MODELO PREDICTIVO PARA LA CLASIFICACIÓN DEL RIESGO DE ALZHEIMER MEDIANTE VARIABLES CLINICAS Y BIOMARCADORES EN PLASMA

Business Analytics

Alumno: Nicolás Montesinos Recio

Tutor: Ana Lazcano de Rojas

Grado en Business Analytics

Curso académico 2024-2025

No me estoy rindiendo, solo me estoy dejando llevar.

- *Glen Campell (2011)*

Agradecimientos

Después de haber terminado este proyecto y contar los días para acabar oficialmente el grado, quiero aprovechar este apartado para dar las gracias no solo a quienes me han ayudado directamente con este trabajo, sino también a todas las personas que han estado a mi lado en este “viaje” universitario.

En primer lugar, a la razón emocional que me impulso a llevar a cabo este proyecto: mis dos abuelas, que padecen de la enfermedad de Alzheimer. Simplemente, GRACIAS, en mayúsculas, por haber sido mi fuente de motivación constante y formar parte de lo que soy hoy. Es una enfermedad dura, que te roba los recuerdos y memorias con los tuyos. Sin embargo, creo firmemente que lo verdadero nunca se borra: no hay enfermedad que borre los sentimientos y estoy seguro de que estarían increíblemente orgullosas de mí.

A mis padres y a mis hermanos, gracias por apoyarme siempre, en cada decisión, sin importar el resultado. Por estar ahí en los buenos y más en los malos momentos.

A mi tutora, Dra. Ana Lazcano, gracias por tu cercanía, tu compromiso con los estudiantes y por acompañarnos como una verdadera guía a lo largo del camino.

También quiero agradecer al Dr. Julio Emilio Sandubete y Dra. María Jesús Gómez por su generosidad y disposición a ayudar en sus respectivas áreas.

A todas las personas que han estado en mi camino, mis amigos y compañeros de clase, sin ellos nada de esto se podría haber llevado a cabo.

Y, por último, a todas las familias que conviven con el Alzheimer. Son un verdadero ejemplo de su lucha y valentía. Entre todos, podemos seguir aportando nuestro grano de arena para encontrar soluciones y dar esperanza.

Gracias a todos.

Resumen

Este trabajo de fin de grado desarrolla un modelo predictivo para clasificar a pacientes según su riesgo de Alzheimer, utilizando datos clínicos y biomarcadores como P-tau181 y Neurofilamente Light (NfL). Se parte de un conjunto de datos real proveniente de la base de datos ADNI con más de mil pacientes y se abordan retos clave como el desbalanceo de clases, la inconsistencia temporal de las pruebas y la selección de variables relevantes. El proyecto se enfoca en un problema de clasificación donde se utilizan diferentes modelos de machine learning para comparar los resultados y analizar cual tiene mejor rendimiento en general.

Este enfoque busca contribuir a la detección temprana del Alzheimer, optimizando el uso de biomarcadores y herramientas digitales de análisis.

Palabras clave: Alzheimer, biomarcadores, clasificación, machine learning, detección temprana, modelo predictivo.

Abstract

This final degree work develops a predictive model to classify patients according to their risk of Alzheimer's disease, using clinical data and biomarkers such as P-tau181 and Neurofilament Light Chain (NfL). It starts from a real dataset from the ADNI database with more than a thousand patients and addresses key challenges such as class imbalance, temporal inconsistency of tests and selection of relevant variables. The project focuses on a classification problem where different machine learning models are used to compare the results and analyse which one performs better overall.

This approach seeks to contribute to the early detection of Alzheimer's disease, optimizing the use of biomarkers and digital analysis tools.

Key words: Alzheimer, biomarker, classification, machine learning, early detection, predictive model.

Índice

Tabla de Ilustraciones

Índice de Tablas

Lista de Acrónimos

Acrónimo	**Significado**
AD	Alzheimer's Disease (Enfermedad de Alzheimer)
CN	Cognitively Normal (Sujeto Cognitivamente Normal)
MCI	Mild Cognitive Impairment (Deterioro Cognitivo Leve)
ANN	Artificial Neural Network (Red Neuronal Artificial)
RF	Random Forest (Bosque Aleatorio)
RL	Logistic Regression (Regresion Logistica)
MLP	Multilayer Perceptron (Perceptrón Multicapa)
FNN	Feedforward Neural Network (Red Neuronal Feedforward)
ADNI	Alzheimer's Disease Neuroimaging Initiative
NfL	Neurofilament Light (Neurofilamento de cadena ligera)
P-tau181	Phosphorylated Tau 181 (Tau fosforilada en la posición 181)
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve (Área Bajo la Curva ROC)
MSE	Mean Squared Error (Error Cuadrático Medio)
ML	Machine Learning (Aprendizaje Automático)
SMOTE	Synthetic Minority Oversampling Technique
DNN	Deep Neural Network (Red Neuronal Profunda)
RNN	Recurrent Neural Network (Red Neuronal Recurrente)

1. INTRODUCCIÓN

A día de hoy, vivimos rodeados de tecnología. Guardamos miles de fotos en móviles, ordenadores o álbumes, convencidos que así preservamos nuestros recuerdos. Sin embargo, la cruda realidad es que, cuando la memoria empieza a desvanecerse, esas imágenes pierden su significado. El Alzheimer no borra los archivos digitales, sino que nos borra la historia vivida en carne y hueso.

