



Modelos de predicción para la detección de Alzheimer

Laura Díaz-Muñoz Manzanares

Proyecto final
Bootcamp en Data Science y Machine Learning

Índice

1. INTRODUCCIÓN.....	3
1.1. OBJETIVOS E IMPORTANCIA DEL PROYECTO.....	3
2. OBTENCIÓN DE DATOS.....	3
3. MODELOS DE CLASIFICACIÓN.....	4
3.1. MACHINE LEARNING.....	4
3.1.1. Variables que conforman el dataset:.....	5
3.1.2. Limpieza y Procesamiento de Datos:.....	6
3.1.3. Aplicación de modelos de machine learning.....	7
3.1.4. Elección del Modelo.....	9
3.1.5. Conclusión.....	10
3.2. DEEP LEARNING	10
3.2.1. Obtención datos	10
3.2.2. Preprocesado y etiquetado	11
3.2.3. Aplicación de redes convolucionales	12
3.2.4. VGG16 CNN.....	13
4. DISCUSIÓN DE RESULTADOS FINALES.....	14
5. CONCLUSIÓN FINAL Y FUTUROS PASOS.....	15
5.1. FUTUROS PASOS	15
6. REFERENCIAS.....	16

1. Introducción

La enfermedad de Alzheimer es un tipo de trastorno cerebral progresivo y degenerativo que produce la muerte de células cerebrales y, en consecuencia, la reducción del tejido cerebral y del tamaño total del cerebro. Además, se manifiesta también como un deterioro cognitivo y un trastorno conductual. Aunque esta enfermedad no presenta un tratamiento de cura, el diagnóstico temprano ha sido señalado como un factor clave para retrasar potencialmente el avance de la enfermedad y de sus síntomas [1]. El Deterioro Cognitivo Leve (DCL) es una etapa intermedia entre la cognición normal y la enfermedad de Alzheimer, donde el 32% de los pacientes con DCL desarrollarán finalmente esta enfermedad. Esto convierte al DCL en un objetivo ideal para la predicción temprana. En este contexto, la aplicación de modelos predictivos se ha vuelto esencial para identificar tempranamente a individuos en riesgo y permitir intervenciones preventivas.

1.1. Objetivos e importancia del proyecto

Este proyecto aborda la necesidad de desarrollar herramientas médicas efectivas que agilicen la identificación temprana de Alzheimer con el fin de, no solo mejorar las posibilidades de tratamiento, sino también brindar a los profesionales de la salud la capacidad de planificar y proporcionar el cuidado adecuado reduciendo su carga de trabajo y priorizando los sujetos con la enfermedad más avanzada.

A lo largo de este proyecto se han utilizado neuroimágenes, así como, datos clínicos y cognitivos relacionados con la enfermedad de Alzheimer, enfrentándonos a desafíos como el desequilibrio en la distribución de clases y la presencia de datos faltantes. Mediante la aplicación de técnicas de preprocesamiento y modelos avanzados de aprendizaje automático, buscamos ofrecer una contribución significativa al campo de la salud y la investigación médica.

Por último, este proyecto no solo se centra en la creación de un modelo eficaz, sino también en comprender la importancia de las métricas específicas en el contexto médico. La capacidad de identificar con precisión casos positivos, en términos de *recall*, se convierte en un factor clave en la evaluación de nuestro modelo y su relevancia en la práctica clínica.

2. Obtención de datos

Los datos utilizados para la creación de los diferentes modelos fueron obtenidos del proyecto "The Open Access Series of Imaging Studies (OASIS)", un proyecto orientado a poner a disposición de la comunidad científica conjuntos de neuroimágenes del cerebro y de datos clínicos reales. En particular, utilizamos el conjunto de datos OASIS-3.

OASIS-3 es una recopilación de datos clínicos, cognitivos y neuroimágenes de 1378 participantes recopilados a lo largo de 30 años. Los participantes incluyen 755 adultos cognitivamente normales y 622 individuos en diversas etapas de deterioro cognitivo, con edades comprendidas entre 42 y 95 años. Esta diversidad en la población de estudio nos brinda la oportunidad de explorar y comprender la variabilidad en las características cerebrales y cognitivas asociadas con diferentes estados de salud mental.

