Avances en Interacción Humano-Computadora, 2025, Volumen 10, No. 1, pp. 12-22
DOI: 10.47756/aihc.y10i1.191

RESEARCH PAPER

Optimización e Interpretabilidad de Modelos Clasificadores para Alzheimer: Un Estudio Basado en el Conjunto de Datos OASIS-2

Optimization and Explainability of Classifier Models for Alzheimer's Disease: A Study Based on the OASIS-2 Dataset

Heber Zapata, Olanda Prieto-Ordaz, Raymundo Cornejo*

Published: 30 November 2025

Resumen
El diagnóstico temprano del Alzheimer es crucial para disminuir su progreso y mejorar la calidad de vida de los pacientes. En este estudio se evalúa el desempeño de tres modelos de aprendizaje automático (regresión logística, máquinas de soporte vectorial y bosques aleatorios) con el conjunto de datos OASIS-2 bajo dos condiciones: considerando únicamente la primera visita de cada participante e incluyendo todas las mediciones longitudinales. Se aplicó un *pipeline* de preprocesamiento, validación cruzada y búsqueda en malla para optimizar los modelos. Los modelos entrenados con el conjunto de datos ampliado alcanzaron una exactitud del 96%, superando resultados reportados en la literatura. Se utilizaron técnicas de explicación, incluyendo *K-Means* y el método de explicaciones locales, interpretables y agnósticas al modelo; para analizar las instancias mal clasificadas, revelando que varios errores se debieron a pacientes "convertido" mal etiquetados más que a deficiencias del modelo. Estos hallazgos destacan que los modelos son sensibles a la calidad de los datos y a la consistencia en el etiquetado. Los resultados enfatizan la necesidad de procedimientos rigurosos de recolección y análisis de datos para garantizar la equidad y la aplicabilidad clínica de los modelos. Para trabajo futuro se sugiere un enfoque en la construcción de conjuntos de datos longitudinales más representativos y explorar técnicas adicionales de explicación para reducir el sesgo y mejorar la confiabilidad de los sistemas de diagnóstico temprano.

Zapata, H., Prieto-Ordaz, O., Cornejo, R.
Universidad Autónoma de Chihuahua
Chihuahua, Chih., México
Email: {p329454, oordaz, rcornejo}@uach.mx

* Corresponding author

Palabras clave:
Alzheimer, aprendizaje automático, aplicaciones en salud, OASIS-2

Abstract
Early diagnosis of Alzheimer's disease is crucial to slowing its progression and improving patients' quality of life. This study evaluates the performance of three machine learning models—logistic regression, support vector machine, and random forest—using the OASIS-2 dataset under two conditions: considering only the first visit of each participant and including all available longitudinal measurements. A standardized preprocessing pipeline, cross-validation, and grid search were applied to optimize the models. Models trained on the extended dataset achieved 96% accuracy, outperforming previously reported results. Explainability techniques, including K-Means clustering and local interpretable model-agnostic explanations, were applied to analyze misclassified instances, revealing that several errors were caused by mislabeled "converted" patients rather than model deficiencies. These findings highlight that classification performance is sensitive to data quality and labeling consistency. The results emphasize the need for rigorous data collection and curation procedures to ensure fairness and clinical applicability of predictive models. Future work should focus on building more representative longitudinal datasets and exploring additional explainability techniques to reduce potential bias and enhance the trustworthiness of early diagnostic systems.

Key words:
Alzheimer, machine learning, healthcare applications, OASIS-2

1 Introducción
El Alzheimer es un trastorno neurológico progresivo que afecta al cerebro, manifestándose en problemas de memoria, funciones ejecutivas y orientación espacial, entre otros [1]. Actualmente, no se dispone de una cura, lo que genera un impacto considerable en los cuidadores debido a la progresión constante de la enfermedad [2], [3]. El aumento en la prevalencia del Alzheimer está vinculado

al crecimiento demográfico y al incremento en la esperanza de vida, afectando principalmente a personas mayores de 65 años [4].

Una estrategia importante para enfrentar esta situación es la detección temprana de la enfermedad, que puede contribuir a controlar los síntomas, ralentizar la progresión de la enfermedad y mejorar la calidad de vida de los afectados [5], [6]. En años recientes, el aprendizaje automático (por sus siglas en inglés, M*achine Learning*, ML) ha mostrado gran potencial en el diagnóstico médico, representando una herramienta prometedora para este desafío [5]. Sin embargo, el uso de estos modelos en el dominio de salud aún, y por consecuencia en el de Interacción Humano Computadora (IHC), es un área de oportunidad debido a los problemas de interpretación, explicación, y equidad. Inclusive, la inconsistencia o falta de información en los conjuntos de datos puede llevar a que los algoritmos en ML tengan sesgos que afectan los diseños en el ámbito de IHC. Recientemente, son cada vez más los diseños tecnológicos en IHC que emplean herramientas o técnicas en ML usando conjuntos de datos y modelos con sesgos e inequidad [7]. Por este motivo, el presente estudio tiene como objetivo mostrar cómo los algoritmos tradicionales de ML pueden ser útiles en el diagnóstico temprano del Alzheimer, y a su vez mantener una equidad reduciendo posibles limitaciones en el preprocesamiento del conjunto de datos empleados.

El presente trabajo se inspira en el trabajo de Khan y Zubair [8] quiénes realizaron un análisis del conjunto de datos "*Open Access Series of Imaging Studies* (OASIS-2)"[1]. El objetivo de este análisis fue encontrar correlaciones entre diversas características y la presencia o ausencia de Alzheimer en los pacientes. Los autores emplearon estadísticas descriptivas y análisis de correlación de las variables. Entre sus hallazgos, destaca que a un mayor nivel educativo y socioeconómico se asocia a una menor probabilidad de Alzheimer. También observaron que el grupo sin demencia presentaba un mayor volumen cerebral en comparación con el grupo con demencia. Adicionalmente, encontraron que el grupo sin demencia obtuvo puntuaciones más altas en *Mini-Mental State Examination* (MMSE) [9], *Clinical Dementia Rating* (CDR)[10] y *Atlas Scaling Factor* [11] con la presencia de demencia.

No obstante, las métricas utilizadas para evaluar el desempeño de los modelos no garantizan una aplicabilidad en la vida real. Esto se debe a que los modelos son precisos en métricas de clasificación, pero carecen de mecanismos adecuados de explicación, interpretación y equidad; esto limita su utilidad en contextos clínicos [12]. Por ejemplo, la falta de equidad puede reproducir sesgos que no solo afectan al desempeño del modelo, sino que también propicia la vulnerabilidad de los sectores poblacionales con Alzheimer. Por otro lado, la falta de explicación e interpretación de los modelos genera una desconfianza por parte del personal de salud [13].

La interacción de estos mecanismos afecta directamente la adopción de tecnologías integradas con ML en el ámbito clínico. En este contexto, es necesario definir los mecanismos involucrados en la adopción de tecnologías que integran ML para la toma de decisiones dentro del área clínica. La interpretación se refiere a la posibilidad de que un humano pueda analizar y entender las predicciones (*outputs)* de un modelo. Por otro lado, la explicación es el uso de técnicas adicionales a las implementadas en el modelo para comprender los *outputs* [14]. Por último, la equidad implica un aspecto más amplio. Para entrenar a un modelo se necesitan datos de entrada, estos datos pueden estar sesgados por diferentes razones, entre ellas la discriminación de grupos vulnerables. Adicionalmente, el preprocesamiento que se le da a los datos también puede generar un sesgo que afecte a sectores poblacionales vulnerables [15].