Hoy en día, esta enfermedad neurodegenerativa afecta a más de 57 millones de personas en el mundo (World Health Organization [WHO], 2025). En España, alrededor de 800,000 personas conviven con ella, y cada año se diagnostican unos 40,000 nuevos casos (Cadena SER, 2025). A pesar de los avances médicos y científicos, la enfermedad de Alzheimer sigue sin tener clara su causa, lo que subraya nuestra limitada comprensión de sus causas y mecanismos. Esto hace aún más urgente la necesidad de desarrollar métodos eficaces para su diagnóstico temprano, lo que permitirá intervenir a tiempo y mitigar los efectos devastadores de la enfermedad.

El proyecto se centra en el desarrollo de un modelo predictivo para clasificar a pacientes con riesgo de Alzheimer utilizando datos clínicos y biomarcadores. El objetivo es mejorar el diagnóstico temprano y la personalización del tratamiento a través del análisis de patrones en estos datos utilizando modelos de Machine Learning.

2. MOTIVACIÓN

Gracias a la cantidad de datos clínicos y biomarcadores que se recopilan constantemente en todo el mundo, y al avance en la tecnología y los modelos de machine learning, a día de hoy, es totalmente posible realizar predicciones más precisas que nos permiten anticipar el resigo de padecer diferentes enfermedades. En este contexto, y dada la creencia popular de personas que son diagnosticadas con Alzheimer, surge la motivación de este trabajo: desarrollar un modelo predictivo que permita predecir el riesgo de padecer Alzheimer en base a una serie de variables clínicas y dos biomarcadores. De tal forma, que podamos contribuir a una posible mejora de diagnóstico temprano y la toma de decisiones clínicas.

3. ESTADO DEL ARTE

En esta sección, se revisarán diversos estudios previos que abordan el diagnóstico y la predicción del Alzheimer, con el fin de contextualizar el enfoque de esta investigación. A través de trabajos previos, se explorarán las metodologías utilizadas y los avances alcanzados en el uso de Machine Learning y Deep Learning para la clasificación temprana de la enfermedad, así como la identificación de posibles biomarcadores relevantes.

3.1 Crecimiento del uso de Deep Learning y Machine Learning en la clasificación del Alzheimer

Diversos estudios han demostrado un crecimiento significativo el uso de Deep learning y Machine Learning para la clasificación temprana del Alzheimer (AD) debido a la capacidad que hay de poder manejar grandes bases de datos. Smith et al. (2018) en una revisión sistemática de estudios publicados entre los años 2013 y 2018, evidenció que el 75% de los estudios utilizan exclusivamente redes neuronales convolucionales (CNN) para el diagnóstico de AD llegando a alcanzar precisiones del 96%.

Por otro lado, en 2019 Zhang et al. (2019) demostró que la combinación de auto-coficadores aplicados (SAE) con técnicas tradicionales de machine learning mejora la precisión en la predicción de la conversión de un paciente con (MCI) a Alzheimer (AD). Los SAE tienen la habilidad de poder extraer representaciones de datos más ventajosos, lo que permitió a estos modelos más tradicionales reducir su dimensión y poder capturar patrones complejos. Su enfoque llego a alcanzar una precisión de hasta 98,8% en la clasificación de AD y un 83,7% en la predicción de la conversión de MCI a AD.

3.2 Machine Learning para predecir la incidencia del Alzheimer

Un estudio realizado hecho por Kim et al. (2020) utilizando los datos administrativos de salud del National Health Insurance Service – National Sample Cohort (NHIS-NCS) en Corea del Sur exploro el uso también de técnicas de machine learning para predecir la incidencia del Alzheimer. El estudio incluyo a más de 40,000 personas mayores de 65 años. Se utilizaron varios modelos, entre ellos el Random Forest (RF) y Regresion Logistica (RL) para poder predecir la aparición de AD en un intervalo de 0 a 4 años. Los resultados para las

predicciones en el corto plazo de 0 años fueron bastante positivos. El RF demostró el mejor rendimiento con un área bajo la curva (AUC) de 0.898 al predecir la enfermedad en 0 años. Sin embargo, la precisión decreció de forma significativa una vez se extendió el horizonte temporal a 4 años con un AUC de 0.662.

3.3 Ventajas del Deep Learning para la clasificación del Alzheimer

Jones et al. (2017) destacaron que el uso de técnicas de deep learning como la CNN, lograron precisiones de hasta el 96% para la clasificación de AD, sin necesidad de procesar las imágenes para la selección de las características. Además, en el estudio se mostró como la combinación de deep learning con biomarcadores líquidos como el p-tau181 y neurofilamento ligero (NfL) mejoro significamente los resultados, llegando a alcanzar una precisión de hasta 98,8% en la clasificación de AD y un 84,2% en la predicción de deterioro cognitivo leve a AD.