El conjunto de datos contiene un total de 2842 sesiones de resonancias magnéticas (MRI), tomografías por emisión de positrones (PET), así como datos clínicos y cognitivos. En particular para este proyecto, se han utilizado las secuencias T1w de las resonancias magnéticas, con un total de 1952 imágenes de sujetos con Alzheimer y 890 imágenes relacionadas con la cognición normal.

La combinación de neuroimágenes y datos clínicos y cognitivos hace que el conjunto de datos de OASIS-3 sea una herramienta excepcionalmente valiosa para abordar nuestro objetivo de desarrollar un modelo preciso de predicción de Alzheimer.

3. Modelos de clasificación

3.1. Machine Learning

Este proyecto se centra en el desarrollo de un modelo de machine learning para predecir la presencia de demencia en pacientes basándose en sus historiales médicos, evaluaciones cognitivas, juicio clínico de síntomas y datos demográficos. Para ello, se utilizaron cuatro conjuntos de datos: "health" (historial médico), "globaldata" (evaluaciones cognitivas), "demo" (datos demográficos) y "uds" (juicio clínico de síntomas).

Se emplearon los algoritmos RandomForest y XGBoost, implementando técnicas como el muestreo estratificado para abordar el desbalance de clases. La evaluación del modelo se llevó

a cabo utilizando métricas de precisión, sensibilidad (*recall*), *F1-score* y curva ROC, proporcionando una visión integral del rendimiento del modelo en un conjunto de datos desequilibrado. El objetivo principal consistía en la optimización de la métrica de *recall* o tasa de verdaderos positivos, cuya fórmula es la siguiente:

$$Recall = \frac{\text{Verdaderos positivos}}{\text{Falsos negativos} + \text{Verdaderos positivos}}$$

Los Verdaderos Positivos hacen referencia a casos positivos (sujetos con Alzheimer) que el modelo ha clasificado correctamente. Por otro lado, los Falsos Negativos son los casos positivos que el modelo ha clasificado incorrectamente como negativos (sujetos que tienen la enfermedad, pero son clasificados como si no la tuvieran). El *recall* es especialmente importante en un contexto médico donde identificar la enfermedad es crucial con el fin de proporcionar un tratamiento temprano. Por lo tanto, en este tipo de contexto se prioriza la minimización de los falsos negativos para evitar que los pacientes no reciban la atención necesaria. Sin embargo, la optimización del *recall* normalmente conlleva una menor precisión, lo que aumentaría el número de falsos positivos (personas que son clasificadas con la enfermedad, pero no la tienen).

Así pues, se exploraron diferentes estrategias para mejorar el equilibrio entre *recall* y precisión y mejorar la capacidad predictiva e interpretativa del modelo.

3.1.1. Variables que conforman el dataset:

Los cuatro conjuntos de datos comparten variables clave como "OASISID", "OASIS_session_label", "days_to_visit" y "age_at_visit". Cada conjunto de datos aporta información específica:

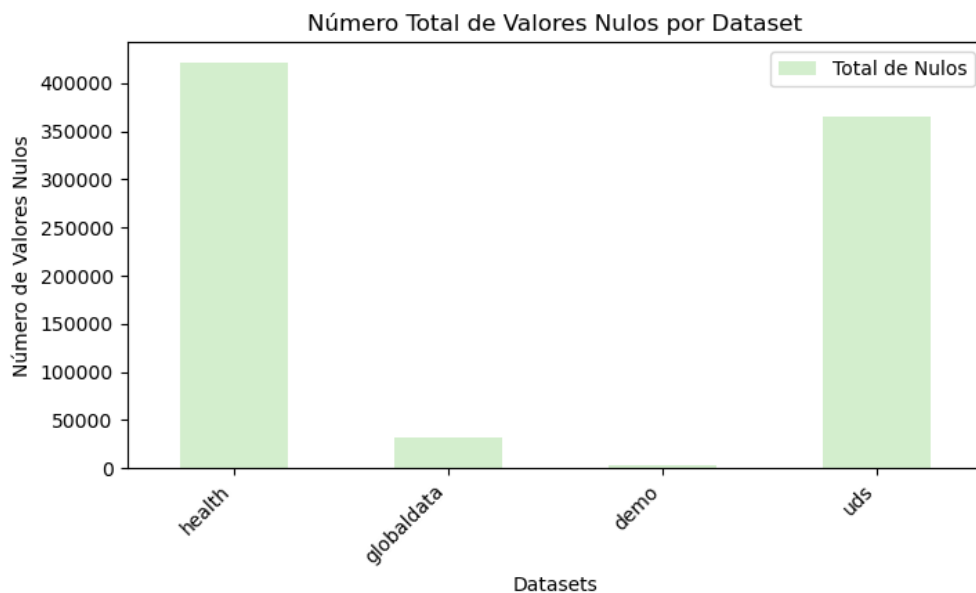
- **Dataset "health" (Historial Médico):** Contiene 72 variables relacionadas con diversos problemas médicos, como enfermedades cardiovasculares, cerebrovasculares, hábitos de tabaquismo y trastornos del sueño.
- **Dataset "globaldata" (Evaluaciones Cognitivas):** Incluye resultados de evaluaciones cognitivas, como el "Clinical Dementia Rating (CDR)" y el "Mini-Mental State Examination (MMSE)", con la creación de la variable objetivo "Outcome" para clasificar casos de demencia.

- **Dataset "demo" (Datos Demográficos):** Se seleccionaron variables como género, años de educación, estatus socioeconómico, raza, y datos genéticos relevantes.
- **Dataset "uds" (Juicio Clínico de Síntomas):** Se eligieron variables con menor cantidad de datos nulos, como declive reportado por el paciente, informante y clínico, así como la edad de inicio del declive cognitivo.

3.1.2. Limpieza y Procesamiento de Datos:

Los conjuntos de datos utilizados presentaron desafíos significativos en términos de calidad y consistencia, ya que se identificó un porcentaje sustancial de valores nulos que requerían atención especial para garantizar la validez y la coherencia de los datos. Además, contábamos con un dataset final que estaba desbalanceado.

Con relación al manejo de valores nulos (Figura 1), fueron necesarias el uso de estrategias específicas para abordar el problema. Por un lado, se llevó a cabo una exhaustiva revisión de las columnas en los datasets para identificar aquellas que eran irrelevantes para el objetivo del proyecto, las cuales fueron eliminadas.



Figura

Figura 1: Gráfico de barras que muestra el número total de valores nulos en cada uno de los dataset utilizados en el proyecto.

Por otro lado, se agruparon variables que mostraban relación entre ellas para consolidar la información pertinente y eliminar valores nulos, facilitando así una representación más efectiva en el modelo.

Con relación a la distribución de nuestra variable objetivo, contábamos con un dataset desbalanceado (Figura 2). El desbalance de clases es una característica común en los conjuntos de datos médicos y fue abordado mediante estrategias de submuestreo, sobremuestreo y muestreo estratificado. Además, también se ajustó el umbral de decisión del modelo a 0.3 (el punto en el que decide clasificar una instancia como positiva o negativa) con el fin de aumentar la probabilidad de que el modelo clasifique una instancia como positiva (variable minoritaria).

