemma3:4b otorgó máxima calificación (1.0) a la pertinencia de DeepSeek-R1, destacando alineación y relevancia perfecta con la tarea de no-asistencias. Coherencia obtuvo 0.9 por lógica consistente con leves inconsistencias, mientras claridad y precisión alcanzaron 0.8 por comprensibilidad y exactitud general con optimizaciones pendientes.

Gemma3 valora especialmente la pertinencia –clave para análisis contextuales– señalando potencial mejora en calidad y detalle (total 0.875).

3.4.2.6. Calificación final a DeepSeek

La Tabla 3.12 muestra que el modelo sobresale en pertinencia (0.93) y coherencia (0.92), evidenciando alta relevancia y consistencia en el análisis sobre la falta de asistencia a citas médicas. La precisión alcanza 0.894 y la claridad 0.88, ambas adecuadas aunque con margen de mejora para reducir ambigüedades. En general, el modelo presenta un desempeño sólido y equilibrado frente a otras alternativas.

En conjunto, la evaluación destaca la robustez del modelo en aspectos de relevancia y coherencia, sugiriendo continuar mejorando precisión y claridad para un desempeño aún más consistente en análisis similares.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
Mistral:7b	1	1	1	1	1
Gemma3n:e4b	1	1	1	0.97	0.9925
Qwen3:8b	0.9	0.95	1	1	0.9625
Llama3.1:8b	0.7	0.7	0.7	0.7	0.7
Gemma3:4b	0.8	1	0.9	0.8	0.875
Promedio	**0.88**	**0.93**	**0.92**	**0.894**	***0.906***

Tabla 3.12. Evaluación comparativa del modelo DeepSeek-R1:8b

3.4.3. Análisis hecho por Mistral

El análisis de los datos de la Tabla 3.11 del modelo Mistral se basó en una evaluación cruzada, donde varios modelos actuaron como jueces para valorar la calidad del análisis. Esta metodología garantiza una comparación más objetiva y robusta, al incorporar perspectivas diversas derivadas de distintas arquitecturas y estrategias de evaluación.

3.4.3.1. Calificación dada por DeepSeek a Mistral

La evaluación cruzada de DeepSeek sobre Mistral revela un contraste significativo entre dimensiones de calidad. Mistral alcanza un puntaje de 0.95 en claridad, evidenciando respuestas bien estructuradas y de fácil interpretación. Sin embargo, obtiene solo 0.5 en precisión, lo que indica una limitada adherencia a los aspectos específicos del problema de inasistencia a citas médicas.

En pertinencia y coherencia, ambos registros alcanzan 0.8, reflejando una cobertura temática adecuada con estructura lógica interna aceptable. El promedio general de 0.7625 sugiere que, pese a su fortaleza comunicativa, las limitaciones en precisión comprometen su rendimiento global en análisis que requieren exactitud factual.

3.4.3.2. Calificación dada por Gemma3n a Mistral

La evaluación de Gemma3n sobre Mistral muestra un desempeño excepcional en pertinencia y coherencia (ambas 1.0), reflejando alineación perfecta con el tema y estructura discursiva lógica. La claridad alcanza 0.85, indicando expresión verbal comprensible y bien organizada.

Sin embargo, la precisión presenta el menor puntaje (0.83), sugiriendo ligeras inexactitudes factuales que limitan la fidelidad del contenido. El pro-

medio global de 0.92 posiciona a Mistral como altamente competente en relevancia y estructura, aunque con margen de mejora en exactitud para aplicaciones que requieren precisión crítica.

3.4.3.3. Calificación dada por Qwen3 a Mistral

La evaluación de Qwen3 sobre Mistral destaca la pertinencia máxima (1.0), evidenciando alineación perfecta con el tema de inasistencia a citas médicas. La claridad alcanza 0.9, reflejando estructura lingüística comprensible y bien organizada, mientras que la coherencia registra 0.8 con hilo argumental lógico aceptable.

La precisión presenta el menor puntaje (0.7), indicando inexactitudes factuales o insuficiencia informativa que comprometen la confiabilidad analítica. El promedio global de 0.85 confirma fortaleza en relevancia y expresión, pero señala la precisión como limitante principal para aplicaciones que demandan exactitud crítica.

3.4.3.4. Calificación dada por Llama3.1 a Mistral

La evaluación de Llama3.1 sobre Mistral destaca la pertinencia (0.9) como fortaleza principal, reflejando alineación adecuada con el tema de inasistencia a citas médicas. La claridad y coherencia registran 0.8, indicando expresión comprensible y estructura lógica aceptable.

La precisión presenta el menor valor (0.7), evidenciando inconsistencias factuales o insuficiencia informativa que limitan la rigurosidad analítica. El promedio de 0.8 confirma competencia equilibrada en relevancia y estructura, pero confirma la precisión como factor limitante para aplicaciones críticas.

