ANÁLISIS DEL DESEMPEÑO DE UN PERCEPTRÓN MULTICAPA PARA EL DIAGNÓSTICO DE ALZHEIMER CONSIDERANDO LA COMBINACIÓN DE TÉCNICAS DE SELECCIÓN DE VARIABLES Y MÉTODOS DE ESCALADO

PERFORMANCE ANALYSIS OF A MULTILAYER PERCEPTRON FOR ALZHEIMER'S DIAGNOSIS CONSIDERING THE COMBINATION OF FEATURE SELECTION TECHNIQUES AND SCALING METHODS

Alejandra Guadalupe Bravo García
Centro Universitario UAEM Valle de México, México
abravog003@alumno.uaemex.mx

Maricela Quintana López
Centro Universitario UAEM Valle de México, México
mquintanal@uaemex.mx

Víctor Manuel Landassuri Moreno
Centro Universitario UAEM Valle de México, México
vmlandassurim@uaemex.mx

Saul Lazcano Salas
Centro Universitario UAEM Valle de México, México
slazcanos@uaemex.mx

Asdrúbal López Chau
Centro Universitario UAEM Zumpango, México
alchau@uaemex.mx

Recepción: 25/noviembre/2025 **Aceptación:** 24/diciembre/2025

Resumen

El diagnóstico del Alzheimer representa un reto clínico, debido a su carácter multifactorial y la necesidad de identificar patrones sutiles en datos clínicos. Este artículo evalúa el desempeño de un Perceptrón Multicapa (MLP) para el diagnóstico de Alzheimer, comparando cuatro técnicas de selección de características: Chi-cuadrado (Chi^2), Información Mutua (MI), Bosque Aleatorio (RF) y Regresión Logística $L1$; combinadas con cuatro métodos de escalado: Min-Max, StandardScaler, RobustScaler y normalización en dos pasos, sobre datos clínicos.

Los resultados muestran que Chi^2 con RobustScaler alcanzó el mejor desempeño global $(exactitud = 0.9069; AUC = 0.9361)$, y que Chi^2 y RF fueron los métodos de selección más estables entre métricas. La comparación evidencia que la elección del escalado influye de manera sustantiva en el rendimiento del clasificador e interactúa con la selección de características. Se concluye que flujos de preprocesamiento bien diseñados, integrando selección y escalado, potencian los modelos MLP para la detección de Alzheimer.

Palabras Clave: Escalado de datos, Perceptrón Multicapa (MLP), selección de variables.

Abstract

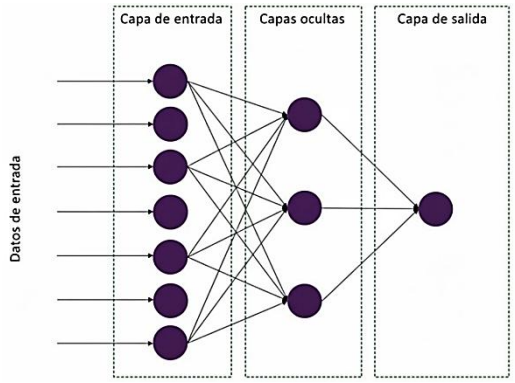
The diagnosis of Alzheimer's disease represents a clinical challenge due to its multifactorial nature and the need to identify subtle patterns in clinical data. This study evaluates the performance of a Multilayer Perceptron (MLP) for Alzheimer's diagnosis, comparing four feature selection techniques: Chi-Squared (Chi^2), Mutual Information (MI), Random Forest (RF), and $L1$ Logistic Regression; combined with four scaling methods: Min-Max, StandardScaler, RobustScaler, and Two-Step Normalization, applied to clinical data. The results show that the Chi^2 with RobustScaler combination achieved the best overall performance ($accuracy = 0.9069; AUC = 0.9361$), and that Chi^2 and RF were the most stable feature selection methods across metrics. The comparison highlights that the choice of scaling significantly affects classifier performance and interacts with feature selection. It is concluded that well-designed preprocessing pipelines integrating selection and scaling enhance the effectiveness of MLP models for Alzheimer's detection.

Keywords: *Data scaling, feature selection, Multilayer Perceptron (MLP).*

1. Introducción

La enfermedad de Alzheimer es un trastorno neurodegenerativo progresivo que afecta profundamente la calidad de vida de millones de personas en todo el mundo, situándose entre las principales causas de discapacidad en adultos mayores

[Cabanillas, 2025]. En este contexto, lograr un diagnóstico temprano y preciso resulta fundamental, ya que permite implementar intervenciones oportunas que pueden ralentizar el avance de la enfermedad y mejorar el pronóstico de los pacientes [El-Sappagh, 2021], [Zhou, 2024]. Con este propósito, la Inteligencia Artificial (IA), y en particular el Machine Learning (ML), se ha consolidado como una herramienta valiosa para potenciar la detección temprana y la predicción de la progresión del Alzheimer [Davuluri, 2020]. Modelos como el Perceptrón Multicapa (Figura 1) son capaces de identificar patrones complejos en extensos conjuntos de datos clínicos y perfiles de biomarcadores, proporcionando un apoyo significativo para realizar pronósticos más precisos [Aguayo, 2023], [El-Sappagh, 2021]. Sin embargo, la efectividad de estos modelos predictivos no depende únicamente de su arquitectura, también, de forma crítica, de la calidad, relevancia y adecuado preprocesamiento de los datos de entrada.