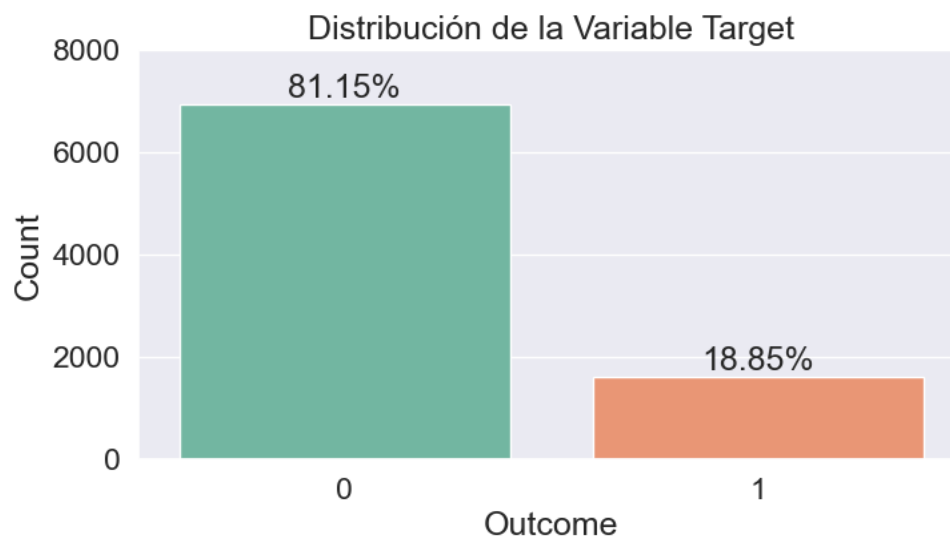


Figura 2: Se muestra la distribución de sujetos con Alzheimer (etiqueta 0) y sin la enfermedad (etiqueta 1).

Se aplicó submuestreo para la clase mayoritaria y sobremuestreo para la clase minoritaria, buscando equilibrar la representación de casos positivos y negativos. Esta medida fue esencial para mejorar la capacidad predictiva del modelo, especialmente en el contexto médico donde la identificación de casos positivos es de vital importancia.

3.1.3. Aplicación de modelos de machine learning

Durante el desarrollo de este proyecto, se llevó a cabo una evaluación de diversos modelos de machine learning con el objetivo de construir un sistema de clasificación efectivo para predecir la presencia de Alzheimer en pacientes. Se exploraron distintas combinaciones de parámetros y estructuras de modelos, considerando también el número total de variables utilizadas. A continuación, se presenta un resumen de los principales resultados obtenidos al probar diferentes enfoques.

-Random Forest (30 variables):

En este modelo se evidenció una dificultad significativa en el *recall* (sensibilidad), indicando que el modelo no lograba identificar adecuadamente casos positivos reales, lo cual es crucial en un contexto médico.

-Random Forest con Submuestreo y Sobremuestreo:

- **Submuestreo:** Este enfoque mostró un buen rendimiento en *recall*, logrando capturar más casos positivos. Sin embargo, se asoció con una menor precisión, lo que implica un posible aumento en los falsos positivos (Tabla 1).

	precision	recall	f1-score	support
Dementia	0.955	0.833	0.8898	1377
Non dementia	0.5481	0.8378	0.6627	333

Tabla 1: Métricas para el modelo Random Forest usando la técnica de submuestreo

- **Sobremuestreo:** Se observó una mayor precisión y exactitud, pero a expensas de un menor *recall*. Aunque el modelo fue más preciso, la capacidad de identificar casos positivos disminuyó (Tabla 2).

	precision	recall	f1-score	support
Dementia	0.9248	0.9557	0.94	1377
Non dementia	0.7875	0.6787	0.729	333

Tabla 2: Métricas para el modelo Random Forest usando la técnica de sobremuestreo

-XGBoost con Muestreo Estratificado (Umbral 0.3, 'scale_pos_weight': 2, 30 variables):

En problemas de clasificación desbalanceada, donde hay una proporción significativamente mayor de una clase en comparación con la otra, el muestreo estratificado ayuda a mantener un equilibrio en las clases objetivo. Esto es beneficioso para asegurar que el modelo aprenda de

manera adecuada de ambas clases. Al utilizar XGBoost con muestreo estratificado, se logró un mayor *recall*, indicando una mejora en la identificación de casos positivos, particularmente de demencia. Sin embargo, se observó un ligero descenso en la precisión, lo cual es un compromiso común al priorizar el aumento del *recall*. También se ajustó el umbral de decisión del modelo a 0.3 (el punto en el que decide clasificar una instancia como positiva o negativa) aumentando la probabilidad de que el modelo clasifique una instancia como positiva, lo que incrementó la sensibilidad (*recall*) de nuestro modelo (Tabla 3).

	precision	recall	f1-score	support
Dementia	0.9417	0.8911	0.9157	1377
Non dementia	0.6314	0.7718	0.6946	333

Tabla 3: Métricas para el modelo de XGboost incluyendo 30 variables.