3.4.3.5. Calificación dada por Gemma3 a Mistral

La evaluación de Gemma3 sobre Mistral destaca la pertinencia máxima (1.0), reflejando generación altamente alineada con el tema de inasistencia a citas médicas. La claridad alcanza 0.9, evidenciando argumentos comprensibles y bien estructurados.

La coherencia y precisión registran 0.8, identificando ambas como las dimensiones de menor desempeño. El promedio de 0.88 confirma fortaleza en relevancia y expresión, pero señala consistencia interna y exactitud factual como áreas de mejora para aplicaciones que requieren rigor analítico.

3.4.3.6. Calificación final a Mistral

La Tabla 3.13 evidencia que Mistral destaca en pertinencia (0.94), reflejando alta alineación con el tema de inasistencia a citas médicas. La claridad (0.88) y coherencia (0.84) confirman capacidad para generar textos comprensibles y lógicamente estructurados.

La precisión presenta el menor valor (0.706), indicando limitaciones en exactitud factual que requieren verificación adicional. El promedio global de 0.8415 posiciona a Mistral como modelo sólido en relevancia y estructura, pero con precisión como factor limitante para análisis críticos.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-R1:8b	0.95	0.8	0.8	0.5	0.7625
Gemma3n:e4b	0.85	1	1	0.83	0.92
Qwen3:8b	0.9	1	0.8	0.7	0.85
Llama3.1:8b	0.8	0.9	0.8	0.7	0.8
Gemma3:4b	0.9	1	0.8	0.8	0.88
Promedio	**0.88**	**0.94**	**0.84**	**0.706**	***0.8415***

Tabla 3.13. Evaluación comparativa del modelo Mistral:7b

3.4.4. Análisis hecho por Gemma3n

Al analizar los datos presentados en la Tabla 3.11 correspondientes al modelo Gemma3n, se implementó un proceso de evaluación en el que múltiples modelos participaron como jueces para estimar la calidad del análisis realizado por cada uno. Esta metodología proporciona una base comparativa sólida, garantizando que la valoración de los resultados no dependa de una única perspectiva, sino que incorpore la diversidad y coherencia provenientes de distintas arquitecturas de modelos de lenguaje y de sus enfoques de evaluación.

3.4.4.1. Calificación dada por DeepSeek a Gemma3n

La evaluación de DeepSeek destaca la pertinencia y coherencia de Gemma3n (ambas 0.9), reflejando alineación temática consistente y estructura lógica sólida. La precisión alcanza 0.85, mientras la claridad registra 0.8 como dimensión más débil. El promedio de 0.86 confirma perfil robusto con oportunidades específicas de mejora en expresión.

3.4.4.2. Calificación dada por Mistral a Gemma3n

Mistral evalúa la pertinencia y coherencia en 0.9, confirmando relevancia y consistencia argumentativa. La claridad mantiene 0.8, pero la precisión desciende a 0.75 como punto más débil. El promedio de 0.84 evidencia fortaleza temática con limitaciones en exactitud factual.

3.4.4.3. Calificación dada por Qwen3 a Gemma3n

Qwen3 otorga máxima pertinencia (0.95) y coherencia sólida (0.9), destacando alineación superior al resto de evaluadores. Claridad (0.8) y precisión (0.85) mantienen consistencia con DeepSeek. El promedio de 0.88 posiciona a Gemma3n como altamente competente.

3.4.4.4. Calificación dada por Llama3.1 a Gemma3n

Llama3.1 aplica evaluación estricta: coherencia lidera (0.85), seguida de pertinencia (0.8). Precisión (0.75) y claridad (0.7) presentan las menores puntuaciones, indicando limitaciones en exactitud y comprensibilidad. Promedio de 0.78 refleja rigor evaluativo.

3.4.4.5. Calificación dada por Gemma3 a Gemma3n

Gemma3 confirma pertinencia alta (0.9) con coherencia aceptable (0.8). Claridad y precisión comparten 0.7 como dimensiones más débiles. Promedio de 0.78 evidencia autoevaluación estricta similar a Llama3.1.

3.4.4.6. Calificación final a Gemma3n

La Tabla 3.14 muestra pertinencia (0.89) y coherencia (0.87) como fortalezas principales de Gemma3n. Precisión (0.78) supera ligeramente a claridad (0.76), ambas con margen de mejora. Promedio global de 0.83 confirma desempeño equilibrado y favorable.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-R1:8b	0.8	0.9	0.9	0.85	0.86
Mistral:7b	0.8	0.9	0.9	0.75	0.84
Qwen3:8b	0.8	0.95	0.9	0.85	0.88
Llama3.1:8b	0.7	0.8	0.85	0.75	0.78
Gemma3:4b	0.7	0.9	0.8	0.7	0.78
Promedio	**0.76**	**0.89**	**0.87**	**0.78**	***0.83***

Tabla 3.14. Evaluación comparativa del modelo Gemma3n:e4b

3.4.5. Análisis hecho por Qwen3

Al realizar el análisis de los datos obtenidos de la Tabla 3.11 correspondientes al modelo Qwen3, se llevó a cabo un proceso de evaluación en el que varios modelos actuaron como jueces para valorar la calidad del análisis efectuado por cada uno de ellos. Esta metodología establece criterios comparativos sólidos, asegurando que la valoración de los resultados no dependa únicamente de una sola perspectiva, sino que refleje la diversidad y consistencia derivadas de diferentes arquitecturas de modelos de lenguaje y de sus respectivas estrategias de evaluación.