Por el otro lado, Hugo Fernández Cobas et al. (2021) de la Universidad Oberta de Catalunya realizo un estudio sobre el análisis de los factores de riesgo de la enfermedad del Alzheimer y su detección temprana mediante Machine Learning el cual utilizo imágenes de resonancia magnética (MRI) como única fuente de datos aplicando una red neuronal convolucional (CNN). Para el tratado de datos y su preparación, así como, para llevar a cabo a cada uno de los análisis y predicciones que contiene el proyecto utilizó el software R y para la detección de posibles enfermos mediante imágenes de resonancia magnética utilizo TensorFlow (entorno de Python). En el proyecto, Hugo utilizo los siguientes algoritmos de machine learning: K-NN, Naive Bayes, Red Neuronal Artificial, SVM, Árbol de decisión y Random Forest. Hugo al final del estudio recalca que lo más probable que si se hubiera tenido en cuenta pruebas neurológicas o marcadores de fluidos daría mejor resultado.

3.4 Uso del dataset ADNI para la clasificación y predicción

No obstante, un estudio realizado por Wee et al. (2023) utilizo datos de la misma base de datos de nuestro proyecto (ADNI) para evaluar la efectividad de algoritmos de aprendizaje automático en la clasificación de individuos cognitivamente normales (CN) frente a pacientes con Alzheimer (AD), y para predecir la conversión de deterioro cognitivo

leve (MCI) a AD. Sin embargo, este estudio utilizo cuatro conjuntos de características basadas en volúmenes hipocampales y de 47 regiones corticales y subcorticales, con y sin datos demográficos utilizando un biomarcador genético APOE4. Se utilizaron 6 diferentes modelos, entre ellos arboles de decisiones, random forest y SVM. Sin embargo, fue un modelo discriminante lineal en conjunto fue el más preciso para clasificar CN frente a AD llegando a una precisión del 92,8% y predecir la progresión de MCI a AD con precisiones que aumentan hasta un 77% en 48 meses.

3.5 Redes Neuronales Artificiales (ANN) para diagnóstico y progresión del Alzheimer

Mas allá de los clasificadores utilizados normalmente por los investigadores, un estudio realizado por Wang et al. (2020) utilizo datos de una base de datos comunitaria de más de 2,400 personas mayores utilizo un modelo de redes neuronales artificiales (ANN). Se utilizaron 89 casos nuevos de AD y 178 CN. Se analizaron una variedad de biomarcadores de orina en sangre (AD7c – NTP), funciones neuropsicológicas, historial médico y características demográficas entre otras. No obstante, descubrieron que el modelo ANN super a otros modelos como la regresión logística, K – Neighbours y SVM, logrando una precisión del 92.13%, una sensibilidad del 87,28% y un área bajo la curva (AUC) de 0,875. Las conclusiones que sacaron fueron que el riesgo de desarrollar Alzheimer con una mayor edad, menor nivel educativo, ingresos familiares bajos, antecedentes familiares y falta de actividad física. Destacaron la falta que hace de tener más investigación en Aβ42 en sangre y como el AD7c-NTP es clínicamente valioso para el diagnóstico temprano.

4. MARCO TEORICO

4.1 La Enfermedad de Alzheimer

El Alzheimer (AD) es un trastorno neurodegenerativo progresivo que afecta principalmente a personas de tercera edad, provocándoles un deterioro gradual en las funciones cognitivas, especialmente de la memoria, el lenguaje y el razonamiento. Actualmente, es una de las principales preocupaciones de salud mundial y se considera la

forma mas común de demencia, representando entre el 60% y el 80% de los casos diagnosticados (Alzheimer's Association, 2023).

Se trata de una enfermedad incurable que, hasta la fecha, no cuenta con un tratamiento capaz de detener su aparición ni frenar su progresión. Desde el momento del diagnóstico, la mediana de supervivencia estimada es aproximadamente 3 a 3.5 años, según el tipo de diagnóstico realizado (Wolfson et al., 2001).

El avance del Alzheimer conlleva una degeneración progresiva del cerebro. Por ello, es imprescindible clasificar a los pacientes en función del estado de evolución de la enfermedad. Esto no solo facilita una mejor compresión clínica del paciente, sino que nos permite poder adaptar los tratamientos en base a su situación y necesidades en cada etapa. Generalmente, se instituyen tres categorías:

- Cognitivamente Normal (CN): Personas sin deterioro cognitivo
- Deterioro Cognitivo Leve (MCI): Etapa intermedia con síntomas leves que no impiden una vida autónoma
- Alzheimer (AD): Etapa avanzada con deterioro cognitivo severo que afecta significamente la funcionalidad diaria.

Es fundamental tener estas tres categorías claras, ya que son las que tomaremos como referencia a lo largo de nuestro estudio.