-XGBoost (34 variables):

Con la inclusión de 4 variables más (relacionadas con los síntomas de los sujetos) a nuestro conjunto de datos y manteniendo los parámetros anteriores, el modelo de XGBoost mostró un mayor *recall* en comparación con los modelos anteriores, indicando una mayor capacidad para identificar casos positivos. Se logró también un buen equilibrio entre precisión y *recall*, como se reflejaba en el F1 Score, una métrica que combina ambas dimensiones de rendimiento (Tabla 4).

	precision	recall	f1-score	support
Dementia	0.9567	0.9136	0.9346	1377
Non dementia	0.6987	0.8288	0.7582	333

Tabla 4: Métricas para el modelo XGBoost con 34 variables.

3.1.4. Elección del Modelo

La decisión final respecto a la selección del modelo se fundamentó en la búsqueda de un equilibrio óptimo entre la sensibilidad (*recall*) y precisión, como reflejado en el *F1 Score*. Se otorgó prioridad a un *recall* elevado con el objetivo de identificar de manera eficaz los casos

positivos, minimizando así la probabilidad de falsos negativos en el diagnóstico de demencia. También, se consideró la importancia de evitar un exceso de falsos positivos, los cuales podrían dar lugar a pruebas adicionales, generando costos económicos y temporales adicionales. De esta manera, la estrategia adoptada se alinea con la necesidad de maximizar la detección de casos positivos, manteniendo a su vez un control sobre la precisión del modelo para evitar clasificaciones erróneas.

3.1.5. Conclusión

El tercer modelo de XGBoost, que incorpora 34 variables, fue seleccionado como la opción final debido a su mejora significativa en *recall* en comparación con modelos previos, manteniendo al mismo tiempo niveles razonables de *precision* y *F1 Score*. Esta elección se ajusta a la estrategia adoptada, que tiene como prioridad la identificación efectiva de casos de Alzheimer, reduciendo así el riesgo de pasar por alto pacientes con esta condición. El equilibrio logrado entre precisión y *recall* respalda la utilidad clínica y la eficiencia del modelo en el diagnóstico de demencia, reforzando su capacidad para proporcionar resultados precisos y relevantes en un contexto médico.

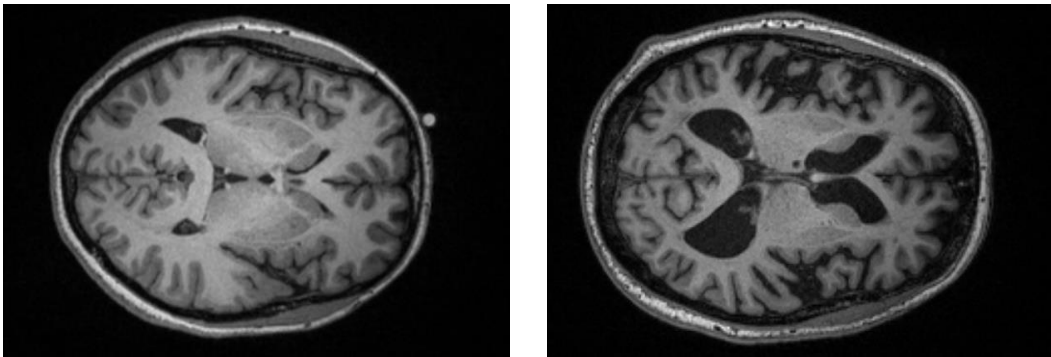
3.2. Deep Learning

Una vez realizada la clasificación basada en los datos clínicos, este estudio pretende ofrecer otra metodología de estudio y clasificación basado en redes convolucionales (CNN) utilizando datos de imagen médica de los mismos pacientes. Es decir, ofrece una segunda visión de clasificación basada esta vez en los efectos físicos y estructurales resultado del impacto del deterioro cognitivo que caracteriza al Alzheimer.