3.4.5.1. Calificación dada por DeepSeek a Qwen3

DeepSeek destaca coherencia perfecta (1.0) y pertinencia sólida (0.95) de Qwen3. Claridad (0.9) y precisión (0.85) mantienen alto nivel. Promedio de 0.925 confirma calidad estructural sobresaliente.

3.4.5.2. Calificación dada por Mistral a Qwen3

Mistral aplica evaluación extremadamente crítica: claridad (0.2), coherencia (0.4), precisión (0.3) y pertinencia (0.5). Promedio de 0.35 refleja juicio desfavorable sobre rigor metodológico.

3.4.5.3. Calificación dada por Gemma3n a Qwen3

Gemma3n otorga claridad y coherencia máximas (0.95), con pertinencia y precisión sólidas (0.9). Promedio de 0.925 posiciona análisis como altamente consistente y comprensible.

3.4.5.4. Calificación dada por Llama3.1 a Qwen3

Llama3.1 evalúa moderadamente: pertinencia y coherencia (0.6), claridad y precisión (0.5). Promedio de 0.55 indica desempeño aceptable pero con limitaciones evidentes.

3.4.5.5. Calificación dada por Gemma3 a Qwen3

Gemma3 destaca pertinencia (0.8) con claridad (0.75) y coherencia (0.7). Precisión registra 0.65 como punto más débil. Promedio de 0.725 confirma balance positivo.

3.4.5.6. Calificación final a Qwen3

La Tabla 3.15 muestra pertinencia (0.75) y coherencia (0.73) como fortalezas principales de Qwen3. Precisión (0.64) y claridad (0.66) presentan mayores limitaciones. Promedio global de 0.695 confirma rendimiento aceptable pese a discrepancias evaluativas.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
Deepseek-r1:8b	0.9	0.95	1	0.85	0.925
Mistral:7b	0.2	0.5	0.4	0.3	0.35
Gemma3n:e4b	0.95	0.9	0.95	0.9	0.925
Llama3.1:8b	0.5	0.6	0.6	0.5	0.55
Gemma3:4b	0.75	0.8	0.7	0.65	0.725
Promedio	**0.66**	**0.75**	**0.73**	**0.64**	***0.695***

Tabla 3.15. Resultados con promedios de los parámetros y por cada modelo hacia Qwen3

3.4.6. Análisis hecho por Llama3.1

En el examen de los datos de la Tabla 3.11 referidos al modelo Llama3.1, se desarrolló una evaluación donde distintos modelos funcionaron como jueces y valoraron la calidad del análisis realizado. Este enfoque permite establecer comparaciones fiables, asegurando que las apreciaciones sobre los resultados integren distintas perspectivas y mantengan la coherencia derivada de diversas arquitecturas de modelos de lenguaje y de sus métodos de evaluación.

3.4.6.1. Calificación dada por DeepSeek a Llama3.1

DeepSeek otorga una evaluación sobresaliente a Llama3.1, resaltando su claridad y coherencia perfectas (1.0), lo que evidencia una estructura argu-

mental impecable y una redacción fluida. La pertinencia alcanza 0.9 y la precisión 0.8, ambas reflejando un dominio temático sólido y respuestas bien fundamentadas. El promedio general de 0.93 confirma un rendimiento global de nivel alto, cercano a la excelencia evaluativa.

3.4.6.2. Calificación dada por Mistral a Llama3.1

Mistral destaca la pertinencia máxima (1.0), señalando alineación total con la consigna y los criterios de evaluación. La claridad (0.85) y la precisión (0.9) refuerzan la consistencia del modelo, mientras que la coherencia (0.8) constituye el punto más bajo, aunque dentro de un rango favorable. El promedio de 0.89 refleja un desempeño equilibrado, con solidez conceptual y buena correspondencia entre forma y contenido.

3.4.6.3. Calificación dada por Gemma3n a Llama3.1

Gemma3n aplica un esquema de evaluación más estricto, otorgando menor puntuación en claridad (0.6) y coherencia (0.7), donde identifica cierta falta de continuidad discursiva. Sin embargo, pertinencia (0.8) y precisión (0.9) mantienen niveles aceptables, destacando la fidelidad de las respuestas al tema principal. El promedio de 0.75 evidencia un enfoque rigurosamente crítico frente al alto estándar de exigencia del evaluador.