El Alzheimer afecta principalmente a las regiones relacionadas con la memoria y las funciones cognitivas superiores, como el hipocampo y la corteza cerebral. Existen hipótesis sobre sus mecanismos, como la acumulación de placas beta – amiloide, que forma placas seniles en el cerebro y la proteína tau, que se agrupa en los llamados ovillos neurofibrilares. Estas alteraciones estructurales interfieren con la comunicación entre las neuronas, alterando las conexiones sinápticas y, en última instancia, provocando la muerte celular. El resultado de este daño es el deterioro gradual de la memoria y a media que la enfermedad avanza, la atrofia cerebral se intensifica, afectando aún más las funciones motoras y emocionales.

4.2 Biomarcadores en el diagnóstico del Alzheimer

Teniendo ya una idea sobre que es el Alzheimer y una posible hipótesis sobre qué es lo que lo causa, es esencial entender el concepto de un biomarcador y cuales vamos a usar en nuestro estudio.

Un biomarcador es una característica biológica que puede medirse objetivamente y que nos indica un proceso biológico o patológico. Los biomarcadores pueden estar presentes en diferentes muestras biológicas, como sangre, orina o líquido cefalorraquídeo (LCR)[1], y su uso nos permite obtener información temprana sobre la enfermedad, incluso antes de que aparezcan los síntomas clínicos. Ente los distintos tipos de biomarcadores, los biomarcadores plasmáticos, es decir, aquellos que pueden detectarse en el plasma de la sangre[2], han ganado protagonismo en los últimos años. Estos permiten acceder a la información relevante sobre el estado cerebral del paciente mediante un procedimiento mínimamente invasivo, como un simple análisis de sangre.

En nuestro estudio, hemos seleccionado dos biomarcadores plasmáticos clave para el diagnóstico del Alzheimer: la proteína tau fosforilada en la posición 181 (P-tau181) y el neurofilamento de cadena ligera (NfL).

4.2.1 P – Tau181

P-tau181 es una forma anómala de la proteína tau, una proteína que, en condiciones normales participa en la estabilización del citoesqueleto neuronal, una red interna de filamentos proteicos que da forma y soporte a las neuronas. Esta estabilización se regula mediante un proceso conocido como fosforilación. La fosforilación es un proceso normal en el que se añaden pequeños grupos químicos (llamados fosfatos) a estas proteínas para regular su actividad.

[1] Es un fluido claro que rodea y protege el cerebro y la médula espinal. Este líquido circula a través de los ventrículos cerebrales, el espacio subaracnoideo (entre las membranas que cubren el cerebro) y la médula espinal.

[2] Cuando hablamos del plasma de la sangre, nos referimos a la parte líquida y transparente de la sangre que queda después de retirar las células sanguíneas (glóbulos rojos, glóbulos blancos y plaquetas).

Sin embargo, en la enfermedad de Alzheimer, sufre un proceso anómalo de hiperfosforilación, es decir, se añaden demasiados grupos de fosfato de forma incorrecta, lo que provoca que se despeguen los microtubos del citoesqueleto neuronal y se su agreguen en el interior de las neuronas acumulando versiones anómalas de la proteína tau hiperfosforilada llamados ovillos neurofibrilares.

Una forma sencilla de entender este proceso es imaginándose que la proteína tau actúa como el pegamento que mantiene firmes los andamios (citoesqueleto neuronal) de una construcción. Si se aplica la cantidad correcta de pegamento (fosforilación), la estructura se mantiene estable y funcional. Pero si empiezas a aplicar pegamento de mas o en lugares equivocados, los andamios se vuelven pegajosos, inestables y terminan colapsando. En el caso del Alzheimer, la tau recibe demasiados grupos fosfato, pierde su capacidad de estabilizar los microtúbulos y comienza a despegarse, lo que provoca el desorden estructural y la acumulación de ovillos neurofibrilares en el interior de la neurona lo que causa que estos montones de tau hiperfosforilada dentro de la neurona bloquen el transporte neuronal. Esto es el misterio del Alzheimer, el entender por qué ocurre esta hiperfosforilación anómala de la proteína tau.

Para facilitar la comprensión de este proceso, se proporciona una imagen descriptiva del mecanismo descrito.

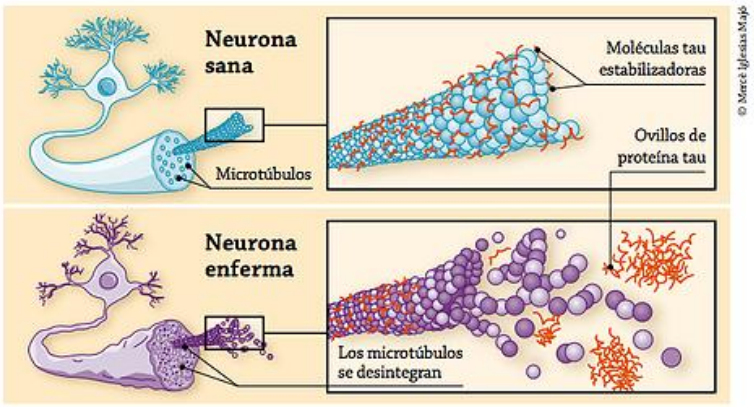

Ilustración 1: P-tau181 Neurona Sana vs Neurona Enferma. *Fuente: Visor de Libros (2020)*

El aumento de P-tau181 en sangre o en LCR se asocia específicamente con la presencia de ovillos neurofibrilares, una de las principales características patológicas del Alzheimer. Esta elevación se considera un indicador altamente especifico de la enfermedad y contribuye significamente a su diagnóstico.