3.2.1. Obtención datos

El acceso al conjunto de archivos (formato .nii.gz) correspondientes a las imágenes médicas extraídas de los 1379 pacientes participantes en este estudio, se obtuvo a través de la API Xnat debido al alto volumen de información disponible. Esto es en parte debido a que en este dataset no solamente se incluyen diferentes técnicas de imagen médica (PET, CT, MRI), y que, a su vez, en casos como el MRI contienen todavía más información ya que contienen diferentes tipologías del mismo (T1, T2, FLAIR...); Sino que además debemos tener en cuenta la naturaleza de estas imágenes, se tratan de imágenes que representan la totalidad del volumen del órgano, siendo formadas por varios slices de “imágenes” realizadas a lo largo de este.

En el caso de este estudio, se extrajo la información correspondiente a los MRI (son una de las formas más populares de detección de Alzheimer) y, específicamente, el T1w, debido a la conveniencia de encontrarse presente en todos los sujetos. De estos a su vez se extrajo (por mantener una generalización coherente) la slice central, lo que permitió obtener imágenes lo más similares posibles para facilitar la clasificación.



Muchos pacientes de este estudio, al tratarse de una investigación sostenida a lo largo de 30 años, se sometieron a más de un MRI, por lo que después de todo el proceso de preparación de datos, se obtienen un total de 2842 imágenes útiles para el modelo.

3.2.2. Preprocesado y etiquetado

Las imágenes resultantes fueron sometidas a un preprocesamiento básico que consistía en:

- Normalización de intensidad
- Escalado de valores
- Redimensionalización a 256x176 (debido a la ligera inconsistencia en el tamaño de algunas de las imágenes)
- Reducción de ruido mediante *Gaussian Blur*

Una vez aplicado el preprocesado, se procedió al etiquetado de las imágenes resultantes. Este proceso tuvo la complejidad añadida de no contar con etiquetados directos en las imágenes o en las sesiones de MRI, como ya se advierte en la página de OASIS. Así, se decidió basar la categorización en los datos clínicos mediante la correlación entre la edad del paciente a la hora

de la sesión de MRI y la edad en el momento de cada consulta clínica, ya que las consultas sí se encuentran ligadas a un diagnóstico.

Siguiendo este método, el resultado obtenido fue de 890 imágenes con demencia y 1952 con cognitividad normal.

Como paso final, se efectuó un proceso de aumento de datos (con selección aleatoria y rotaciones de $\pm 20^\circ$), mediante el cual se acabó con un total de 2000 imágenes de cada categoría.

3.2.3. Aplicación de redes convolucionales

Se exploraron dos vías a la hora de aplicación de redes convolucionales: la creación de una red propia y el uso de una red pre entrenada, específicamente la VGG16.

Para la red propia, que también sirvió como método de investigación de las especificidades de las redes para el caso, se siguió un proceso de acierto y error:

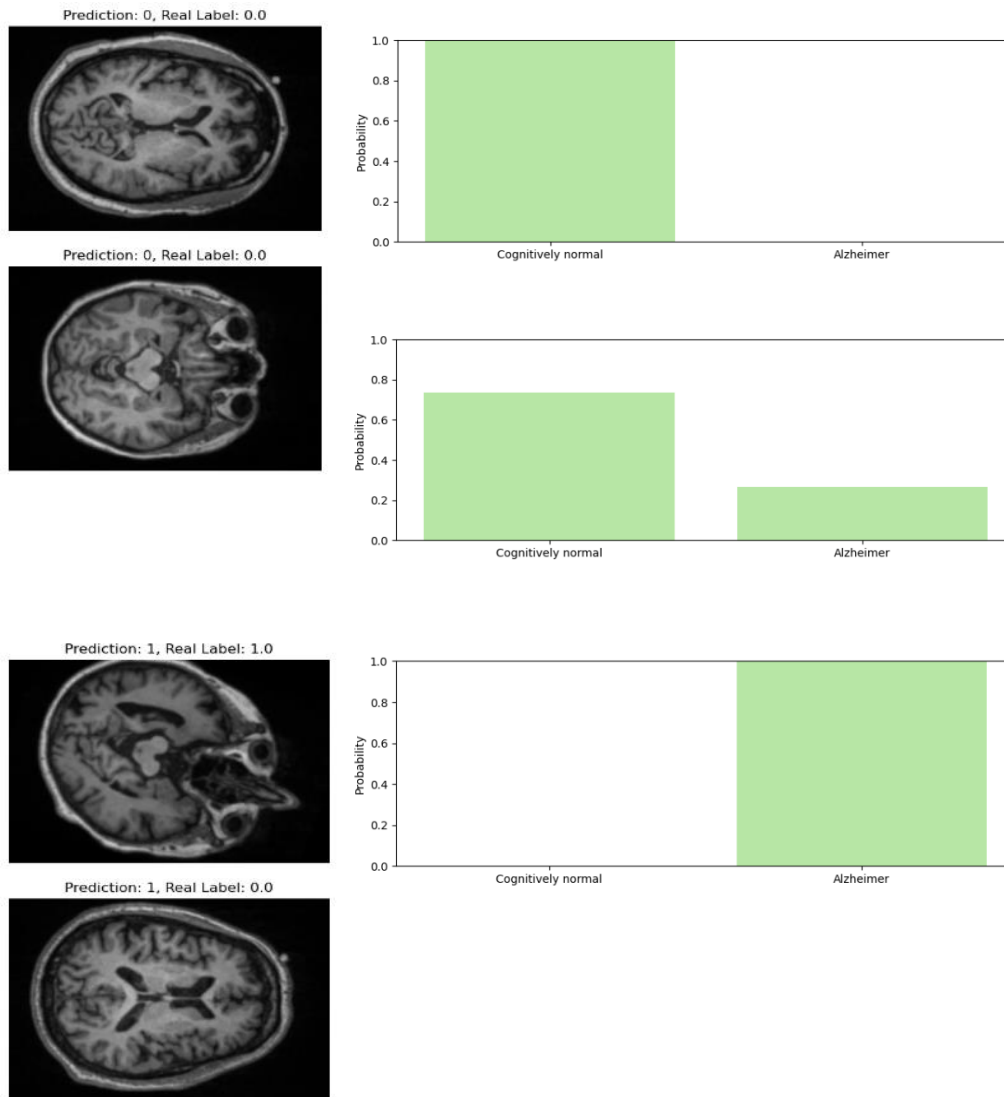
<i>Epochs</i>	<i>Batch size</i>	<i>Otras modificaciones</i>	<i>Accuracy en test set</i>
10	100	-	0.692
20	32	-	0.691
11	64	-	0.726
11	64	Ajustes preprocesamiento de imágenes	0.690
11	64	Ajustes de filtro y de <i>max pooling</i>	0.720
11	64	Data augmentation	0.728
11	64	Data augmentation + Ajuste de filtro	0.736

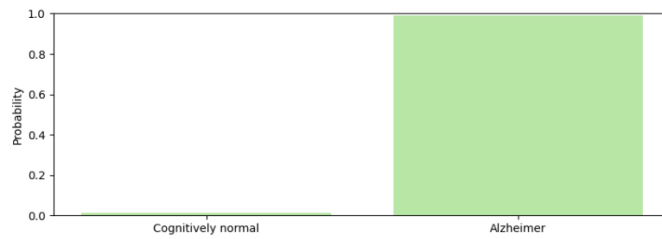
Mientras que con la red VGG16 solamente se tuvo que tener en cuenta la falta de canales. Las imágenes de MRI solamente tenían un canal al tratarse de imágenes en escala de grises; Sin embargo, esta red requería la existencia de 3 canales, por lo que se procedió a su conversión.

<i>Epochs</i>	<i>Batch size</i>	Otras modificaciones	<i>Accuracy en test set</i>
12	64	-	0.825
11	64	-	0.855

3.2.4. VGG16 CNN

El proceso de establecer una red propia sirvió también como forma de establecer un preprocesado de imágenes todo lo óptimo posible, aunque, como se esperaba, la red pre entrenada demostró un mejor rendimiento, por lo que ese es el método que se utilizó para llevar a cabo el proceso de detección y clasificación. Tal como se muestra en la tabla anterior, después de entrenar la red, la precisión en el test set fue de aproximadamente el 85%.