3.4.6.4. Calificación dada por Qwen3 a Llama3.1

Qwen3 valora especialmente la claridad (0.95) y coherencia (0.9), reflejando una lectura fluida y bien estructurada del contenido. La precisión se mantiene en 0.8, respaldando una correcta interpretación de los datos o argumentos tratados. La pertinencia (0.85) aparece como el indicador más débil, aunque sigue dentro de parámetros destacados. Con un promedio final de 0.88, Qwen3 confirma un desempeño robusto y sostenido por parte de Llama3.1.

3.4.6.5. Calificación dada por Gemma3 a Llama3.1

Gemma3 resalta la precisión y la pertinencia (ambas con 0.9), subrayando la capacidad del modelo para responder de forma enfocada y técnicamente exacta. La coherencia alcanza 0.8, consolidando una estructura argumental congruente, mientras que la claridad (0.7) aparece como el aspecto más débil por leves ambigüedades. El promedio de 0.83 indica un rendimiento favorable y balanceado, con margen de mejora en formulación expresiva.

3.4.6.6. Calificación final a Llama3.1

La Tabla 3.16 posiciona la pertinencia (0.89) como la principal fortaleza de Llama3.1, reflejando una respuesta altamente alineada con las consignas y objetivos evaluativos. Le siguen la precisión (0.86), que evidencia un manejo técnico consistente, y la coherencia (0.84), asociada a una buena organización discursiva. La claridad (0.82) completa el conjunto con una comunicación fluida, aunque con margen para optimización en expresividad. El promedio global de 0.85 confirma un rendimiento equilibrado y sólido ante evaluadores diversos, destacando la versatilidad del modelo y su estabilidad en distintas dimensiones de calidad.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-R1:8b	1	0.9	1	0.8	0.93
Mistral:7b	0.85	1	0.8	0.9	0.89
Gemma3n:e4b	0.6	0.8	0.7	0.9	0.75
Qwen3:8b	0.95	0.85	0.9	0.8	0.88
Gemma3:4b	0.7	0.9	0.8	0.9	0.83
Promedio	**0.82**	**0.89**	**0.84**	**0.86**	***0.85***

Tabla 3.16. Evaluación comparativa del modelo Llama3.1:8b

3.4.7. Análisis hecho por Gemma3

En el análisis de los datos obtenidos de la Tabla 3.11 correspondientes al modelo Gemma3, se llevó a cabo un procedimiento de evaluación en el que diversos modelos actuaron como árbitros para ponderar la calidad del análisis desarrollado por cada uno de ellos. Esta metodología establece parámetros comparativos rigurosos, garantizando que la valoración de los resultados no dependa exclusivamente de una única óptica, sino que refleje la pluralidad y la consistencia derivadas de distintas arquitecturas de modelos de lenguaje y de sus respectivas estrategias evaluativas.

3.4.7.1. Calificación dada por DeepSeek a Gemma3

DeepSeek evalúa a Gemma3 con un desempeño general de alto nivel, destacando la claridad y la pertinencia (0.9) como dimensiones bien desarrolladas. La coherencia alcanza 0.95, evidenciando una estructura lógica fluida y argumentación cohesiva. La precisión (0.85) se mantiene en un rango aceptable, aunque con ligeros márgenes de mejora. El promedio de 0.9 confirma

un rendimiento consistente y equilibrado, con especial fortaleza en aspectos comunicativos y de consistencia interna.

3.4.7.2. Calificación dada por Mistral a Gemma3

Mistral otorga máximas puntuaciones en pertinencia y coherencia (1.0), destacando una alineación completa entre contenido y consigna, así como continuidad narrativa impecable. La claridad (0.9) complementa este resultado con buena organización expresiva. En contraste, la precisión (0.8) aparece como el punto más débil, asociada a detalles menores en exactitud conceptual. El promedio de 0.93 evidencia una fortaleza estructural sobresaliente y una ejecución general muy sólida.

3.4.7.3. Calificación dada por Gemma3n a Gemma3

Gemma3n ofrece una visión equilibrada pero más crítica, asignando 0.8 tanto en claridad como en pertinencia, con una coherencia sólida (0.9) que refleja un manejo adecuado del hilo argumental. Sin embargo, la precisión desciende a 0.7, mostrando mayor exigencia respecto al detalle técnico. El promedio final de 0.8 representa una evaluación balanceada, aunque con limitaciones detectadas en la exactitud de las respuestas.

3.4.7.4. Calificación dada por Qwen3 a Gemma3

Qwen3 distribuye sus puntuaciones de forma uniforme en claridad, pertinencia y coherencia (0.9), reconociendo la consistencia global en la estructura y el contenido del modelo. La precisión (0.8) mantiene buenas condiciones, aportando estabilidad al promedio general. Con un resultado de 0.88, Qwen3 confirma un desempeño sostenido de Gemma3 y una ejecución confiable en todas las dimensiones analizadas.