Fuente: elaboración propia

Figura 1 Arquitectura de un Perceptrón Multicapa.

En este sentido, la selección de características desempeña un papel esencial al reducir el ruido, eliminar variables irrelevantes o redundantes y mejorar así tanto la precisión como la capacidad de generalización de los modelos de ML [Pudjihartono, 2022], [Venkatesh, 2019]. Técnicas como Chi-cuadrado, Información Mutua y métodos basados en la importancia de los atributos, como Bosque Aleatorio, son ampliamente utilizados para cumplir este objetivo [Davuluri, 2020], [Pudjihartono, 2022]. Su correcta aplicación asegura que los modelos se centren en patrones clínicamente significativos, aspecto especialmente relevante en aplicaciones

médicas. Por otro lado, el desarrollo de sistemas predictivos confiables también depende del proceso de escalado de datos. Métodos como MinMaxScaler y StandardScaler estandarizan los rangos de las variables, favoreciendo la convergencia durante el entrenamiento del modelo y minimizando el impacto de las diferencias de escala en la dinámica de aprendizaje [Sharma, 2022]. Un escalado apropiado no solo acelera la optimización, sino que también mejora la estabilidad del proceso de aprendizaje. Es importante señalar que el término escalado engloba tanto la normalización como la estandarización, aunque estos conceptos no son intercambiables y se refieren a procesos distintos. La normalización generalmente implica transformar los valores de las variables para que se ubiquen dentro de un rango específico, típicamente entre 0 y 1, como en el caso del método MinMaxScaler. En cambio, la estandarización transforma los datos para que tengan una media de 0 y una desviación estándar de 1, lo cual es útil cuando los datos siguen una distribución aproximadamente normal, como ocurre con StandardScaler [Sree, 2018]. La elección entre una u otra técnica depende del modelo utilizado y de la naturaleza de los datos, ya que cada método tiene un impacto diferente en la dinámica del aprendizaje.

2. Métodos

La metodología de este estudio fue diseñada para evaluar de manera sistemática cómo diferentes técnicas de selección de atributos y escalado de datos afectan el desempeño de una MLP en el diagnóstico temprano del Alzheimer. Se inicia describiendo las características del conjunto de datos clínicos utilizado. Posteriormente, se detallan el esquema experimental para el entrenamiento, la prueba del modelo y el análisis comparativo de combinaciones de preprocesamiento.

Dataset

El conjunto de datos empleado en este estudio está compuesto por registros clínicos de $2{,}149$ pacientes, que incluyen un total de 35 variables que abarcan distintos dominios, Tabla 1.

Tabla 1 Atributos y descripción del conjunto de datos.

Categoría	**Atributo**	**Descripción**
Identificación	PatientID	Un identificador único asignado a cada paciente
Detalles demográficos	Age	Edades $(60 - 90\ años)$
	Gender	0 = Hombre, 1 = Mujer
	Ethnicity	0 = Caucásico, 1 = Afroamericano, 2 = Asiático, 3 = Otros.
	EducationLevel	0 = Ninguno, 1 = Secundaria, 2 =Preparatoria, 3 = Superior.
Factores de estilo de vida	BMI	Índice de Masa Corporal $(15 - 40)$.
	Smoking	0 = No, 1 = Sí.
	AlcoholConsumption	Consumo semanal de alcohol $(0 - 20$ unidades).
	PhysicalActivity	Actividad física semanal $(0 - 10$ horas).
	DietQuality	Puntuación de calidad de la dieta $(0 - 10)$.
	SleepQuality	Puntuación de la calidad de sueño $(4 - 10)$.
Historial médico	FamilyHistoryAlzheimers	0 = No, 1 = Sí.
	CardiovascularDisease	0 = No, 1 = Sí.
	Diabetes	0 = No, 1 = Sí.
	Depression	0 = No, 1 = Sí.
	HeadInjury	0 = No, 1 = Sí.
	Hypertension	0 = No, 1 = Sí.
Estudios clínicos	SystolicBP	Presión arterial sistólica $(90 - 180\ mmHg)$.
	DiastolicBP	Presión arterial diastólica $(60 - 120\ mmHg)$.
	CholesterolTotal	Colesterol $(150 - 300\ mg/dL)$.
	CholesterolLDL	LDL Colesterol $(50 - 200\ mg/dL)$.
	CholesterolHDL	HDL Colesterol $(20 - 100\ mg/dL)$.
	CholesterolTriglycerides	Triglicéridos $(50 - 400\ mg/dL)$.
Evaluaciones cognitivas y funcionales	MMSE	Puntuación del Mini-Examen del Estado Mental.
	FunctionalAssessment	Puntuación Funcional $(0 - 10)$.
	MemoryComplaints	0 = No, 1 = Sí.
	BehavioralProblems	0 = No, 1 = Sí.
	ADL	Puntuación de actividades de la vida diaria $(0 - 10)$.
Síntomas	Confusion	0 = No, 1 = Sí.
	Disorientation	0 = No, 1 = Sí.
	PersonalityChanges	0 = No, 1 = Sí.
	DifficultyCompletingTasks	0 = No, 1 = Sí.
	Forgetfulness	0 = No, 1 = Sí.
Diagnóstico	Diagnosis	0 = No Alzheimer, 1 = Alzheimer.
ID Confidencial	DoctorInCharge	'XXXConfid' para todos los pacientes.