2 Trabajo relacionado

A medida que ML se integra en el diagnóstico médico, surgen preguntas sobre su comportamiento contextos reales. Especialmente, el uso de la base de datos OASIS-2 ha permitido entrenar modelos de ML con resultados prometedores en métricas de clasificación, aunque sin garantizar aspectos como la equidad o la interpretabilidad. En esta sección se revisan estudios que han utilizado esta base de datos, se analizan casos donde los sesgos han afectado el diseño de prototipos clínicos, y se presenta el método de explicaciones locales, interpretables y agnósticas al modelo (por sus siglas en inglés, *local interpretable model-agnostic explanations*, LIME) y *K-means* como herramientas para mitigar falta de explicación y equidad en modelos de ML.

2.1 Algoritmos ML y Alzheimer

Autores han llegado a conclusiones similares a Khan y Zubair [16]. A partir de un análisis de las principales características (PCA), seleccionaron MMSE, CDR, MR *delay* (tiempo de retardo antes de la obtención de la imagen en tiempo real) y WBV (resultado normalizado del volumen cerebral total); luego entrenaron un modelo de máquinas de soporte vectorial (por sus siglas en inglés, *support vector machine*, SVM). Su objetivo fue optimizar los parámetros de SVM y mejorar la precisión de clasificación en comparación a [17], quienes utilizaron la base de datos *Alzheimer's Disease Neuroimaging Initiative* (ADNI) y entrenaron un modelo SVM con un *kernel* de función de base radial. En el año 2019, Battineni, Chintalapudi y Amenta, determinaron que los parámetros óptimos para SVM son 1.0E-4 y 100, para gamma y regularización (C) respectivamente. Las métricas del modelo alcanzaron un 70% para exactitud (o *accuracy* en inglés), 65% para sensibilidad (o *recall* en inglés) y 82% en precisión (o *precision* en inglés).

Por otro lado, en [18] compararon el desempeño de diferentes algoritmos de ML para la predicción de Alzheimer con la base de datos de OASIS. En este estudio se implementaron y evaluaron los modelos de SVM, regresión logística (por sus siglas en inglés, logistic regression, LR), árbol de decisión (por sus siglas en inglés, *decision tree*, DT) y bosques aleatorios (por sus siglas en inglés, *random forest*, RF). Posteriormente se obtuvieron los mejores parámetros mediante la función de búsqueda en cuadrícula (por sus siglas en inglés, *greed search*, GS) y validación cruzada (*cross-validation*) para cada modelo y se evaluaron nuevamente. El modelo que reportó mejor desempeño fue SVM con ajuste, aunque no mencionan los mejores parámetros encontrados. Las métricas fueron 92% para accuracy, y 91% para recall y área bajo la curva (por sus siglas en inglés, *area under the curve*, AUC).

Mientras que los artículos previos no reportan de manera explícita todas las estrategias que se utilizaron para el procesamiento de datos, en [19] si se reporta un flujo de trabajo (*pipeline)* que abarca desde la recolección de datos y su preprocesamiento hasta la evaluación de los modelos utilizados. En este artículo, los autores describen paso a paso los métodos utilizados para la recolección de datos, el procesamiento, el entrenamiento de los modelos y su evaluación. El *pipeline* se

[1] https://sites.wustl.edu/oasisbrains/

dividió en 5 pasos: 1) la preparación de los datos, que incluye la recopilación de datos, su visualización, la selección de características y su transformación. 2) La separación de los datos, donde los datos se dividen en conjuntos de entrenamiento, prueba y validación. 3) el modelado, que implica el entrenamiento, la evaluación, la validación cruzada y el ajuste de hiper parámetros de varios algoritmos de ML. 4) La predicción, que incluye la generación de predicciones de cada modelo con el conjunto de datos de prueba. 5) La evaluación del modelo, que es la obtención de las métricas para cada modelo. El modelo con mejores métricas reportado fue RF, con 86.84% de *accuracy*, 80% de *recall*, 88% para *specificity*, 94.11% de *precisión* y 87.22% para AUC.

Los métodos implementados anteriormente, si bien reportan métricas relevantes, no aseguran la equidad de los modelos. En la siguiente sección se presentan casos de modelos entrenados con sesgos y cómo esto impactó en el diseño de prototipos.

2.2 Casos de inequidad en prototipos

En el dominio de salud existen intentos por implementar prototipos basados en ML para el diagnóstico de distintas enfermedades. A continuación, se describen dos casos en los que el uso de modelos de ML entrenados con datos sesgados generó un impacto negativo en la generación de prototipos para el diagnóstico clínico.

DermaScan[2] es una aplicación basada en ML para generar diagnósticos de lesiones en la piel. Aunque es una aplicación que apoya al médico al generar un diagnóstico, los creadores de DermaScan aseveran que el modelo tiene un sesgo racial. El sesgo se debe a que el conjunto de imágenes con el que fue entrenado, DermaScan (HAM10000), contiene menos del 5% de imágenes de personas con tono oscuro de piel. Por lo tanto, el modelo no solo falla al identificar a personas con tono oscuro de piel, sino que el sesgo provoca que condiciones dermatológicas propias de esta población no sean consideradas en la clasificación [20].

Un ejemplo similar es la implementación de un algoritmo para predicción de costos de atención médica, ampliamente utilizado en el sistema de atención médica estadounidense para identificar y ayudar a pacientes con necesidades de salud complejas [21]. Los sistemas de salud y las aseguradoras dependen de este tipo de algoritmos para dirigir a los pacientes a programas de "manejo de atención de alto riesgo". Se estima que esta clase de herramientas de predicción de riesgo se aplica a aproximadamente 200 millones de personas en los Estados Unidos cada año [21]. El sesgo de este algoritmo es predecir los costos futuros de atención médica en lugar de la enfermedad o las necesidades de salud. Debido al acceso desigual a la atención, se gasta menos dinero en el cuidado de las personas con tono oscuro de piel en comparación a las personas con tono claro de piel, incluso con el mismo nivel de necesidad de salud. Así, la predicción precisa de costos resulta en una subestimación de la enfermedad en personas con tono oscuro de piel, lo que, si se corrigiera, aumentaría la fracción de personas negras que reciben ayuda adicional del 17.7% al 46.5% [21].

Los casos presentados en esta sección evidencian cómo los sesgos en los datos pueden generar impactos negativos en aplicaciones de salud basadas en ML, afectando la equidad y la confianza de los profesionales clínicos. Esta problemática resalta la necesidad de incorporar métodos que permitan interpretar y explicar las decisiones de los modelos, con el fin de facilitar su adopción en la práctica médica.

2.3 *K-means* y LIME

Como se mostró en los casos, la aplicabilidad de ML al dominio médico aún es un tema controversial. Los expertos en salud señalan que sin explicaciones claras, es poco probable que ML se convierta en parte de la práctica médica rutinaria [22].

Para abordar la problemática de las explicaciones poco claras, en artículos anteriores, se ha utilizado LIME [23], [24]. Al proporcionar explicaciones locales e intuitivas para predicciones individuales, LIME ayuda a reducir la brecha de confianza entre los sistemas de Inteligencia Artificial (IA) y los profesionales de salud, mejorando así la IHC, y potencialmente los resultados del paciente.