3.4.7.5. Calificación dada por Llama3.1 a Gemma3

Llama3.1 aplica un criterio más riguroso, mostrando menor benevolencia ante las debilidades observadas. La claridad (0.7) y la precisión (0.6) son los puntos más bajos, indicando posibles dificultades en la formulación y el rigor técnico. No obstante, la pertinencia (0.9) y la coherencia (0.8) reflejan buena comprensión temática y cohesión discursiva. El promedio de 0.75 evidencia una mirada evaluativa estricta y minuciosa sobre el desempeño de Gemma3.

3.4.7.6. Calificación final a Gemma3

La Tabla 3.17 resume las tendencias observadas entre evaluadores, posicionando la coherencia (0.91) y la pertinencia (0.9) como las principales fortalezas de Gemma3, al demostrar alta estabilidad en el discurso y adecuación conceptual. En contraste, la precisión (0.75) se identifica como debilidad crítica, sugiriendo inconsistencias o errores menores en detalle técnico, seguida por la claridad (0.84), que presenta áreas de mejora moderadas. El promedio global de 0.85 confirma un rendimiento sólido, aunque con margen para optimizar la exactitud y uniformidad de resultados bajo criterios más exigentes.

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-R1:8b	0.9	0.9	0.95	0.85	0.9
Mistral:7b	0.9	1	1	0.8	0.93
Gemma3n:e4b	0.8	0.8	0.9	0.7	0.8
Qwen3:8b	0.9	0.9	0.9	0.8	0.88
Llama3.1:8b	0.7	0.9	0.8	0.6	0.75
Promedio	**0.84**	**0.9**	**0.91**	**0.75**	***0.85***

Tabla 3.17. Evaluación comparativa del modelo Gemma3:4b

3.4.8. Resultados del análisis de los modelos

La Tabla 3.18 confirma que DeepSeek-r1:8b lidera como el modelo más sólido (promedio 0.906), destacando en pertinencia (0.93) y coherencia (0.92) por su estructura argumental consistente y lógica superior en el análisis Dempster-Shafer del dataset de no-shows médicos. Llama3.1:8b y Gemma3:4b empatan en segundo lugar (0.85), con desempeños equilibrados en todas las métricas. Mistral:7b (0.8415), Gemma3n:e4b (0.83) y Qwen3:8b (0.695) siguen con resultados decrecientes, donde Qwen3 presenta la mayor variabilidad evaluativa debido a la crítica extrema de Mistral (0.35) contrastando con puntuaciones altas de DeepSeek (0.925) y Gemma3n (0.925).

	Claridad	Pertinencia	Coherencia	Precisión	Promedio
DeepSeek-r1:8b	0.88	0.93	0.92	0.894	0.906
Mistral:7b	0.88	0.94	0.84	0.706	0.8415
Gemma3n:e4b	0.76	0.89	0.87	0.78	0.83
Qwen3:8b	0.66	0.75	0.73	0.64	0.695
Llama3.1:8b	0.82	0.89	0.84	0.86	0.85
Gemma3:4b	0.84	0.90	0.91	0.75	0.85

Tabla 3.18. Calificación final promedio otorgada por los jueces a cada modelo en métricas evaluadas (Medical Appointment No Shows)

Capítulo 4

Discusión y Conclusiones

4.1. Discusión de Resultados

Los resultados obtenidos confirman plenamente la hipótesis central de esta tesis: la integración de la Teoría de Dempster-Shafer (DST) con técnicas de explicabilidad como SHAP, materializada en el framework DSExplainer, supera significativamente los métodos convencionales al proporcionar una cuantificación explícita y realista de la incertidumbre epistémica mediante intervalos de creencia-plausibilidad (Bel-Pl).

En los experimentos realizados con datasets canónicos como Titanic, Iris y Breast Cancer, DSExplainer demuestra que mantiene la interpretabilidad aditiva característica de SHAP mientras incorpora un diagnóstico robusto de incertidumbre que los métodos tradicionales no pueden igualar. La Tabla 3.1 del capítulo anterior muestra de manera clara esta superioridad: los anchos de intervalo Bel-Pl generados por DSExplainer son sistemáticamente más amplios y realistas que las bandas bootstrap de SHAP. Por ejemplo, en el dataset Titanic con interacciones de orden $k = 2$, DSExplainer reporta un ancho promedio de 0.505, frente a solo 0.124 de SHAP bootstrap. Esta diferencia es crucial porque evita la sobreconfianza artificial que surge cuando la variabilidad se reparte en espacios de mayor dimensionalidad.

El comportamiento de ambos métodos al incrementar el orden de interacción k de 1 a 2 es particularmente revelador. Mientras SHAP tiende a contraer artificialmente sus bandas percentilares –lo que puede llevar a inter-

pretaciones engañosamente precisas–, DSExplainer expande legítimamente los intervalos Bel-Pl. Este comportamiento reconoce una verdad fundamental: las hipótesis de interacción de orden superior conllevan intrínsecamente mayor incertidumbre residual, y el framework lo refleja de manera transparente.