Fuente: elaboración propia.

La heterogeneidad de estas características es de suma importancia, dado que el Alzheimer es una enfermedad multifactorial que involucra la interacción de componentes biológicos, cognitivos y conductuales. Al integrar este tipo de variables, el modelo MLP puede aprender patrones sutiles que podrían indicar estadios tempranos de Alzheimer. Además, esta diversidad resalta la necesidad de un preprocesamiento robusto, pues un manejo inadecuado podría amplificar el ruido o enmascarar relaciones críticas, impactando directamente la capacidad del modelo para generalizar.

Preprocesamiento

La variable objetivo de este estudio fue Diagnosis, que diferencia entre pacientes diagnosticados con Alzheimer (1) y aquellos sin diagnóstico (0). Esta variable se utilizó como clase de salida en todos los experimentos de clasificación. El conjunto de datos estaba completamente limpio, sin valores faltantes, por lo que no fue necesario aplicar técnicas de imputación. Se eliminaron columnas como DoctorInCharge debido a su falta de valor predictivo y a que contenían identificadores confidenciales. Del mismo modo, se excluyó PatientID al ser un identificador único sin relación con la progresión de la enfermedad. Las variables categóricas se codificaron en formato numérico cuando fue necesario, para permitir su procesamiento por los algoritmos de aprendizaje automático. Posteriormente, se realizó una partición consistente del conjunto de datos, destinando el 80% para entrenamiento y el 20% para prueba, utilizando una semilla fija ($random\_state = 42$) para garantizar la reproducibilidad en todos los experimentos.

Selección de características

Para examinar la influencia de diversas estrategias de reducción de variables en el rendimiento del modelo, este estudio aplicó cuatro técnicas de selección de características ampliamente utilizadas: Chi^2, Información Mutua, Bosques Aleatorios y Regresión Logística $L1$. Cada uno de estos métodos emplea principios matemáticos distintos que permiten identificar los atributos más relevantes del conjunto de datos, influyendo directamente en la capacidad del MLP para generalizar y detectar patrones clínicos sutiles.

La prueba Chi-cuadrado es un método tipo filtro que mide la dependencia estadística entre variables categóricas y la clase binaria mediante la comparación de frecuencias observadas y esperadas. Aquellas características con puntuaciones Chi^2 más altas muestran una asociación más fuerte con el diagnóstico, como se ha demostrado en investigaciones recientes sobre ingeniería de características híbrida [Silpa, 2024], facilitando la priorización de variables potencialmente indicativas del estado de Alzheimer. La Información Mutua cuantifica cuánto se reduce la incertidumbre al conocer una característica respecto a la variable objetivo [Zhou,

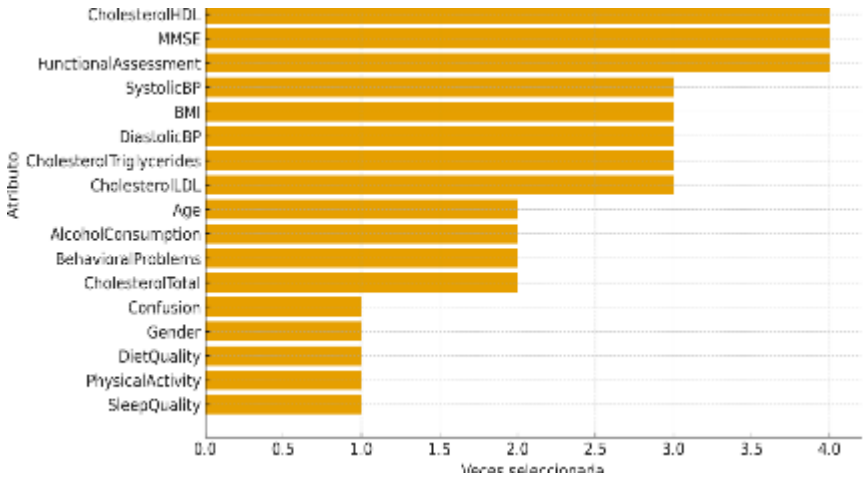
2024]. A diferencia de la correlación, captura dependencias tanto lineales como no lineales, aspecto clave en enfermedades multifactoriales como el Alzheimer, donde las interacciones entre variables suelen ser complejas. Recientes avances en su estimación han demostrado que la MI permite identificar relaciones de alta dimensionalidad incluso en sistemas biológicos complejos, al modelar estructuras de dependencia no lineales mediante representaciones latentes aprendidas [Gowri, 2024], reforzando su utilidad como métrica robusta para el análisis de interacciones entre variables clínicas y cognitivas.

Por otro lado, Random Forest utiliza un ensamble de árboles de decisión para evaluar cuánto disminuye la impureza de cada atributo a lo largo de los nodos de división. Este enfoque embebido no solo ordena las variables según su contribución predictiva, sino que también considera interacciones y efectos no lineales, resultando especialmente adecuado para datos clínicos heterogéneos. Diversos estudios han demostrado su eficacia en el desarrollo de modelos predictivos médicos, particularmente en escenarios longitudinales y con alta dimensionalidad, permitiendo capturar dependencias complejas entre variables clínicas y resultados binarios [Speiser, 2021]. Dicho enfoque ha sido aplicado con éxito en la predicción de discapacidades de movilidad y otros desenlaces clínicos relevantes, evidenciando su potencial en contextos de diagnóstico médico asistido por aprendizaje automático.