Se ha aplicado LIME con éxito en diversos dominios médicos, incluido el diagnóstico de la enfermedad de Parkinson [25], la evaluación del riesgo de diabetes [26], y el análisis de imágenes de cáncer de mama [27]. En el diagnóstico temprano del Alzheimer, se ha implementado con diferentes tipos de datos, incluidas las imágenes médicas [28], el análisis de expresión genética [29] y la interpretación de datos clínicos multimodales [30] . LIME ayuda a los médicos a comprender las decisiones del modelo al generar explicaciones interpretables localmente, lo cual es fundamental para generar confianza en las herramientas de diagnóstico impulsadas por IA [31].

Adicionalmente, los métodos de aprendizaje no supervisado, como *K-means*, pueden ayudar a identificar patrones ocultos y simplificar conjuntos de datos de atención médica complejos, lo que facilita una toma de decisiones más transparente [32]. Específicamente, *K-means* ha sido utilizado para detectar anomalías en datos clínicos, por ejemplo, para localizar valores atípicos en la predicción de enfermedades cardíacas mediante la evaluación de puntos de datos distantes de los centroides de los clústeres [33]. Otro estudio aplicó la agrupación combinatoria de *K-means* para separar a los pacientes con diabetes en grupos clínicos distintos, lo que demuestra su potencial para el reconocimiento visual de patrones [34].

Por lo tanto, en el presente trabajo se realizó un análisis mediante *K-Means* y LIME para obtener un grado de explicación de las instancias mal clasificadas; además, se analizaron las instancias mal clasificadas para comprender el porqué de su errónea clasificación y brindar un mayor grado de equidad en modelos futuros. La interpretación de los modelos fue un aspecto que no se abordó en este estudio.

3 Materiales y métodos

En esta sección se incluye una descripción detallada del conjunto de datos longitudinal OASIS-2, el cuál fue utilizado considerando dos estrategias para conformar diferentes conjuntos de datos. La primera versión incluye 150 instancias del conjunto de datos mientras que la segunda versión incluye 373 instancias. Posteriormente, se describe el preprocesamiento realizado al conjunto de datos que incluye la normalización y partición de datos en entrenamiento y prueba. A continuación, se incluyen los modelos de ML seleccionados especificando los hiperparámetros utilizados, así como la configuración de cada uno de los experimentos realizados a lo largo de este estudio. Finalmente, se incluye el proceso de evaluación realizado a cada uno de los modelos implementados complementando con un análisis e interpretación de los resultados mediante técnicas de visualización y explicabilidad como *K-Means* y LIME.

[2] https://dermascanai.com/es/

3.1 Base de datos

OASIS-2 es un conjunto de datos recolectados de manera longitudinal de 150 participantes, con edades entre 60-96 años. Cada participante fue escaneado en dos o más visitas, separadas por al menos un año para un total de 373 sesiones de imagen. Para cada participante, se incluyen 3 o 4 exploraciones individuales de resonancia magnética ponderada en T1 obtenidas en sesiones únicas de exploración. Todos los sujetos son diestros e incluyen tanto hombres como mujeres. 72 de los participantes se caracterizaron como no dementes a lo largo del estudio. 64 de los participantes se caracterizaron como dementes en el momento de sus visitas iniciales y siguieron siéndolo en las exploraciones posteriores, incluidos 51 individuos con enfermedad de Alzheimer de leve a moderada. Otros 14 participantes fueron caracterizados como no dementes en el momento de su visita inicial y posteriormente fueron caracterizados como dementes en una visita posterior [11]. La Tabla 1 muestra una breve descripción del contenido de cada columna de OASIS-2.

Tabla 1. Descripción de las columnas del conjunto de datos OASIS.

Columna	Descripción
Subject ID	Identificador del paciente
MRI ID	Identificador de las imágenes del paciente
Group	*Demented*, *nondemented*, o *converted*
Visit	Número de visitas de cada paciente
MR Delay	Tiempo de retardo dado antes de la obtención de la imagen en tiempo real
M/F	Género
Hand	Mano predominante
Age	Edad al momento de la toma de la muestra
EDUC	Nivel educativo
SES	Nivel socioeconómico
MMSE	Puntuación en el *Mini-Mental State Examination*
CDR	Puntuación en *Clinical Dementia Rate*
eTIV	Resultado estimado del volumen intracraneal total
nWBV	Resultado normalizado del volumen cerebral total
ASF	*Atlas Scaling Factor*

3.2 Preprocesamiento

El preprocesamiento de los datos, así como el entrenamiento de los modelos y los análisis realizados en este trabajo se realizaron en un entorno de Google Colab y se utilizó Python como lenguaje de programación. Para la manipulación de estructuras de datos se emplearon Numpy y Pandas; adicionalmente, se utilizaron librerías de scikit-learn para el entrenamiento y análisis de datos. Por último, se emplearon matplotlib y seaborn para generar las figuras presentadas en la sección de resultados. Resulta relevante añadir que el acceso a la base de datos OASIS-2 se hizo desde su página oficial y se firmó digitalmente la solicitud de acceso a la base de datos.

OASIS-2 reporta datos de 150 pacientes. La recolección de datos de esta base de datos fue de manera longitudinal. Sin embargo, es importante señalar que los pacientes tuvieron diferentes puntos de medición. En total se tienen 373 datos, que son distintos puntos de medición por paciente, pero hay pacientes que solo cuentan con una visita y hay pacientes hasta con 4 visitas. Por lo tanto, en la literatura revisada solo se utiliza la primera visita de cada paciente para entrenar los modelos [8], [19]. Esta estrategia se toma debido a que no todos los pacientes tienen más de un punto de medición y eso podría sesgar los resultados. Sin embargo, explorar cómo se comportan los modelos al utilizar todos los puntos de medición nos puede indicar cómo se comportarían los algoritmos con una cantidad de datos mayor a la usada en la literatura. Por lo tanto, se comparará el desempeño de diferentes algoritmos de ML entrenados con los datos de la primera visita (150 datos) contra el desempeño de los algoritmos entrenados con todos los puntos de medición (373 datos), tomando en cuenta cada punto de medición como un paciente distinto.

Similar a [19], el preprocesamiento se dividió en 2 etapas: preparación de los datos y separación de los datos. La preparación de los datos consiste en: 1) ingesta de datos: obtener acceso al conjunto de datos de manera ética, descarga del conjunto de datos y lectura de este en *jupyter notebook*. 2) Visualización de datos: análisis visual de los datos con fin exploratorio. Se realizó identificación de valores atípicos, revisión de la distribución de los datos y la correlación entre características. 3) Transformación de los datos: se crearon dos funciones, una para preparar el conjunto de datos con 150 pacientes y otra para crear el conjunto de datos de 373 pacientes, además, estas funciones eliminan columnas innecesarias ('Subject ID', 'MRI ID', 'Visit' y 'Hand'), a los pacientes etiquetados como 'Converted' los cambia a 'Demented' y convierte la columna objetivo ('Group') en una variable binaria ('Nondemented': 0, 'Demented': 1). Luego se generó un *pipeline* que para las variables numéricas maneja los datos faltantes imputándolos con la media, y escala los datos aplicando el método de *StandardScaler*. Para las variables no numéricas las representa en números con el método *OneHotEncoder*.