Como se ha mencionado previamente, existen dos biomarcadores ampliamente estudiados y considerados entre los más relevantes en el diagnóstico del Alzheimer: P-tau181 y la relación Aβ42/40 (Abeta42/40). En nuestro caso hemos optado por incluir P-tau181 debido a su alta especificidad en la detección de ovillos neurofibrilares. Sin embargo, en lugar de usa la relación Aβ42/40, se ha decidido incorporar el biomarcador Neurofilamento de cadena ligera (NfL). Esto debido a que se le quería dar un enfoque personal y diferenciador con el objetivo de proporcionar una perspectiva distinta al análisis del Alzheimer.

4.2.2 Neurofilamento de cadena ligera (NfL)

El neurofilamento de cadena ligera (NfL) es una proteína estructural que al igual que el P-tau181 forma parte del citoesqueleto de las neuronas, específicamente en los axones, que son las prolongaciones encargadas de transmitir los impulsos eléctricos entre las células nerviosas.

El NfL contribuye a mantener la estabilidad, forma y funcionalidad del axón. En condiciones normales, esta proteína es esencial para preservar la integridad estructural del sistema nervioso central y asegurar que la comunicación entre las neuronas sea correcta. También participa en el transporte de sustancias dentro de la neurona, asegurando su funcionamiento adecuado.

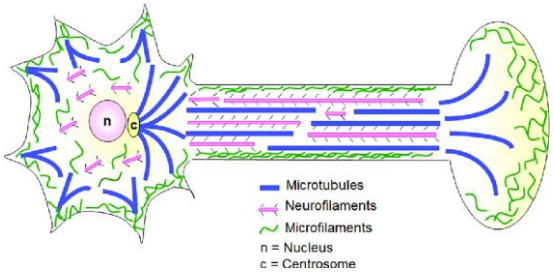

Ilustración 2: NfL. *Fuente: Research Gate (2022)*

Sin embargo, en el contexto de enfermedades neurodegenerativas como el Alzheimer, a medida que la neurona comienza a sufrir daños patológicos como la acumulación de proteínas toxicas (por ejemplo, tau hiperfosforilada), este deterioro afecta especialmente a los axones. Como parte del citoesqueleto axonal, el neurofilamento de cadena ligera (NfL) se descompone y se libera al espacio extracelular cuando la neurona se daña o muere. Desde allí, puede propagarse hacia el líquido cefalorraquídeo (LCR), que rodea el cerebro y la medula espinal y posteriormente llega a la sangre.

Por este motivo, al igual que con el biomarcador P-tau181, la presencia de niveles elevados de NfL en fluidos biológicos se considera un marcador directo pero sensible del daño neuronal y por lo tanto de la progresión de enfermedades neurodegenerativas. Esta afirmación se respalda con un estudio publicado por Mattsson et. al (2021) en Jama Neurology en la que se examinó la relación entre NfL plasma y la progresión clínica en pacientes de Alzheimer. Los investigadores encontraron que niveles elevados de NfL en plasma estaban asociados con una mayor tasa de deterioro cognitivo y funcional.

Cabe señalar que el NfL no es un biomarcador específico del Alzheimer, ya que su evaluación puede observarse en otras patologías neurodegenerativas. Esto nos aporta un valor extra a nuestro proyecto debido a que cuando se combina con otros indicadores más específicos, como el P-tau181, permite mejorar el diagnostico diferencial y la precisión en la identificación temprana de la enfermedad porque tenemos dos perspectivas distintas. El NfL indica el daño neuronal de forma general mientras que el P-tau181 está directamente más relacionado con la fisiopatología específica del Alzheimer permitiéndonos diferenciarla de otras enfermedades neurodegenerativas, como la demencia frontotemporal o la enfermedad de Parkinson.

4.3 Explicación teórica de los modelos a utilizar

4.3.1 Estrategia de Analisis - Machine Learning

Con el objetivo de resolver nuestro problema principal, vamos a hacer uso de una rama de la inteligencia artificial llamada Machine Learning. Esta es un campo de la inteligencia artificial que permite a las maquinas aprender de los datos a través de algoritmos y hacer predicciones o decisiones sin necesidad de una programación explicita.

4.3.2 Modelos Supervisados

La decisión de emplear solamente modelos supervisados en lugar de modelos no supervisados se basa principalmente en la naturaleza de nuestro problema y nuestro conjunto de datos. Los modelos supervisados requieren que los datos estén etiquetados, es decir, que se conozcan las categorías de salida para cada observación. En nuestro caso, contamos con etiquetas claras y definidas como CN, MCI y AD, lo que hace que modelos supervisados como Regresion Logistica, Random Forest o ANN sean más adecuados para este tipo de casos. Gracias a conocer las etiquetas, los modelos van a ser capaces de aprender las características de cada paciente (como los biomarcadores o los datos clínicos) y por ente, realice predicciones sobre la categoría de diagnóstico más probable.