Finalmente, Regresión Logística $L1$ introduce una penalización sobre los valores absolutos de los coeficientes durante la optimización, reduciendo a cero los pesos de las características menos informativas. De esta manera, actúa como un mecanismo interno de selección, simplificando el modelo y disminuyendo el riesgo de sobreajuste en conjuntos con predictores potencialmente correlacionados [Rahman, 2020].

Para cada algoritmo se seleccionaron las 10 características más relevantes de acuerdo con el criterio interno de ranking propio de cada técnica. Los atributos más frecuentemente identificados por estos métodos, que fueron los puntajes ADL, quejas de memoria, colesterol HDL, MMSE y presión sistólica, mostrados en la Figura 2. Esta selección colectiva evidencia cómo distintas metodologías priorizan

consistentemente marcadores clínicos y cognitivos estrechamente ligados a la progresión del Alzheimer.

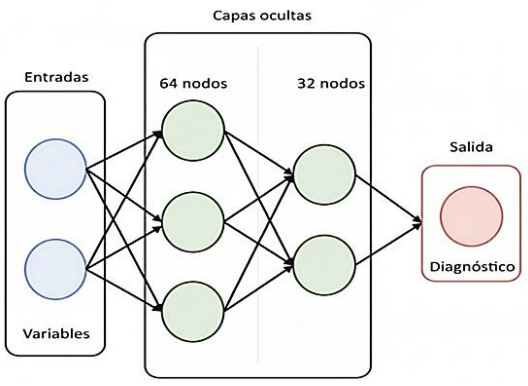


Fuente: elaboración propia

Figura 2 Frecuencia de los atributos seleccionados.

Arquitectura del modelo

El perceptrón multicapa diseñado para este estudio fue estructurado cuidadosamente para equilibrar la complejidad del modelo con su capacidad de generalización, garantizando que las diferencias observadas en el rendimiento se atribuyeran principalmente a las variaciones en el preprocesamiento y no a la arquitectura misma. La red constó de una capa de entrada cuya dimensionalidad coincidía con el número de características seleccionadas por cada combinación de preprocesamiento, seguida de dos capas ocultas con 64 y 32 neuronas, Figura 3.



Fuente: elaboración propia

Figura 3 Representación del MLP empleado.

Cada capa oculta utilizó la función de activación ReLU (Rectified Linear Unit), ampliamente empleada en redes neuronales por su capacidad de modelar relaciones no lineales complejas y mitigar problemas como el desvanecimiento del gradiente.

Para fomentar la generalización y reducir el riesgo de sobreajuste, se incorporaron capas de $dropout$ con una tasa de 0.3 después de cada capa oculta, desactivando aleatoriamente neuronas durante el entrenamiento. Esta técnica de regularización evita que la red dependa excesivamente de una sola ruta a través de su estructura. La capa de salida consistió en una única neurona activada mediante una función sigmoide, adecuada para clasificación binaria al mapear las salidas a probabilidades entre 0 y 1. El entrenamiento del modelo empleó la función de pérdida binary cross entropy, que mide la divergencia entre las probabilidades predichas y las etiquetas binarias verdaderas, optimizada usando el algoritmo Adam, elegido por sus tasas de aprendizaje adaptativas y su eficacia demostrada en problemas de optimización a gran escala.

Además, el entrenamiento se configuró hasta un máximo de 100 épocas, con EarlyStopping y una división de prueba del 20% para monitorizar el sobreajuste, conservando los pesos del modelo que lograron el mejor desempeño en el conjunto de prueba. Esta arquitectura y esquema de entrenamiento consistentes en todas las ejecuciones experimentales garantizaron que las diferencias en el rendimiento predictivo fueran impulsadas por las estrategias de selección y escalado aplicadas.

Configuración experimental

Para evaluar de forma sistemática el impacto del preprocesamiento en el rendimiento del modelo, este estudio implementó un diseño experimental que combinó cuatro técnicas distintas de selección de características con cuatro métodos de escalado de datos, dando lugar a 16 combinaciones de preprocesamiento únicos, como se observan en la Tabla 2. Esta configuración integral fue esencial para determinar cómo cada combinación influía en la capacidad del perceptrón multicapa para diferenciar a los pacientes diagnosticados con Alzheimer de los que no.

Tabla 2 Matriz de combinaciones.

	Métodos de escalado			
Técnicas de selección	**Min-Max**	**StandardScaler**	**RobustScaler**	**Normalización 2 pasos**
Chi²	Chi^2 con Min-Max	Chi^2 con StandardScaler	Chi^2 con RobustScaler	Chi^2 con Normalización 2 pasos
Mutual Info	Mutual Info con Min-Max	Mutual Info con StandardScaler	Mutual Info con RobustScaler	Mutual Info con Normalización 2 pasos
Random Forest	Random Forest con Min-Max	Random Forest con StandardScaler	Random Forest con RobustScaler	Random Forest con Normalización 2 pasos
Regresión Logística	Regresión Logística con Min-Max	Regresión Logística con StandardScaler	Regresión Logística con RobustScaler	Regresión Logística con Normalización 2 pasos

Fuente: elaboración propia

En particular, el pipeline incluyó la selección de características mediante Chi-cuadrado, Información Mutua, importancia en Bosques Aleatorios o Regresión Logística con regularización $L1$. Cada conjunto de características seleccionadas se procesó posteriormente con uno de cuatro métodos de escalado: Min-Max Scaling, StandardScaler, RobustScaler o el enfoque de normalización en dos pasos.