Para la separación de los datos se mantuvo la metodología de [19] (véase Figura 1), utilizando una proporción 80% - 20% para entrenamiento y prueba, respectivamente. Se utilizó el método *train_test_split* con los parámetros *random_state* = 42 y *stratify* de acuerdo con la etiqueta.

3.3 Modelos

Guiados por la bibliografía revisada se escogieron dos modelos, RF [19] y SVM [16], [18]; además, se implementó un modelo de LR debido a su bajo coste computacional. Se utilizó un diseño factorial 2 (conjunto de datos 150 y conjunto de datos 373) x 3 (SVM, RF y LR) por lo que cada modelo se entrenó con la base de datos de 150 pacientes y con la base de datos de 373 pacientes, quedando un cruce como el que se muestra en la Tabla 2. Se utilizó un diseño 2x3 con el objetivo de analizar cómo las métricas de los modelos (variables dependientes), varían en función de la interacción entre el modelo utilizado y el número de instancias con las que fue entrenado.

Para cada modelo se utilizó el método *KFold* con los parámetros *n_splits*=5, *shuffle*=True, *random_state*=42. Luego, se utilizó *GridSearchCV* para encontrar los mejores parámetros de cada modelo, con los siguientes parámetros: *param_grid*, *cv=kfold*, *scoring='accuracy'*. Donde *kfold* es el método KFold establecido y el *param_grid* de cada modelo se describe en las tablas 3,4 y 5.

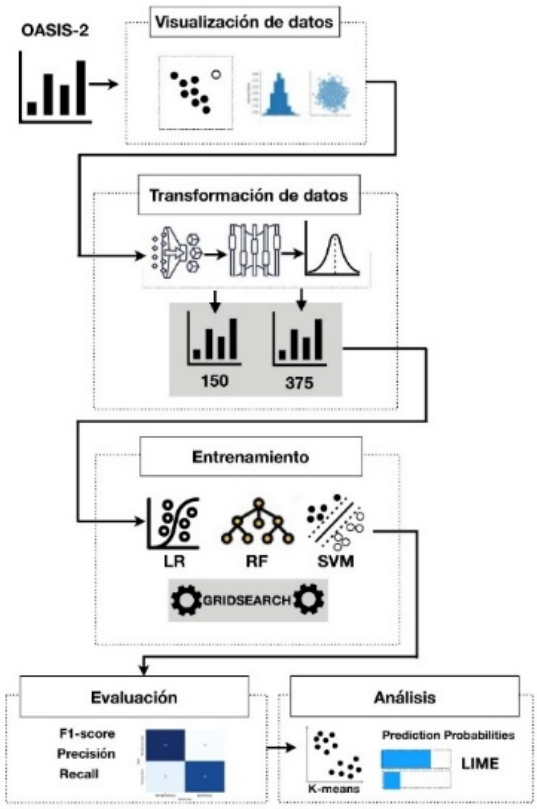

Figura 1. Esquema general del procedimiento.

Tabla 2. Diseño experimental.

		Modelos		
		SVM	RF	LR
Conjunto de datos	150	SVM_150	RF_150	LR_150
	373	SVM_373	RF_373	LR_373

Tabla 3. *Param_grid* de SVM.

C	[0.1, 1, 10, 100]
kernel	['linear', 'poly', 'rbf', 'sigmoid']
degree	[2, 3, 4]
gamma	['scale', 'auto']

Tabla 4. *Param_grid* de RF.

n_estimators	[100, 200, 300]
criterion	['gini', 'entropy']
max_depth	[None, 10, 20, 30]
max_features	[None, 'sqrt', 'log2']

Tabla 5. *Param_grid* de LR.

C	[0.001, 0.01, 0.1, 1, 10, 100]
penalty	['l1', 'l2']
solver	['liblinear', 'saga']
max_iter	[1000, 5000, 10000, 15000]

3.4 Métricas

Cada modelo fue evaluado con *accuracy*, que mide la proporción de aciertos sobre el total de muestras; *precision*, que indica cuántos de los casos predichos como positivos realmente lo son; *recall*, que refleja la capacidad del modelo para detectar todos los casos positivos reales; F1-score, que equilibra *precision* y *recall* mediante su promedio armónico; y Fβ-score con β = 2, que prioriza el *recall* sobre la *precision*, útil cuando es más importante identificar correctamente los casos positivos incluso si hay más falsos positivos.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

3.5 *K-means* y LIME

Después de obtener las métricas, se identificaron las instancias clasificadas como falsos positivos en cada modelo. Posteriormente, se aplicó *K-means* con el objetivo de analizar cómo se agrupan estos casos, evaluando si presentan patrones comunes que pudieran explicar su incorrecta clasificación.

K-means se implementó para agrupar a los participantes en función de las variables preprocesadas. Para determinar el número óptimo de *clústeres* se evaluaron diferentes valores de k (2 a 10) mediante el método del codo (inercia o suma de errores cuadrados; por sus siglas en inglés, *sum of squared errors*, SSE) y el índice de *silhouette*. El modelo final se entrenó con k = 2, con los parámetros *random_state* = 42 y *n_init* = 10 para asegurar la reproducibilidad y la estabilidad de la solución.

Por último, se implementó LIME con el objetivo de obtener explicaciones locales de las predicciones de cada modelo entrenado (LR_150, SVM_150, RF_150, LR_373, SVM_373 y RF_373). Para ello, se utilizó *LimeTabularExplainer* con los datos de entrenamiento preprocesados, especificando los nombres de las variables y de las clases ("*Nondemented*" y "*Demented*"), y el parámetro *mode* = "classification". Para el análisis se seleccionaron las instancias clasificadas como falsos positivos por cada modelo, generando explicaciones sobre la contribución de cada característica a dichas predicciones erróneas.

3.6 Análisis manual de instancias

Finalmente, para complementar el análisis de las instancias clasificadas como falsos positivos, se realizó una revisión manual

de cada caso con el fin de identificar posibles patrones de manera visual. Este procedimiento buscó aportar elementos adicionales que permitieran reconocer sesgos y favorecer un preprocesamiento más equitativo en futuros modelos.

4 Resultados

Esta sección presenta los resultados derivados del diseño experimental, iniciando con la especificación de los hiperparámetros óptimos determinados para cada uno de los seis modelos mediante *grid search*. A continuación, se detalla el desempeño de los modelos LR, SVM y RF en ambos conjuntos de datos (150 y 373 instancias) a través de un análisis comparativo de métricas (*Accuracy, Precision, Recall*, F1-*score* y Fbeta-*score* con beta = 2), contrastando los hallazgos con la literatura previa. Finalmente, se aborda la interpretación y el análisis de errores mediante los resultados de la agrupación de falsos negativos con *K-means* y las explicaciones de predicción local proporcionadas por LIME, complementados con un análisis manual que identifica el origen de las clasificaciones erróneas.

4.1 Mejores parámetros

Las tablas 6, 7 y 8, muestran los mejores parámetros obtenidos para cada modelo a partir de la búsqueda realizada con *GridSearchCV*.

Tabla 6. Mejores parámetros para SVM.

Modelo	C	kernel	degree	gamma
SVM_150	0.1	2	scale	linear
SVM_373	0.1	2	scale	linear

Tabla 7. Mejores parámetros para RF.