El resultado de clasificación se estableció por probabilidades y con un *threshold* de decisión de 0.5, aunque este podría ser modificado siguiendo la línea aplicada en el proceso de detección mediante datos clínicos (donde el *threshold* era menor para adaptarlo al contexto médico y aumentar el *recall*). En este caso, se mantuvo el *threshold* de 0.5 debido a que las métricas de la matriz de confusión mostraron resultados satisfactorios.

Confusion Matrix:			
[[219 58]			
[39 284]]			
Classification Report:			
	precision	recall	
0.0	0.85	0.79	
1.0	0.83	0.88	

4. Discusión de Resultados Finales

La comparación entre el modelo de machine learning (Random Forest y XGBoost) y el modelo de deep learning basado en redes convolucionales (VGG16) revela diferentes consideraciones clave. En el caso de los modelos de machine learning, específicamente el XGBoost con 34 variables, destacó por ofrecer un equilibrio significativo entre precisión y recall. Este equilibrio es esencial en un contexto médico, ya que prioriza tanto la identificación eficiente de casos positivos como la minimización de los falsos positivos y falsos negativos.

Por otro lado, el modelo de deep learning, utilizando la red convolucional VGG16, demostró una precisión del 85% en el conjunto de prueba, superando al modelo de machine learning en términos de rendimiento puro. Sin embargo, la interpretabilidad de las decisiones de la red neuronal puede ser un desafío, ya que las CNNs tienden a ser modelos complejos que destacan

por su arquitectura profunda. Aunque se logró un rendimiento notable, la comprensión detallada de cómo la red llega a sus conclusiones puede requerir más análisis.

5. Conclusión final y futuros pasos

Ambos enfoques, machine learning y deep learning, han demostrado ser prometedores para la detección temprana de la enfermedad de Alzheimer. El modelo de machine learning se destaca por su interpretabilidad y equilibrio en la precisión y *recall*, mientras que la red convolucional VGG16 logra un rendimiento óptimo a través de las redes preprocesadas.

La combinación de estos enfoques podría ser clave para mejorar aún más la precisión y la eficacia del sistema de detección. Integrar variables obtenidas de las imágenes, como intensidad y contraste, puede enriquecer la información disponible para el modelo. Además, estandarizar las pruebas y clasificar a los sujetos según el punto en el que se encuentran en el Clinical Dementia Rating (CDR) puede proporcionar más detalles sobre la progresión de la enfermedad y mejorar la personalización del tratamiento.

5.1. Futuros Pasos

- **Incorporar variables de imágenes:** Agregar variables derivadas de las imágenes, como intensidad y contraste para capturar información más detallada y relevante.
- **Pruebas en bases de datos estándar:** Probar los modelos en bases de datos de imágenes (MRI T1w) más estandarizadas para evaluar su rendimiento en entornos más diversos.
- **Clasificación según Clinical Dementia Rating (CDR):** Explorar la clasificación de sujetos según el punto en el que se encuentran en el Clinical Dementia Rating para una comprensión más completa de la progresión de la enfermedad.
- **Integración de Enfoques:** Investigar la posibilidad de combinar los enfoques de machine learning y deep learning para aprovechar las fortalezas de ambos y lograr una detección más robusta y precisa.

Al seguir estos pasos, se puede avanzar hacia un sistema más completo y personalizado para la detección temprana de la enfermedad de Alzheimer, ofreciendo a los profesionales de la salud herramientas más efectivas y contribuyendo al avance en la investigación médica.

6. Referencias

- [1] Prince, M., Bryce, R., & Ferri, C. (2011). Alzheimer's Disease International World Alzheimer Report 2011 The benefits of early diagnosis and intervention. *Alzheimer's Disease International*. <https://www.alzint.org/u/WorldAlzheimerReport2011.pdf>

- [2] Velazquez, M., Lee, Y., & Initiative, A. D. N. (2021). Random Forest model for feature-based Alzheimer's Disease conversion prediction from early mild cognitive impairment subjects. *PLOS ONE*, 16(4), e0244773. <https://doi.org/10.1371/journal.pone.0244773>