Estos experimentos permitieron explorar rigurosamente cómo tanto la reducción de dimensionalidad como el escalamiento de los datos, contribuyen a los resultados predictivos. Todos los modelos emplearon una arquitectura consistente, garantizando que las diferencias observadas en el desempeño fueran atribuibles exclusivamente a las variaciones en el preprocesamiento. Cada modelo fue evaluado utilizando cinco métricas clave, exactitud, precisión, recall, $F1-score$ y área bajo la curva ROC (AUC), ofreciendo así, una valoración de la precisión de la clasificación. La Tabla 3 presenta las características más frecuentemente seleccionadas por cada técnica de selección de atributos, donde 1 significa “seleccionada” y 0 “no seleccionada”.

3. Resultados

En este estudio se presenta cómo afecta al desempeño de una MLP, la combinación de técnica de selección de características y método de escalado que se elija. Para ello, se evaluaron 16 combinaciones mediante cinco métricas clave (exactitud, precisión, recall, $F1-score$ y AUC). Los hallazgos obtenidos se pueden

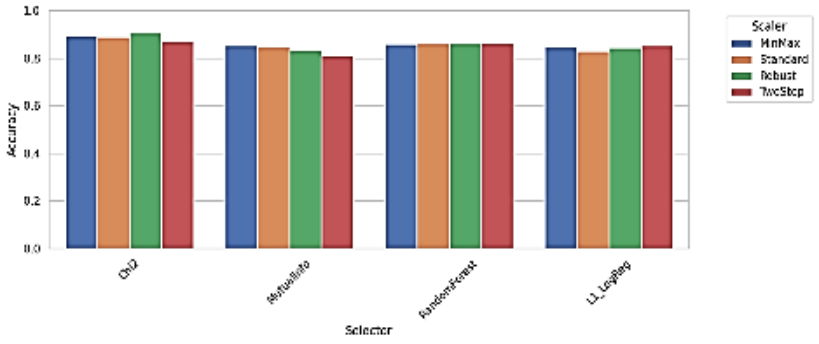
observar en los gráficos de barras (Figuras 4 a 8) y el mapa de calor comparativo (Figura 9), haciendo posible una interpretación detallada para comprender qué combinaciones ofrecen resultados con mayor valor clínico.

Tabla 3 Atributos seleccionados por técnica de selección de atributos.

		Chi^2	**MutualInfo**	**RandomForest**	$Reg - Log - L1$
Atributos	ADL	1	1	1	1
	Age	0	1	0	1
	AlcoholConsumption	0	1	1	0
	BMI	1	0	1	1
	BehavioralProblems	1	0	1	0
	CholesterolHDL	1	1	1	1
	CholesterolLDL	1	0	1	1
	CholesterolTotal	0	0	1	1
	CholesterolTriglycerides	1	0	1	1
	Confusion	0	1	0	0
	DiastolicBP	0	1	1	1
	DietQuality	0	0	1	0
	FunctionalAssessment	1	1	1	1
	Gender	0	1	0	0
	MMSE	1	1	1	1
	MemoryComplaints	1	1	1	1
	PhysicalActivity	0	0	1	0
	SleepQuality	0	0	1	0
	SystolicBP	1	0	1	1

Fuente: elaboración propia

La Figura 4 muestra la exactitud alcanzada por cada combinación. Destaca que el uso de la selección de características mediante Random Forest junto con la normalización en dos pasos fue el que logró la mayor exactitud, superando el 94%, lo que subraya cómo los rankings de importancia basados en árboles combinados con un método de escalado que armoniza las escalas absolutas y relativas de las variables pueden mejorar la corrección general de la clasificación.

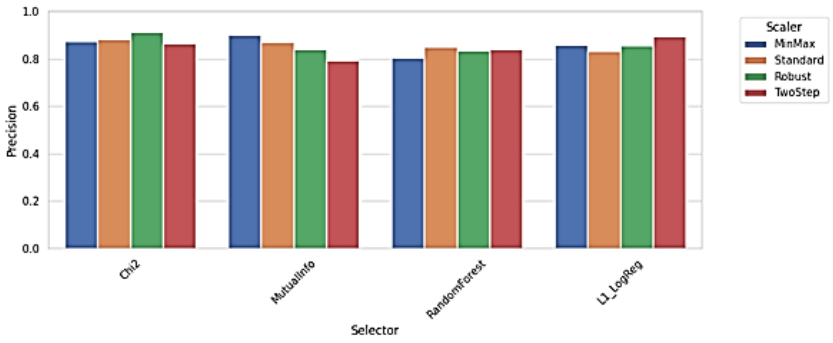


Fuente: elaboración propia

Figura 4 Comparación de exactitud entre métodos de selección y escalamiento.

En contraste, las combinaciones que incluyeron Información Mutua con RobustScaler obtuvieron la exactitud más baja, por debajo del 90%, lo que sugiere que en este conjunto de datos la Información Mutua no filtró suficientemente las variables menos relevantes, especialmente al emplearse con un escalador enfocado principalmente en mitigar valores atípicos. No obstante, en diagnósticos médicos, la exactitud por sí sola puede ocultar deficiencias críticas si el modelo falla en identificar correctamente los casos positivos.