4.3.3 Enfoque complementario de análisis

El objetivo principal de esta fase es no solo clasificar correctamente a los pacientes, sino también identificar qué características tienen mayor impacto en la predicción del riesgo de Alzheimer. Se investigará cómo ciertos biomarcadores, como P-tau181 y NfL, o variables categóricas como MH14ALCH podrían estar relacionados con el avance de la enfermedad, verificando empíricamente su relación con el diagnóstico mediante los modelos seleccionados. Para ello, se han escogido los modelos que cubren tanto nuestro objetivo principal como los secundarios.

4.3.4 Regresión Logistica Multinomial

La regresión logística multinomial es una extensión de la regresión logística apropiado para problemas de clasificación donde la variable dependiente Y puede tener tres

o mas clases no ordenadas. Dado un conjunto de predictores X, estimas las probabilidades para cada una de las categorías de la variable salida. La variable dependiente debe ser una variable categórica, mientras que las variables independientes pueden ser factores o covariables. En general, los factores deben ser variables categóricas y las covariables deben ser variables continuas.

Dado un conjunto de variables independientes $X = (x_1, x_2, x_3, \dots, x_p)$ y una variable Y que puede tomar J clases o categorías posibles, la regresión logística multinomial permite modelar la probabilidad de pertenencia a cada una de las clases. Estima la probabilidad de que una observación (un paciente, por ejemplo) pertenezca a la clase $j = 1, \dots, J$, usando una clase de referencia J.

La regresión logística multinomial se basa en el logit, que es el algoritmo de la razón de probabilidad (odds ratio). Es decir, la razón de probabilidad de un evento es la probabilidad de que el evento ocurra, dividida por la probabilidad de que no ocurra. Si tenemos una variable dependiente Y con J categorías se define un conjunto de log-odds para cada una de las clases j respecto a una clase base J. El modelo sigue la siguiente forma (1):

$$\log\left(\frac{\pi_j}{\pi_j}\right) = \alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jK}X_k$$

Ecuación 1: Logit Formula

π_j representa la probabilidad de que una observación pertenezca a la clase j. π_j, por el otro lado, es la probabilidad de que una observación pertenezca a la clase base J.

Este cálculo se realiza para cada clase $j = 1,2, \dots, J-1$, siendo a la clase J la referencia o base. La fórmula del logit nos da un valor que es un logaritmo de la relación de las probabilidades, pero no es una probabilidad en sí misma. Para obtener las probabilidades reales de pertenecer a cada clase, aplicaremos la función softmax. En este caso los valores que nos interesan son π_j ya que nos dice la probabilidad de que una observación pertenezca a cada clase.

Formula softmax:

Esta fórmula transforma los logits en probabilidades reales, que suman 1.

$$\pi_j = \frac{e^{\log\left(\frac{\pi_j}{\pi_j}\right)}}{1 + \sum_{i=e}^{J-1} e^{\log\left(\frac{\pi_i}{\pi_j}\right)}}$$

Ecuación 2: Softmax

Para cada clase j, usamos el valor del logit calculado previamente y aplicamos la función exponencial e^x para obtener una probabilidad no normalizada. Luego, sumamos todas las probabilidades no normalizadas y dividimos cada una de ellas por esta suma para normalizar los resultados y asegurar que las probabilidades de todas las clases sean igual a 1

4.3.5 Random Forest Multinomial

El Random Forest (RF) es un algoritmo compuesto por un conjunto de árboles de decisión denominados como bosque que se entrenan de forma independiente. Al igual que la regresión logística multinomial, este tipo de algoritmo se emplea cuando la variable dependiente tiene más de dos clases ($J > 2$). Se busca modelar la función de predicción $\hat{y}$ para una variable categórica Y con J categorías, dada un conjunto de variables independientes (predictoras) $X = (x_1, x_2, x_3, \dots, x_p)$.

Cada clase $j \in \{1,2 \dots J\}$ tiene una probabilidad de ser asociada de que una observación i (un paciente en nuestro caso) pertenezca a la clase j, condicionada por los valores de la variable independiente.

Arboles de Decisión

Antes de nada, debemos de entender que es un árbol de decisión y que función tiene en el random forest. Los árboles de decisión son un modelo que divide recursivamente el espacio de las variables predictoras X en varias regiones, tomando decisiones basadas en el valor de una variable X_k para una observación dada. En el caso de la clasificación multinomial, cada una de estas regiones es representada por un nodo (hoja), y cada hoja está asociada con una clase j de la variable Y. La clase de un nodo es, por lo tanto, determinada en función de la clase más frecuente entre los datos que llegan a ese nodo.

El objetivo del árbol de decisión en este contexto es encontrar un conjunto de reglas de partición[3] $T = (t_1, t_2, t_3, \dots t_n)$ que dividan el espacio de las características en varias regiones. Cada partición t_n corresponde a una condición sobre las variables independientes que divide el espacio en dos subregiones disjuntas. Básicamente, el modelo crea una jerarquía de particiones basadas en condiciones sobre las variables predictoras, y a media que una observación va recorriendo el árbol, se le asigna una de las clases en función de las particiones por las que pasa.