Modelo	n_estimators	criterion	max_depth	max_features
RF_150	entropy	None	None	200
RF_373	gini	None	sqrt	200

Tabla 8. Mejores parámetros para LR.

Modelo	C	penalty	solver	max_iter
LR_150	10	1000	11	liblinear
LR_373	0.1	1000	11	liblinear

4.2 Métricas

La Tabla 9 muestra una comparativa de los resultados obtenidos con los seis modelos entrenados y los mejores modelos reportados en la literatura revisada. En los encabezados de la tabla, se utilizan las siguientes abreviaturas: *Accuracy* –> Acc, *Precision* –> P, y *Recall* -> R, Fβ-*score* -> F2. Asimismo, se incluyen las métricas utilizadas por cada autor y las aplicadas en este trabajo. Luego, en las Figuras 2,3 y 4 se incluyen las matrices de confusión de cada modelo entrenado.

En la Tabla 9 se presentan los resultados de los modelos entrenados con el subconjunto de datos de 150 participantes, junto con los modelos de estudios previos. Esto, con el objetivo de mostrar que nuestros modelos son consistentes con el estado del arte. Lo que sugiere que se siguió una metodología alineada con la literatura y, además, confirma que nuestros análisis se basan en modelos semejantes a los del estado del arte. Por otro lado, en la Tabla 10 se encuentra una comparación entre los modelos entrenados con el conjunto completo de datos de 373 participantes y el subconjunto de datos de 150 participantes. Se puede observar que el aumento de datos aumenta el desempeñó de los modelos.

Tabla 9. Comparativa con estado del arte.

	Acc	P	R	F1	F2	AUC
Battineni et al. (2019) SVM	.70	.82	.65	-	-	-
Antor et al. (2021) SVM	.92	-	.91	-	-	.91
Khan y Zubair (2022) RF	.86	.94	.80	-	-	.87
LR_150	.90	1	.81	.89	.84	-
SVM_150	.90	1	.81	.89	.84	-
RF_150	.86	.92	.81	.86	.83	-

Tabla 10. Comparativa de modelos 150 vs 373.

	Acc	P	R	F1	F2
LR_150	.90	1	.81	.89	.84
SVM_150	.90	1	.81	.89	.84
RF_150	.86	.92	.81	.86	.83
LR_373	.96	1	.91	.95	.93
SVM_373	.96	1	.91	.95	.93
RF_373	.96	1	.91	.95	.93

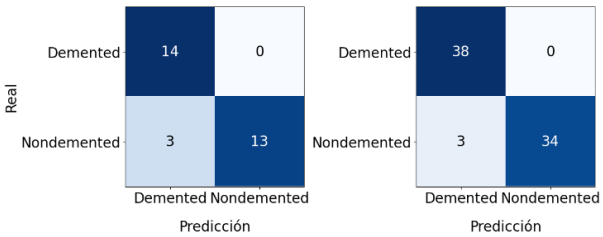

Figura 2. Matrices de confusión RF (la matriz del lado izquierdo pertenece al modelo SVM_150 y la del lado derecho al modelo SVM_373).

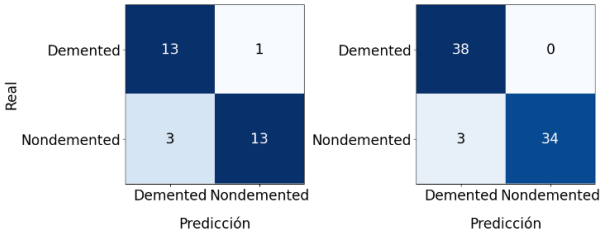

Figura 3. Matrices de confusión RF (la matriz del lado izquierdo pertenece al modelo RF_150 y la del lado derecho al modelo RF_373).

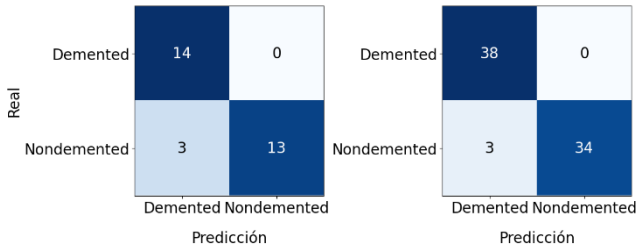

Figura 4. Matrices de confusión LR (la matriz del lado izquierdo pertenece al modelo LR_150 y la del lado derecho al modelo LR_373).

En la Figura 3, en la matriz de confusión del modelo RF_150 se observa un falso positivo. Este dato no se aborda en el presente estudio debido a que los falsos negativos son de mayor relevancia en el área clínica. Esta decisión también se refleja en el uso de la métrica Fβ-score con β = 2.

4.3 *K-means* y LIME

Como se observa en la Tabla 9, el *recall* de los modelos entrenados siempre es menor que la precisión, lo que indica la presencia de falsos negativos (FN). Además, en la misma tabla, el Fβ-score es menor que el F1-score, lo que también evidencia FN, dado que Fβ con β = 2 penaliza más los falsos negativos. Al identificar las instancias FN para cada modelo, se encontró un patrón: los modelos entrenados con el conjunto de datos de 150 participantes presentan las mismas instancias mal clasificadas, independientemente del modelo; de manera similar, los modelos entrenados con el conjunto de 373 participantes muestran las mismas instancias mal clasificadas, también independientemente del modelo.

A continuación, se presentan los resultados obtenidos con *K-Means* y LIME, mostrando la agrupación de los FN y las contribuciones de cada característica a estas predicciones, respectivamente. Estos análisis permiten identificar patrones y comprender de manera más detallada los errores de clasificación de los modelos.

Las Figuras 5 y 6 muestran la visualización de la agrupación de participantes obtenida con *K-Means* tras reducir la dimensionalidad de los datos mediante PCA. Para ello, se aplicó PCA sobre los conjuntos de datos preprocesados (150 pacientes y 373 pacientes), reduciendo las variables originales a dos componentes principales, lo que permitió representar cada instancia en un plano 2D. Posteriormente, se entrenó *K-Means* con el número de clústeres previamente determinado y se asignaron las etiquetas de grupo a cada paciente. En las figuras 5 y 6 se representan los pacientes coloreados según su clúster asignado y los centroides de cada grupo, mientras que las instancias clasificadas como FN se resaltan con un color distinto para identificar visualmente los patrones de error dentro de los clústeres.

Por último, se implementó la técnica de LIME con el objetivo de generar explicaciones locales para cada modelo entrenado (LR_150, SVM_150, RF_150, LR_373, SVM_373 y RF_373). Para ello, se utilizó la librería *LimeTabularExplainer* con los datos de entrenamiento preprocesados, especificando los nombres de las variables y de las clases ("*Nondemented*" y "*Demented*"), y el parámetro *mode* = "*classification*". Para este análisis, se seleccionaron las instancias clasificadas como FN, es decir, casos cuya etiqueta corresponde a "*Demented*", pero que los modelos predicen como "*Nondemented*".