La Figura 5 ilustra las puntuaciones de precisión, donde la combinación de Información Mutua con Min-Max Scaling sobresale, superando el 95%. Esto indica que este enfoque reduce eficazmente los falsos positivos, aunque podría no capturar todos los casos reales, una inquietud que se evidencia al analizar el recall.



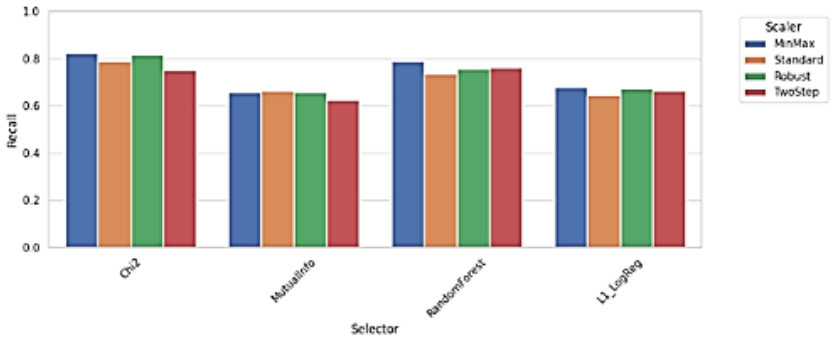
Fuente: elaboración propia

Figura 5 Comparación de precisión.

La Figura 6 destaca los valores de recall, que miden la sensibilidad del modelo para identificar pacientes con Alzheimer. Aquí, Random Forest con normalización en dos pasos vuelve a liderar, alcanzando casi el 93%, lo que demuestra su solidez para minimizar falsos negativos, un objetivo esencial en la detección temprana del Alzheimer donde omitir un diagnóstico puede retrasar intervenciones clave. En contrapartida, aunque Mutual Information con Min-Max logró alta precisión, registró un recall significativamente inferior, reafirmando la existencia de un compromiso entre reducir falsos positivos y evitar dejar pasar verdaderos casos positivos.

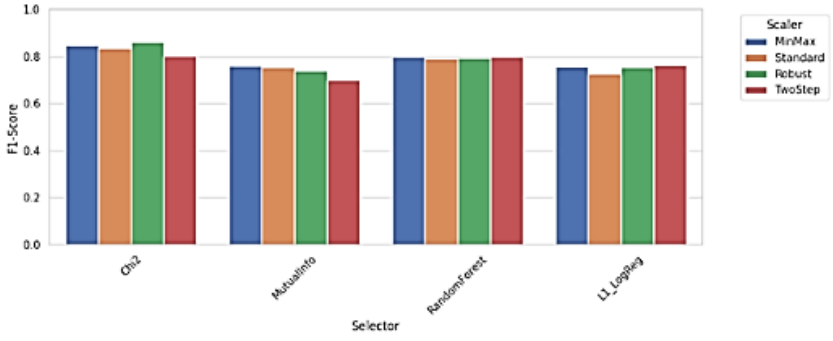
Las puntuaciones $F1$, presentadas en la Figura 7, que equilibran precisión y recall, enfatizan aún más este compromiso. El mayor $F1$ fue nuevamente obtenido por

Random Forest combinado con la normalización en dos pasos, seguido de cerca por la Regresión Logística con regularización $L1$ junto con StandardScaler. Esto sugiere que regularización $L1$, combinada con un escalado estándar, permite aislar señales predictivas relevantes manteniendo un balance entre identificar casos positivos y reducir alarmas falsas. La Figura 8 presenta el AUC, ofreciendo una perspectiva global del rendimiento de clasificación a través de distintos umbrales. Las combinaciones que incluyeron Random Forest o $L1$ mostraron consistentemente valores de AUC superiores a 0.93, respaldando su utilidad en marcos de decisión clínica donde sensibilidad y especificidad deben optimizarse conjuntamente. Finalmente, la Figura 9 consolida estos hallazgos en un mapa de calor comparativo que refuerza visualmente cómo los pipelines que emplearon Random Forest o $L1$ junto con métodos de normalización robustos como el enfoque en dos pasos o StandardScaler ofrecieron resultados más equilibrados y sólidos en todas las métricas evaluadas.



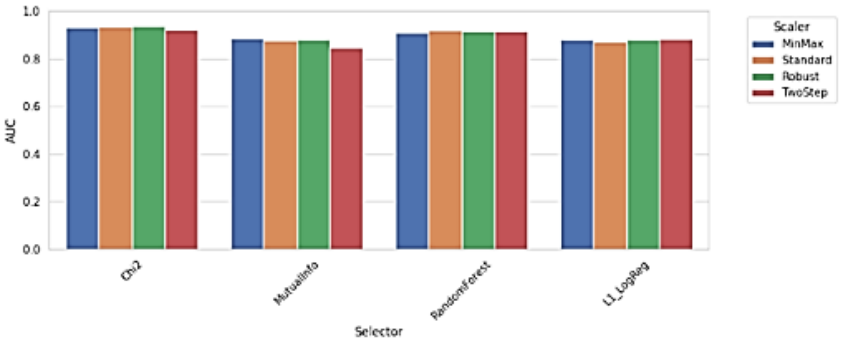
Fuente: elaboración propia

Figura 6 Comparación de Recall.