Proceso de entrenamiento

El proceso de entrenamiento se basa en la técnica de modelos ensamblaje llamada Bagging. La idea es crear un conjunto de árboles independientes y entrenados sobre subconjuntos del conjunto de entrenamiento, mediante muestreo con reemplazo. La división porcentual de entrenamiento y test se describirá en la metodología.

Dado un conjunto de datos de entrenamiento $D = \{(X_i, Y_i)\}_{i=1}^{n}$, se genera un subconjunto para cada árbol $D_b = \{(X_i^b, Y_i^b)\}_{i=1}^{m}$ para cada árbol b, donde X_i^b son las observaciones seleccionadas al azar con reemplazo, e Y_i son sus etiquetas correspondientes. Este proceso es repetido para B arboles sucesivamente, generando un bosque de árboles de decisiones.

Predicción en Random Forest Multinomial

Una vez que ya se tiene el conjunto de arboles $\{T_1, T_2, T_3, \dots T_B\}$, cada árbol predice una clase (CN, MCI o AD) para una nueva observación X_i. El Random Forest recoge todas las predicciones de los árboles y predice la clase final mediante el voto mayoritario de estos. De tal forma que se expresa de la siguiente forma:

$$\widehat{Y_i} = \arg max_j \left(\sum_{b=1}^{B} I(\widehat{Y_i}^{(b)} = j)\right)$$

Ecuación 3: Random Forest Y

[3] Las reglas de partición en los árboles de decisión son condiciones o decisiones que se utilizan para dividir el espacio de las características o variables predictoras en subregiones más pequeñas

Donde $I(\hat{Y}_i^{(b)} = j)$ es una función indicadora que toma el valor 1 si el árbol B predice la clase j y 0 si no la predice y la predicción final $\hat{Y}_i$ es la clase j que tiene la mayor cantidad de votos.

Sin embargo, además de la clasificación de la observación, Random Forest multinomial puede estimar la probabilidad de X_i observación, pertenezca a una clase j. Para ello, se calcula la fracción de arboles que predicen esa clase (2).

$$\hat{\pi}(X_i) = \frac{1}{B}\sum_{b=1}^{B} I((\hat{Y}_i^{(b)} = j)$$

Ecuación 4: Random Forest Fracción

4.3.6 Red Neuronal Artificial (ANN – FNN/MLP)

4.3.6.1 Estructura

Las Redes Neuronales Artificiales (ANN) son modelos computacionales inspirados en el funcionamiento del cerebro humano. Están compuestas por una capa de entrada, una o varias capas ocultas y una capa de salida la cual es responsable de mostrar el resultado.

Cuando estas redes incluyen múltiples capas ocultas, pasan a formar parte del campo de Deep Learning, una rama del Machine Learning representada por el uso de arquitecturas mas profundas y capaces de capturar patrones complejos y no lineales.

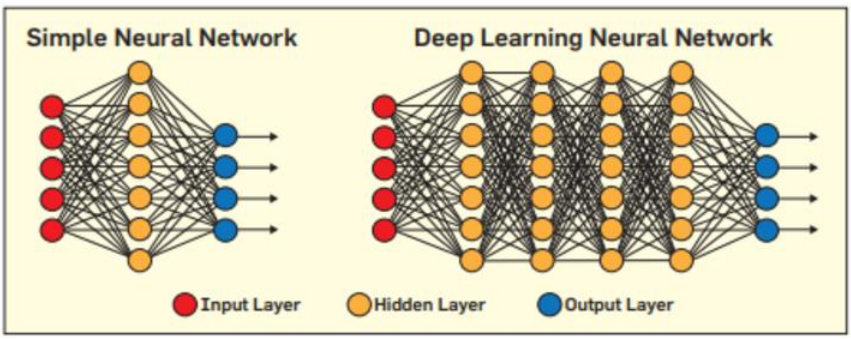

Ilustración 3: ANN. *Fuente: Research Gate (2023)*

En una red neuronal completamente conectada, cada nodo (unidad básica de procesamiento de la red neuronal) está conectado a todas las neuronas de la capa anterior. La salida (output) de los nodos de la capa previa viajan hasta la siguiente capa como la entrada (inputs) para esos nodos. Para comprender mejor el funcionamiento de las redes neuronales, resulta útil visualizar cada neurona como una regresión lineal que recibe una serie de entradas (3).