Para ejemplificar este análisis, se incluye una explicación generada para una instancia que corresponde a un FN utilizando el modelo LR. Como se puede observar, la Figura 7 muestra la interpretación resultante de la técnica LIME. Las barras en color rojo indican las características que contribuyen a la predicción de la clase "*nondemented*", mientras que las barras en color verde muestran las características a favor de la clase "*demented*". La longitud de cada barra refleja el peso relativo de la característica en la predicción local. De acuerdo con estos resultados se infiere que el modelo otorga mayor relevancia a las características de ASF, CRD, EDUC, M/F_M, y nWBV cuyos valores favorecen la clasificación de "*nondemented*". En contraste, las características de MMSE, Age, SES y eTIV aportan una ponderación a favor de la clase "*demented*". La predicción final se obtiene mediante la suma ponderada de estas contribuciones, por lo cual se deduce que las características estructurales del cerebro y cognitivas favorecen a una predicción correcta. Mediante este análisis se identificaron dos hallazgos relevantes, el primero fue que la mayoría de los modelos entrenados con 150 instancias otorga mayor peso a las características cognitivas como ASF, EDU, CRF para clasificar instancias en la clase "*nondemented*"; mientras que los modelos entrenados con 373 instancias otorgan más peso a las características estructurales del cerebro como eTIV, MMSE y a la edad. El segundo fue que, al trabajar con cada una de las instancias identificadas como FN, se identificó que antes del preprocesamiento, tenían asignada la etiqueta de "*Converted*". Sin embargo, al seguir el preprocesamiento de datos propuesto en la literatura [8],[9], estas instancias fueron reclasificadas como "*Demented*". Lo anterior nos lleva a mencionar que los modelos realizan una clasificación correcta de acuerdo con los datos evaluados ya que con los datos iniciales estos pacientes no presentaban demencia.

En la Tabla 11 se muestran las probabilidades finales asignadas por LIME para cada instancia de los FN de los modelos entrenados con el conjunto de datos de 150 pacientes. De manera similar, en la Tabla 12 se muestran las probabilidades asignadas por LIME para cada instancia de los FN de los modelos entrenados con el conjunto de datos de 373 pacientes. Estas probabilidades corresponden a la estimación de LIME sobre la confianza de cada modelo en que una instancia pertenezca a cada clase ("*Demented*" o "*Nondemented*"), mostrando cómo el modelo pondera las características para generar su predicción en cada caso.

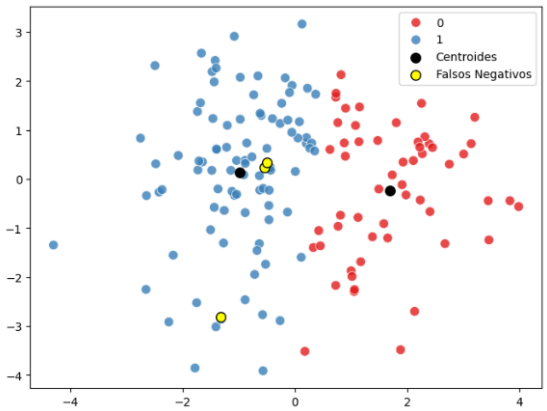

Figura 5. K-Means del conjunto de datos de 150 pacientes (0 es la etiqueta 'Nondemented' y 1 es la etiqueta 'Demented').

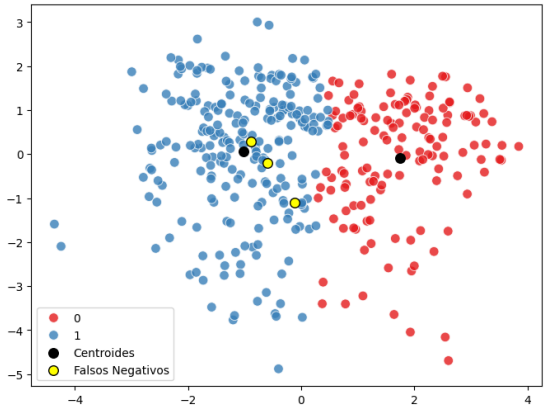

Figura 6. *K-Means* del conjunto de datos de 373 pacientes (0 es la etiqueta 'Nondemented' y 1 es la etiqueta 'Demented').

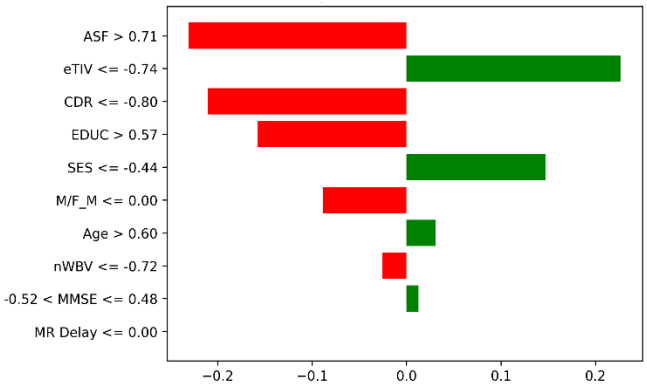

Figura 7. Ejemplo de FN.

Tabla 11. Probabilidades de LIME para el conjunto de datos de 150 pacientes.

Instancia	'Nondemented'			'Demented'		
	SVM	RF	LR	SVM	RF	LR
1	.84	.83	.80	.16	.17	.20
2	.84	.76	.87	.16	.24	.13
3	.84	.98	.99	.16	.02	.01

Tabla 12. Probabilidades de LIME para el conjunto de datos de 373 pacientes.

Instancia	'Nondemented'			'Demented'		
	SVM	RF	LR	SVM	RF	LR
1	.90	.96	.90	.10	.04	.10
2	.90	.95	.89	.10	.04	.11
3	.90	.91	.89	.10	.10	.11

4.4 Análisis manual de instancias

En el análisis manual de las instancias se identificó que el problema de clasificación observado, tanto en el conjunto de datos de 150 participantes como en el de 373, se debía a que estos pacientes estaban etiquetados como *converted*. Según la literatura [8], [19], los pacientes con esta etiqueta suelen reclasificarse como *demented* durante el preprocesamiento, procedimiento que también se siguió en este estudio. Como recapitulación, OASIS-2 recolectó datos longitudinales de los pacientes; aquellos etiquetados como *converted* corresponden a individuos que no presentaban Alzheimer en la primera toma de muestra, pero sí en mediciones posteriores. Por lo tanto, la instancia mal etiquetada corresponde a la primera visita de algunos pacientes, momento en el cual no tenían la enfermedad. En este sentido, el modelo está realizando una clasificación correcta, y el problema radica en el preprocesamiento de los datos. Adicionalmente, aunque existen más pacientes con la etiqueta *converted*, en los análisis se identificaron únicamente tres instancias en el conjunto de 150 participantes y tres instancias diferentes en el conjunto de 373 participantes, independientemente del modelo utilizado.

5 Discusión

Al usar el conjunto de datos OASIS-2 para la clasificación de pacientes con Alzheimer, se muestra que los modelos aumentan sus métricas al ser entrenados con el conjunto de datos completo (373 participantes) y no con la manera tradicional de hacerlo, con la primera visita (150 participantes). Esto sugiere que las características incluidas en este conjunto de datos son relevantes para la clasificación de Alzheimer. En general, los modelos entrenados con el conjunto de datos completo alcanzaron un *accuracy* de 96%, superando a los modelos entrenados con solo la primera visita, que en promedio obtuvieron un 90%.