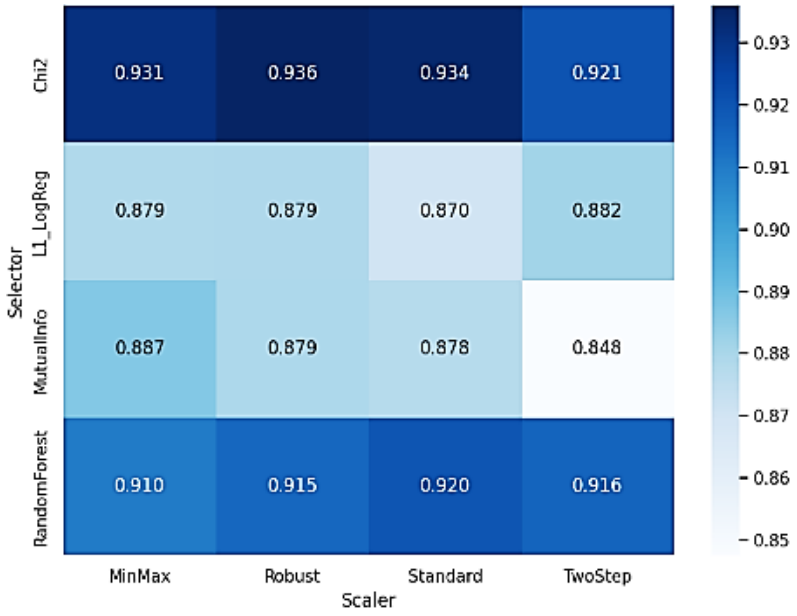
Fuente: elaboración propia

Figura 7 Evaluación F1-Score.



Fuente: elaboración propia

Figura 8 Comparación del AUC.



Fuente: elaboración propia

Figura 9 Mapa de calor de puntuaciones de precisión.

En conjunto, estos resultados muestran que, contar con pipelines de preprocesamiento robustos, no es un simple ajuste técnico, sino un requisito indispensable para construir modelos precisos y clínicamente significativos en el diagnóstico del Alzheimer. Al combinar sistemáticamente técnicas de selección de características que aíslan variables realmente informativas con métodos de escalado de datos, este estudio demuestra un camino claro para mejorar tanto la sensibilidad diagnóstica como la confiabilidad general del modelo, contribuyendo directamente al desarrollo de sistemas de apoyo a la decisión en enfermedades neurodegenerativas.

4. Discusión

Los hallazgos de este estudio evidencian con claridad que la elección del pipeline de preprocesamiento ejerce una influencia determinante sobre el éxito predictivo de un modelo MLP aplicado a datos clínicos para el diagnóstico del Alzheimer.

Las diferencias de rendimiento entre combinaciones no fueron triviales; en métricas como F1-score y AUC se observaron mejoras superiores al 10% simplemente al modificar la técnica de selección de características o el método de escalado, lo que subraya que las decisiones de preprocesamiento son tan cruciales como la propia arquitectura del modelo.

En esta evaluación comparativa, las técnicas de selección basadas en Chi^2 y en la importancia de RF mostraron consistentemente un desempeño superior. Esto probablemente se deba a su capacidad inherente para identificar características con fuertes asociaciones directas o indirectas con la variable de diagnóstico. En particular, Bosque Aleatorio, al emplear un ensamble de árboles, capturó eficazmente interacciones no lineales y priorizó atributos que más contribuyeron a reducir la impureza, algo fundamental dado el carácter multifactorial del Alzheimer.

En contraste, aunque la Regresión Logística con regularización $L1$ es conocida por inducir esparsidad, presentó rendimientos más bajos en varias métricas. Una explicación aceptable es que Regresión Logística $L1$ tiende a seleccionar de forma arbitraria entre predictores correlacionados, fenómeno común en conjuntos de datos clínicos, lo que puede llevar a excluir variables complementarias necesarias para caracterizar completamente el estado del paciente.

En cuanto a la normalización de datos, tanto RobustScaler como StandardScaler se adaptaron bien la presencia de valores atípicos típicos en medidas clínicas reales. Llamativamente, la normalización en dos pasos, pese a su ventaja teórica de controlar secuencialmente la escala y la magnitud vectorial, no superó a métodos más simples. Esto podría atribuirse a la naturaleza heterogénea de los datos; cuando existen valores extremos junto con variables estrechamente agrupadas, una normalización demasiado sofisticada puede distorsionar gradientes clínicos relevantes, disminuyendo la sensibilidad. Especialmente revelador fue el comportamiento de la selección basada en Información Mutua.

Si bien MI es eficaz para capturar dependencias complejas y potencialmente no lineales, no considera de manera intrínseca la redundancia entre variables. Esto podría explicar su recall comparativamente más bajo, pues podría favorecer atributos con información única que, sin embargo, no son necesariamente los más predictivos de forma aislada, conduciendo a modelos que omiten patrones sutiles, pero clínicamente cruciales. Estos compromisos son críticos en entornos médicos, donde no identificar verdaderos casos positivos tiene consecuencias graves, especialmente en enfermedades neurodegenerativas con ventanas de intervención temprana limitadas.