En el análisis global de los datasets, las interacciones socio-demográficas como "sexo-edad" y "tarifa-cabina" en Titanic concentran la mayor creencia (Bel) con intervalos estrechos que reflejan estabilidad evidencial robusta. En Iris, las medidas morfológicas de pétalos dominan tanto la creencia como la plausibilidad, validando la capacidad del framework para capturar la estructura geométrica subyacente que separa las especies. Finalmente, en Breast Cancer se identifican marcadores morfológicos clave de malignidad como "radio_peor-perimetro_peor" con alta creencia, aunque con expansiones plausibilísticas que destacan sutiles dependencias contextuales no evidentes en el análisis SHAP estándar.

La evaluación cruzada de Large Language Models (LLMs) en datasets médicos reales –Alzheimer y Medical Appointment No-Shows– corrobora la aplicabilidad práctica de toda la metodología propuesta. Los resultados posicionan a Llama3.1 como líder absoluto en coherencia argumentativa (promedio 0.92), mientras Qwen3 destaca en pertinencia temática (0.95 evaluando DeepSeek). Esta variabilidad sistemática entre evaluadores –con Llama3.1 demostrando un rigor notablemente superior (solo 0.4 en coherencia para DeepSeek)– valida completamente el enfoque innovador de evaluación distribuida presentado en la sección 2.11. Al fusionar estas evaluaciones mediante la regla de combinación Dempster, se obtiene una valoración agregada con bajo conflicto K que mitiga sesgos individuales de manera elegante.

Tabla 4.1. Desempeño relativo de LLMs por dataset y fortaleza principal

Dataset	**Mejor LLM (Promedio)**	**Fortaleza Principal**
Alzheimer	Qwen3 (0.85)	Pertinencia contextual
No-Shows	Llama3.1 (0.92)	Coherencia argumentativa
DSExplainer (Global)	N/A	Diagnóstico Bel-Pl

4.2. Cumplimiento de Objetivos

El objetivo general de la tesis –desarrollar un framework integral para inteligencia artificial explicable bajo condiciones de incertidumbre mediante

DST– se cumple de manera ejemplar con la creación de DSExplainer. Este framework no solo extiende las capacidades de SHAP preservando su interpretabilidad aditiva característica, sino que incorpora de manera natural la gestión de masas de probabilidad básica (BPA) derivadas de intervalos bootstrap, proporcionando una representación mucho más rica de la incertidumbre subyacente.

Los objetivos específicos delineados en la sección 2.4.1 se materializan de la siguiente manera concreta:

El preprocesamiento robusto descrito en la sección 2.9 –que incluye Label Encoding para variables categóricas, normalización Min-Max/Z-score para escalas heterogéneas, y generación sistemática de variables combinadas– habilita directamente la asignación de masas basada en varianza que constituye el núcleo del enfoque Dempster-Shafer (sección 2.7).

La validación distribuida mediante evaluación cruzada de LLMs con criterios estructurados de coherencia, pertinencia, precisión y claridad (sección 2.8), fusionados mediante la regla Dempster, supera con creces las matrices de evaluación humana tradicionales tanto en objetividad como en escalabilidad.

La hipótesis central del framework DSExplainer (sección 2.3.8) –que los intervalos Bel-Pl serían más realistas que los bootstrap estándar de SHAP– queda confirmada empíricamente tanto por la Tabla 3.1 como por el análisis de estabilidad mediante coeficiente Jaccard sobre las top-N hipótesis bootstrap.

Finalmente, las aplicaciones prácticas en datasets del mundo real (sección 2.6) demuestran éxito rotundo: en Medical Appointment No-Shows, el framework identifica con precisión los factores socioeconómicos dominantes de la inasistencia a citas médicas; en Alzheimer, las interacciones entre MMSE y ADL emergen como predictoras clave con cuantificación explícita de su incertidumbre asociada.

4.3. Limitaciones del Estudio

Aunque los resultados son inequívocamente positivos, es importante reconocer las limitaciones estructurales y metodológicas que delimitan el alcance actual del trabajo:

La limitación más significativa es la escalabilidad computacional inherente al enfoque combinatorio. Cuando se incrementa el orden máximo de

interacción k más allá de 2, el espacio de hipótesis explota exponencialmente –llegando a 41,448 hipótesis en el dataset Alzheimer con $k = 4$, según la Tabla 3.2–. Esto restringe la aplicabilidad práctica a datasets de muy alta dimensionalidad sin implementar mecanismos de preselección inteligente de interacciones.

La validación mediante LLMs, aunque innovadora, mostró sensibilidad notable a la especificidad de los prompts de evaluación (sección 3.2.5). Si bien la adición de glosarios mínimos y convenciones explícitas de escalado eliminó la mayoría de errores, persiste una variabilidad evaluativa que refleja las limitaciones inherentes de modelos no fine-tuneados específicamente para tareas Dempster-Shafer.

Una limitación metodológica importante es la ausencia de validación humana externa por parte de expertos dominio-específicos –neurólogos para Alzheimer, epidemiólogos para No-Shows–. Aunque la evaluación cruzada LLM representa un avance significativo hacia la descentralización evaluativa, aún requiere confrontación con ground-truth humano para aplicaciones clínicas reales.