Por otro lado, con los análisis de explicación (*K-Means* y LIME) se logró identificar cuál es el sesgo que tienen los modelos al clasificar. En este caso, las instancias clasificadas como *'converted'* son las que generan este error de clasificación y producen falsos negativos. Esta etiqueta representa participantes cuya transición clínica ocurrió a lo largo de las visitas. Esto indica que los errores de clasificación no se deben a la calidad de los datos o a inconsistencias del algoritmo, sino a un preprocesamiento que no enfatizó la explicación y equidad de los modelos.

En la literatura revisada, al tomar la decisión de cómo tratar estos datos, no se contempló la explicación y equidad de los modelos. Por lo tanto, la presencia de estos falsos negativos es un sesgo presente en el estado del arte. En este contexto, analizar la explicación y equidad de los modelos entrenados con OASIS-2 es el principal aporte del presente estudio. Aunque en el estado del arte se encuentran estudios en los que se busca la optimización de modelos de ML entrenados con OASIS-2, no se encontró literatura que implemente algoritmos especializados para la comprensión de las inconsistencias de clasificación. Esta contribución nos permite cuestionar el cómo se procesan los datos, previo al entrenamiento de los modelos. En este sentido, en el estado del arte se expusieron casos que evidencian los problemas sociales que implican no incluir análisis de explicación en los modelos. Por lo tanto, este trabajo contribuye a la subárea de IHC en Inteligencia Artificial (IHC/IA) al promover la investigación de técnicas de explicación. Se plantea que LIME y *K-means* son herramientas eficientes para construir sistemas transparentes y confiables en el dominio sensible de la salud. Estas técnicas aseguran que los avances tecnológicos se traduzcan en beneficios reales y equitativos para los pacientes.

Sin embargo, la principal limitación del estudio fue tomar como distintos participantes a diferentes tomas de muestra de un mismo participante. Esto implica varias suposiciones de nuestra parte, y esto implica otro tipo de sesgos en los resultados del modelo. Por ejemplo, la decisión es útil para explorar el

comportamiento del modelo con más instancias, pero ignora la dependencia estadística entre mismos participantes. Por otro lado, el análisis de explicación no se abordó a profundidad, solo se identificaron las probabilidades de las instancias clasificadas como falsos negativos. Adicionalmente, a pesar de haber utilizado *K-Means* para identificar falsos negativos, no se implementaron métricas cuantitativas de equidad. Estas deficiencias generan sesgos no explorados en los modelos.

Partiendo de lo anterior, se propone generar modelos que en su entrenamiento consideren estructuras de datos que permitan abordar datos longitudinales. Esto eliminaría el sesgo del presente estudio, ya que cada participante sería incluido en su totalidad. Adicionalmente, no solo abordaría las inconsistencias de procesamiento de datos aquí planteadas, sino que también favorecería la equidad del modelo. Por otro lado, se sugiere implementar explicaciones de los aditivos Shapley (por sus siglas en inglés, *SHapley Additive exPlanations*, SHAP) para aumentar la profundidad del análisis de explicación. SHAP explica las predicciones mostrando cuánto aporta cada instancia a la decisión del modelo, esto mediante principios de teoría de juegos. Por lo tanto, es una técnica que puede generar análisis más completos sobre los modelos entrenados con OASIS-2.

En síntesis, los resultados permiten reconocer inconsistencias en el estado del arte sobre el procesamiento de los datos en *pipelines* que integran a OASIS-2. Nuestra propuesta soluciona estas inconsistencias y propone soluciones viables para las incógnitas que surgieron durante el análisis.

6 Conclusiones

El presente trabajo plantea las limitaciones de explicación y equidad de los modelos de ML en el diagnóstico de Alzheimer, teniendo profundas implicaciones para la comunidad de IHC. La falta de transparencia en los modelos (el problema de la "caja negra") y existencia de sesgos no analizados comprometen directamente el diseño de sistemas éticos y confiables para el apoyo a la decisión clínica. Para que las herramientas de diagnóstico basadas en ML se adopten por médicos y especialistas, es crucial que los sistemas no solo ofrezcan alta precisión, sino que también proporcionen explicaciones claras y accionables que permitan al usuario (médico o investigador) comprender la justificación de la predicción y mitigar los riesgos de errores o inequidad.

Los resultados indican que, aunque los modelos de aprendizaje automático pueden alcanzar un desempeño alto incluso con conjuntos de datos limitados, su integración en prototipos orientados al diagnóstico de Alzheimer requiere consideraciones que van más allá de la precisión del modelo. Incorporar estos modelos en prototipos funcionales y escenarios reales, es importante asegurarse de que se evalúen en términos de equidad y explicación. Si bien, este estudio únicamente se enfoca en la explicación y equidad, es recomendable que, en caso de no existir evaluaciones previas, se deban realizar antes de la implementación o comunicar de manera transparente al usuario las limitaciones del modelo. Además, resulta relevante comprender cómo se perciben los resultados emitidos por el modelo, especialmente en poblaciones vulnerables, dado que la interpretación de una predicción puede influir en la confianza, la adherencia al tratamiento y la aceptación del sistema. Estos elementos subrayan la importancia de seguir investigando en cómo integrar principios de equidad, transparencia y centrado en el usuario al desarrollar prototipos de HCI en salud, garantizando que los avances tecnológicos se traduzcan en beneficios reales y equitativos para los pacientes.

7 Referencias

[1] Swanberg, M. M., Tractenberg, R. E., Mohs, R., Thal, L. J. and Cummings, J. L. “Executive Dysfunction in Alzheimer Disease”, *Arch Neurol*, vol. 61, núm. 4, p. 556, abr. 2004, doi: 10.1001/archneur.61.4.556.

[2] Frias, C. E., Cabrera, E. and Zabalegui, A. “Informal Caregivers’ Roles in Dementia: The Impact on Their Quality of Life”, *Life*, vol. 10, núm. 11, p. 251, oct. 2020, doi: 10.3390/life10110251.

[3] Goren, A., Montgomery, W., Kahle-Wrobleski, K., Nakamura, T. and Ueda, K. “Impact of caring for persons with Alzheimer’s disease or dementia on caregivers’ health outcomes: findings from a community based survey in Japan”, *BMC Geriatr*, vol. 16, núm. 1, p. 122, dic. 2016, doi: 10.1186/s12877-016-0298-y.

[4] Fathi, S., Ahmadi, M. and Dehnad, A. “Early diagnosis of Alzheimer’s disease based on deep learning: A systematic review”, *Computers in Biology and Medicine*, vol. 146, p. 105634, jul. 2022, doi: 10.1016/j.compbiomed.2022.105634.

[5] Tan, W. Y., Hargreaves, C., Chen, C. and Hilal, S. “A Machine Learning Approach for Early Diagnosis of Cognitive Impairment Using Population-Based Data”, *JAD*, vol. 91, núm. 1, pp. 449–461, ene. 2023, doi: 10.3233/JAD-220776.

[6] Diogo, V. S., Ferreira, H. A., Prata, D. and for the Alzheimer’s Disease Neuroimaging Initiative, “Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach”, *Alz Res Therapy*, vol. 14, núm. 1, p. 107, dic. 2022, doi: 10.1186/s13195-022-01047-y.

[7] Ferrara, E. “The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness”, *Machine Learning with Applications*, vol. 15, p. 100525, mar. 2024, doi: 10.1016/j.mlwa.2024.100525.