Es destacable que las pruebas con mejor desempeño no solo lograron una alta exactitud, sino que también mantuvieron valores de AUC superiores a 0.93, lo que indica una sólida capacidad de discriminación entre pacientes con y sin Alzheimer. Además, la selección recurrente de variables como los puntajes ADL, MMSE, quejas de memoria, colesterol HDL y evaluaciones funcionales en los modelos de mejor rendimiento refuerza su reconocida relevancia clínica para la detección temprana.

En conjunto, estos resultados enfatizan la necesidad de evaluar sistemáticamente diferentes estrategias de preprocesamiento, en lugar de asumir la superioridad de métodos complejos o novedosos sin una validación adecuada. Las técnicas tradicionales, bien implementadas y alineadas con las características específicas del conjunto de datos, dan lugar a modelos de IA que no solo son robustos técnicamente, sino también clínicamente significativos para apoyar en el diagnóstico temprano del Alzheimer.

5. Conclusiones

Este estudio confirma que las decisiones relacionadas con la selección de características y la normalización de datos tienen un impacto directo en el rendimiento de los modelos MLP aplicados al diagnóstico temprano de la enfermedad de Alzheimer. Entre las configuraciones evaluadas, las combinaciones que involucraron métodos de selección como Chi^2 y Bosque Aleatorio, junto con escaladores como RobustScaler y StandardScaler, demostraron resultados consistentes y efectivos en métricas clave como exactitud, sensibilidad y AUC. En

contraste, técnicas como la regularización $L1$ mostraron un rendimiento limitado dentro de este contexto clínico específico. Estos hallazgos resaltan la importancia de explorar sistemáticamente múltiples configuraciones de preprocesamiento en el desarrollo de sistemas de diagnóstico automatizado, ya que la elección de dichas técnicas puede ser tan crítica como la selección del propio modelo. Con base en estos resultados, se ha establecido una base sólida para una segunda fase de investigación centrada en el uso de datos clínicos sintéticos dentro de un esquema MLP.

La siguiente etapa tiene como objetivo evaluar si la integración de conjuntos de datos sintéticos puede mejorar la robustez y la capacidad de generalización del modelo, manteniendo o incluso superando los niveles de rendimiento alcanzados con datos reales.

6. Referencias y Bibliografía

[1] Aguayo, G. A., Zhang, L., Vaillant, M., Ngari, M., Perquin, M., Moran, V., Huiart, L., Krüger, R., Azuaje, F., Ferdynus, C., Fagherazzi, G., (2023). Machine learning for predicting neurodegenerative diseases in the general older population: A cohort study. BMC Medical Research Methodology, Vol. 23, No. 1, Article 1. https://doi.org/10.1186/s12874-023-01837-4.

[2] Cabanillas, M., Zapata, J., (2025). Evaluation of machine learning models for the prediction of Alzheimer's: In search of the best performance. Brain, Behavior, & Immunity Health, Vol. 44, 100957. https://doi.org/10.1016/j.bbih.2025.100957.

[3] Davuluri, R., (2020). A Survey of Different Machine Learning Models for Alzheimer Disease Prediction. International Journal of Emerging Trends in Engineering Research, Vol. 8, No. 7, Article 7. https://doi.org/10.30534/ijeter/2020/73872020.

[4] El-Sappagh, S., Alonso, J. M., Islam, S. M. R., Sultan, A. M., Kwak, K. S., (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Scientific Reports, Vol. 11, No. 1, 2660. https://doi.org/10.1038/s41598-021-82098-3.

[5] Gowri, G., Lun, X. K., Klein, A. M., Yin, P. Approximating mutual information of high-dimensional variables using learned representations, 2024.

[6] Pudjihartono, N., Fadason, T., Kempa, A. W., O'Sullivan, J. M., (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, Vol. 2. https://doi.org/10.3389/fbinf.2022.927312.

[7] Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., Razak, R. A., (2020). A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder. Brain Sciences, Vol. 10, No. 12, Article 12. https://doi.org/10.3390/brainsci10120949.

[8] Sharma, V., (2022). A Study on Data Scaling Methods for Machine Learning. International Journal for Global Academic & Scientific Research, Vol. 1, No. 1. https://doi.org/10.55938/ijgasr.v1i1.4.

[9] Silpa, N., Swain, S. K., Rao V. V. R., (2024). Revolutionizing feature engineering for robust ensemble machine learning by hybridizing MRMR insight and CHI2 independence. Proceedings on Engineering Sciences, Vol. 6, No. 3, pp. 1337–1348. https://doi.org/10.24874/PES.SI.24.03.017.

[10] Speiser, J. L., (2021). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. Journal of Biomedical Informatics, Vol. 117, 103763. https://doi.org/10.1016/j.jbi.2021.103763.

[11] Sree, K. D., Bindu, C. S., (2018). Data analytics: Why data normalization. International Journal of Engineering and Technology (UAE), Vol. 7, No. 4.6 (Special Issue 6). https://doi.org/10.14419/ijet.v7i4.6.20464.

[12] Venkatesh, B., Anuradha, J., (2019). A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies, Vol. 19, No. 1, pp. 3–26. https://doi.org/10.2478/cait-2019-0001.

[13] Zhou, H., Wang, X., Zhang, Y., (2024). Feature selection based on weighted conditional mutual information. Applied Computing and Informatics, Vol. 20, No. 1–2, Article 1–2, https://doi.org/10.1016/j.aci.2019.12.003.