La asignación de masas de creencia proporcional a la varianza de variables combinadas (sección 2.7), aunque teóricamente fundamentada, introduce un sesgo potencial hacia características volátiles que podría subestimar predictores estables pero clínicamente críticos.

Finalmente, todo el análisis se limita a datasets tabulares estáticos. No se prueba la extensibilidad del framework a escenarios multimodales –combinando imágenes médicas con informes clínicos, por ejemplo– ni a despliegues en tiempo real.

4.4. Trabajo Futuro

Cada una de las limitaciones identificadas sugiere direcciones naturales y prioritarias para la extensión del trabajo:

Optimizar DSExplainer mediante algoritmos de preselección inteligente de interacciones, utilizando métricas como información mutua o correlaciones parciales para restringir el espacio combinatorio y habilitar análisis con $k = 3+$ en datasets del mundo real.

Desarrollar LLMs específicamente fine-tuneados con funciones de pérdida que penalicen inconsistencias evidenciales y manejen explícitamente la semántica de creencia/plausibilidad, reduciendo así los sesgos evaluativos

observados en los experimentos actuales.

Establecer colaboraciones con centros médicos para implementar validación empírica robusta: ground-truth humano proporcionado por neurólogos y epidemiólogos, midiendo el impacto clínico real de las explicaciones DSExplainer en la toma de decisiones.

Extender el framework hacia multimodalidad mediante fusión jerárquica Dempster que combine evidencias de diferentes modalidades –rayos-X con informes clínicos, series temporales fisiológicas con notas de enfermería– manteniendo trazabilidad completa de incertidumbre.

Desarrollar versiones optimizadas para despliegue industrial real-time, integrando DSExplainer con frameworks de edge computing para monitoreo predictivo continuo en manufactura y salud.

Alinear explícitamente el framework con regulaciones emergentes como el UE AI Act, implementando auditorías automáticas que utilicen los intervalos Bel-Pl como proxy cuantitativo de confianza calibrada para sistemas de alto riesgo.

4.5. Conclusiones Finales

Esta tesis representa un avance significativo y sistemático en el campo de la inteligencia artificial explicable al fusionar de manera elegante la Teoría de Dempster-Shafer con técnicas state-of-the-art de explicabilidad como SHAP y evaluación mediante Large Language Models.

El framework DSExplainer emerge no solo como prueba de concepto, sino como herramienta madura y reproducible que resuelve tres limitaciones fundamentales de los enfoques existentes:

Preserva completamente la interpretabilidad aditiva característica de SHAP mientras incorpora un diagnóstico explícito y cuantitativo de incertidumbre epistémica mediante intervalos Bel-Pl sistemáticamente más realistas que cualquier baseline bootstrap.

Demuestra éxito empírico rotundo tanto en datasets canónicos de machine learning (Titanic, Iris, Breast Cancer – Tabla 3.1) como en aplicaciones médicas del mundo real críticas (Alzheimer, Medical No-Shows – Tablas 3.4-3.10), identificando patrones predictivos clínicamente relevantes con trazabilidad completa de su incertidumbre asociada.

Innova metodológicamente al introducir evaluación cruzada distribuida de LLMs fusionada mediante Dempster, superando las limitaciones de la

evaluación humana tradicional tanto en escalabilidad como en mitigación de sesgos individuales.

Las contribuciones concretas de esta investigación incluyen el framework DSExplainer completamente reproducible –con mapeo multi-variante SHAP→BPA, fusión escalar Dempster, baselines comparativas exhaustivas y funciones auxiliares de métricas–, una nueva metodología de evaluación híbrida LLM+DST para IA explicable, y evidencia empírica irrefutable de la superioridad diagnóstica del enfoque Dempster-Shafer sobre métodos probabilísticos convencionales en la gestión de evidencia incierta y conflictiva.

DSExplainer se posiciona así como una herramienta lista para adopción práctica en dominios donde la confianza bajo incertidumbre es paramount: diagnóstico médico asistido, monitoreo predictivo industrial, sistemas de salud pública, y cualquier escenario donde las decisiones automatizadas deben ser no solo precisas, sino comprensiblemente justificadas ante usuarios humanos y reguladores.

Esta tesis no pretende cerrar el ciclo de investigación en DST-XAI, sino catalizar su aceleración hacia inteligencia artificial verdaderamente confiable, interpretable y alineada con los imperativos éticos y regulatorios del siglo XXI.

"La evidencia no es estática; es un proceso dinámico de refinamiento continuo bajo incertidumbre."

Referencias bibliográficas

[1] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[2] S. Kaur and B. Kaur, "Neural networks and their applications," *International Journal of Electronics
Communication Technology*, vol. 3, no. 4, pp. 414–417, 2012.

[3] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[4] S.-M. Udrescu and M. Tegmark, "Ai feynman: A physics-inspired method for symbolic regression," *Science Advances*, vol. 6, no. 16, p. eaay2631, 2020.