[8] Khan A. and Zubair, S. “Longitudinal Magnetic Resonance Imaging as a Potential Correlate in the Diagnosis of Alzheimer Disease: Exploratory Data Analysis”, *JMIR Biomed Eng*, vol. 5, núm. 1, p. e14389, abr. 2020, doi: 10.2196/14389.

[9] Cockrell, J. R. and Folstein, M. F. “Mini-Mental State Examination (MMSE)”, *Psychopharmacol Bull*, vol. 24, núm. 4, pp. 689–692, 1988.

[10] Morris, J. C. *et al.*, “Clinical Dementia Rating training and reliability in multicenter studies: The Alzheimer’s Disease Cooperative Study experience”, *Neurology*, vol. 48, núm. 6, pp. 1508–1510, jun. 1997, doi: 10.1212/WNL.48.6.1508.

[11] Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. “Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults”, *Journal of Cognitive Neuroscience*, vol. 22, núm. 12, pp. 2677–2684, dic. 2010, doi: 10.1162/jocn.2009.21407.

[12] Chorąś, M., Pawlicki, M., Puchalski, D. and Kozik, R. “Machine Learning – The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness?”, en *Computational Science – ICCS 2020*, vol. 12140, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, y J. Teixeira, Eds.,

en Lecture Notes in Computer Science, vol. 12140., Cham: Springer International Publishing, 2020, pp. 615–628. doi: 10.1007/978-3-030-50423-6_46.

[13] Cutillo, C. M. *et al.*, "Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency", *npj Digit. Med.*, vol. 3, núm. 1, p. 47, mar. 2020, doi: 10.1038/s41746-020-0254-2.

[14] Dosilovic, F. K., Brcic, M. and Hlupic, N. "Explainable artificial intelligence: A survey", en *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija: IEEE, may 2018, pp. 0210–0215. doi: 10.23919/MIPRO.2018.8400040.

[15] Barocas, S. and Selbst, A.D. "Big Data's Disparate Impact", 2016, doi: 10.15779/Z38BG31.

[16] Battineni, G., Chintalapudi, N. and Amenta, F. "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)", *Informatics in Medicine Unlocked*, vol. 16, p. 100200, 2019, doi: 10.1016/j.imu.2019.100200.

[17] Sørensen, L. and Nielsen, M. "Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination", *Journal of Neuroscience Methods*, vol. 302, pp. 66–74, may 2018, doi: 10.1016/j.jneumeth.2018.01.003.

[18] Bari Antor*, M. et al.*, "A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease", *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, jul. 2021, doi: 10.1155/2021/9917919.

[19] Khan A. and Zubair, S. "An Improved Multi-Modal based Machine Learning Approach for the Prognosis of Alzheimer's disease", *Journal of King Saud University - Computer and Information Sciences*, vol. 34, núm. 6, pp. 2688–2706, jun. 2022, doi: 10.1016/j.jksuci.2020.04.004.

[20] Morales-Forero, A., Rueda Jaime, L., Gil-Quiñones, S.R., Barrera Montañez, M.Y., Bassetto, S. and Coatanea, E. "An insight into racial bias in dermoscopy repositories: A HAM10000 data set analysis", *JEADV Clinical Practice*, vol. 3, núm. 3, pp. 836–843, jul. 2024, doi: 10.1002/jvc2.477.

[21] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, vol. 366, núm. 6464, pp. 447–453, oct. 2019, doi: 10.1126/science.aax2342.

[22] Vellido, A. "The importance of interpretability and visualization in machine learning for applications in medicine and health care", *Neural Comput & Applic*, vol. 32, núm. 24, pp. 18069–18083, dic. 2020, doi: 10.1007/s00521-019-04051-w.

[23] Hassan, S. U., Abdulkadir, S. J., Zahid, M. S. M. and Al-Selwi, S. M. "Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review", *Computers in Biology and Medicine*, vol. 185, p. 109569, feb. 2025, doi: 10.1016/j.compbiomed.2024.109569.

[24] Shaikh, A. S., Samant, R. M., Patil, K. S., Patil, N. R. and Mirkale, A. R. "Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain", *IJCA*, vol. 185, núm. 45, pp. 18–23, nov. 2023, doi: 10.5120/ijca2023923263.

[25] Magesh, P. R., Myloth, R. D. and Tom, R. J. "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery", *Computers in Biology and Medicine*, vol. 126, p. 104041, nov. 2020, doi: 10.1016/j.compbiomed.2020.104041.

[26] Falvo F. R. and Cannataro, M. "Explainability techniques for Artificial Intelligence models in medical diagnostic", en *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal: IEEE, dic. 2024, pp. 6907–6913. doi: 10.1109/BIBM62325.2024.10821826.

[27] Guzmán Ponce, A., López-Bautista, J. and Fernandez-Beltran, R. "Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME", *IRCI*, vol. 2, núm. 2, p. 15, jul. 2024, doi: 10.36677/ideaseningenieria.v2i2.23952.

[28] Varghese, A., Sherimon, V., Raja, X. C., Ephrem, B. G. and Gouda, P. "Neural Imaging for Alzheimer's Prediction using AI: Exploring CNNs and LIME Explanations", *Alzheimer's & Dementia*, vol. 20, núm. S4, p. e088802, dic. 2024, doi: 10.1002/alz.088802.

[29] Kamal, Md. S., Northcote, A., Chowdhury, L., Dey, N., Crespo, R.G. and Herrera-Viedma, E. "Alzheimer's Patient Analysis Using Image and Gene Expression Data and Explainable-AI to Present Associated Genes", *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–7, 2021, doi: 10.1109/TIM.2021.3107056.

[30] Soladoye, A.A., Aderinto, N., Osho, D. and Olawade, D.B. Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data", *International Journal of Medical Informatics*, vol. 204, p. 106093, ago. 2025, doi: 10.1016/j.ijmedinf.2025.106093.

[31] Loveleen, G., Mohan, B., Shikhar, B. S., Nz, J., Shorfuzzaman, M. and Masud, M. "Explanation-Driven HCI Model to Examine the Mini-Mental State for Alzheimer's Disease", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, núm. 2, pp. 1–16, feb. 2024, doi: 10.1145/3527174.

[32] Eckhardt, C.M. *et al.*, "Unsupervised machine learning methods and emerging applications in healthcare", *Knee surg. sports traumatol. arthrosc.*, vol. 31, núm. 2, pp. 376–381, feb. 2023, doi: 10.1007/s00167-022-07233-7.

[33] Ripan, R. C., Sarker, I. H., Hasan Furhad, Md., Musfique Anwar, M. and Hoque, M. M. "An Effective Heart Disease Prediction Model Based on Machine Learning Techniques", en *Hybrid Intelligent Systems*, A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios, y T.-P. Hong, Eds., en Advances in Intelligent Systems and Computing, vol. 1375. Cham: Springer International Publishing, 2021, pp. 280–288. doi: https://doi.org/10.1007/978-3-030-73050-5_28.

[34] Nedyalkova, M., Madurga, S. and Simeonov, V. "Combinatorial K-Means Clustering as a Machine Learning Tool Applied to Diabetes Mellitus Type 2", *IJERPH*, vol. 18, núm. 4, p. 1919, feb. 2021, doi: 10.3390/ijerph18041919.

DOI: 10.47756/aihc.y10i1.191