[5] A. P. Dempster, *Upper And Lower Probabilities induced by multivalued mapping*. Hardvard University, 1967.

[6] Z. Tong, P. Xu, and T. Denœux, "An evidential classifier based on dempster-shafer theory and deep learning," *Neurocomputing*, vol. 450, pp. 275–293, 2021.

[7] S. Peñafiel, N. Baloian, H. Sanson, and J. A. Pino, "Applying dempster-shafer theory for developing a flexible, accurate, and interpretable classifier," *Expert Systems with Applications*, vol. 148, p. 113262, 2020.

[8] M. F. Rosli, L. M. Hee, and M. S. Leong, "Integration of artificial intelligence into dempster shafer theory: A review on decision making in condition monitoring," *Applied Mechanics and Materials*, vol. 773-774, pp. 154–157, 2015.

[9] P. Aggarwal, D. Bhatt, V. Devabhaktuni, and P. Bhattacharya, "Dempster shafer neural network algorithm for land vehicle navigation application," *Information Sciences*, vol. xxx, pp. xxx–xxx, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2013.08.039

[10] K. H. Hui, C. S. Ooi, M. H. Lim, and M. S. Leong, "A hybrid artificial neural network with dempster-shafer theory for automated bearing fault diagnosis," *Journal of Vibroengineering*, vol. 18, no. 7, pp. 4409–4418, 2016. [Online]. Available: https://doi.org/10.21595/jve.2016.17024

[11] K. Yu, T. R. Lin, and J. Tan, "A bearing fault and severity diagnostic technique using adaptive deep belief networks and dempster–shafer theory," *Structural Health Monitoring*, vol. xx, no. xx, pp. 1–22, 2019. [Online]. Available: https://doi.org/10.1177/1475921719841690

[12] M. Ladjal, M. Bouamar, M. Djerioui, and Y. Brik, "Performance evaluation of ann and svm multiclass models for intelligent water quality classification using dempster-shafer theory," in *2nd International Conference on Electrical and Information Technologies (ICEIT)*. IEEE, 2016, pp. xx–xx. [Online]. Available: https://doi.org/10.1109/ICEIT.2016.8469

[13] T. Denoeux, "A neural network classifier based on dempster-shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 30, no. 2, pp. 131–150, 2000. [Online]. Available: https://doi.org/10.1109/3468.840040

[14] ——, "Logistic regression, neural networks and dempster-shafer theory: A new perspective," *Knowledge-Based Systems*, 2019, preprint arXiv:1807.01846v3.

[15] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghasem-Aghaee, "Sentiment prediction based on dempster-shafer theory of evidence," *Mathematical Problems in Engineering*, vol. 2014, p. Article ID 361201, 2014.

[16] L. Ang, M. H. Yim, J.-H. Do, and S. Lee, "A novel method in predicting hypertension using facial images," *Applied Sciences*, vol. 11, no. 5, p. 2414, 2021.

[17] M. Khouja, "The single-period (news-vendor) problem: Literature review and suggestions for future research," *Omega*, vol. 27, no. 5, pp. 537–553, 1999.

[18] L. A. Zadeh, "A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination," *AI Magazine*, vol. 7, no. 2, pp. 85–90, 1986.

[19] N. Baloian, J. Frez, J. A. Pino, and G. Zurita, "Supporting collaborative preparation of emergency plans," *Proceedings*, vol. 2, no. 1254, pp. 1–11, 2018. [Online]. Available: https://doi.org/10.3390/proceedings2191254

[20] P. Bhattacharya, "On the dempster-shafer evidence theory and non-hierarchical aggregation of belief structures," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 30, no. 5, pp. 526–535, 2000.

[21] Y. Mulyani, E. F. Rahman, Herbert, and L. S. Riza, "A new approach on prediction of fever disease by using a combination of dempster shafer and naïve bayes," in *2016 2nd International Conference on Science in Information Technology (ICSITech)*. IEEE, 2016, pp. 367–372.

[22] Le, Trung H., Nguyen, Huynh A.D., Ha, Quang P., Tran, Minh Q., Ahmed, Masrur, Kong, Jing, Barthelemy, Xavier, Duc, Hiep, Jiang, Ningbo, Azzi, Merched, and Riley, Matthew, "Dempster-shafer ensemble learning framework for air pollution nowcasting," *E3S Web Conf.*, vol. 626, p. 01003, 2025. [Online]. Available: https://doi.org/10.1051/e3sconf/202562601003

[23] J. Frez, "DSExplainer," 2026. [Online]. Available: https://github.com/jfrez/DSExplainer

[24] R. E. Kharoua, "Alzheimer's disease dataset," 2024. [Online]. Available: https://www.kaggle.com/dsv/8668279

[25] Joniarroba, "Medical appointment no shows dataset," https://www.kaggle.com/datasets/joniarroba/noshowappointments, 2017.