$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 \ldots + bias$$

Ecuación 5: Red Neuronal Lineal

w_i representa el peso asignado a cada entrada, que puede interpretarse como la importancia de esa entrada para la activación de la neurona

x_i corresponde al valor de salida de la neurona de la capa anterior

El bias es el termino independiente

Cada salida lineal generada por la neurona mediante la *Ecuación 5: Red Neuronal Lineal* se transforma mediante una función de activación. Su cargo es introducir no linealidad en el modelo y determina si esa neurona se activa o no. Un ejemplo muy común de la función de activación es la función escalón (4):

$$f(x) = \begin{cases} 1, & si \sum w_i x_i + bias \geq 0 \\ 0, & si \sum w_i x_i + bias \leq 0 \end{cases}$$

Ecuación 6: Función de Activación

No obstante, en la práctica se hace uso de otro tipo de funciones de activación, las cuales son más suaves, como ReLU, sigmoide o tanh. Hay que aclarar que se usara una u otra dependiendo del modelo y el tipo de objetivo:

ReLU (1): Transforma cualquier valor de entrada negativo en cero y deja pasar sin cambios los valores positivos. De tal forma que, las redes neuronales no lineales aprendan relaciones no lineales al activar solo ciertas neuronas según la magnitud de la entrada.

$$f(x) = \max(0, x)$$

Ecuación 7: ReLU

Sigmoide (2): Convierte cualquier valor real en un rango entre 0 y 1, lo que la hace útil para representar probabilidades.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Ecuación 8: Sigmoide

- Tanh (3): Transforma los valores en un rango de -1 a 1, siendo útil para centrar los datos

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Ecuación 9: Tanh

4.3.6.2 Entrenamiento del Modelo

Una vez ya habiendo comprendido lo esenciales del funcionamiento de una red neuronal, es fundamental entender entrenamiento. Este consiste en que la red neuronal busque encontrar los valores óptimos para los pesos de cada entrada y el sesgo de cada neurona, con el objetivo de minimizar el error en las predicciones. El entrenamiento de una red neuronal se basa en dos conceptos principales: propagación hacia adelante y propagación hacia atrás.

Las redes neuronales trabajan con información mediante propagación hacia adelante, alimentando al neuron de la siguiente capa. Durante el entrenamiento, los pesos y los sesgos se inicializan aleatoriamente. Luego, se evalúa el error del modelo usando la función de error cuadrático medio (MSE), que comparara la salida obtenida $\hat{y}$ con la salida observada y (deseada) para ajustar los parámetros. Para facilitarles la compresión les pongo una ilustración de como seria en nuestro contexto.

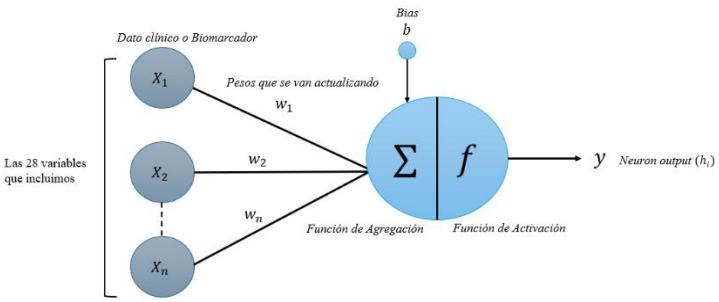

Ilustración 4: Proceso de Neurona (Caso de Uso) *Fuente: Elaboración propia*

Función del error (MSE) (4)

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{y}-y)^2$$

Donde:

- m representa el número de muestras usadas
- $\hat{y}$ valor predicho
- y valor deseado

Ecuación 10: MSE

La función Mean Square Error (MSE) mide el rendimiento de una red neuronal calculando la medida de los errores al cuadrado entre la salida generada y la esperada. Se define como $MSE = (\hat{y} - y)$. Por lo tanto, un valor bajo indicaría que el modelo esta mejorando su entrenamiento. El objetivo, durante la fase de entrenamiento es minimizar esta función ajustando los pesos y los sesgos de cada neurona mediante el algoritmo de descenso del gradiente (5).

$$gradiente = \frac{\vartheta(MSE)}{\vartheta w(h)} = \frac{\vartheta(MSE)}{\vartheta(h)} * \frac{\vartheta(h)}{\vartheta w(h)}$$

Ecuación 11: Gradiente (ANN)

Este algoritmo, calcula el gradiente, lo que es la derivada de la función del error respecto a cada peso. Así, determina en que dirección (positiva o negativa) deben modificarse los parámetros para acercarse al mínimo del error. Si el gradiente es positivo, estará hacia la izquierda; si es negativo, el mínimo esta hacia la derecha.

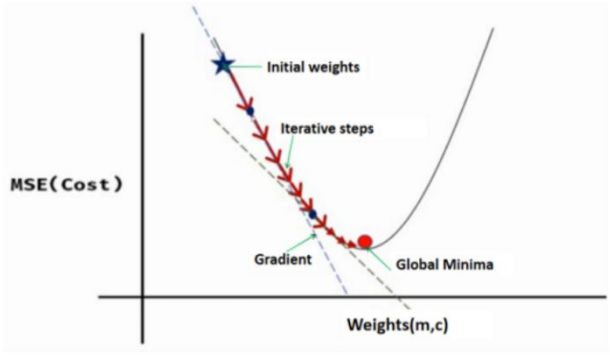

Ilustración 5: Derivada (ANN) *Fuente: Towards Data Science (2025)*

Una vez se realiza la propagación hacia adelante, y se obtiene el error, comienza la retropropagacion, donde se van actualizando los pesos y sesgos desde la capa de salida hacia

la de entrada. Para lograrlo, se utiliza la regla de la cadena, dividiendo la derivada en partes más simples por cada capa. Este proceso se repite sucesivamente en varias iteraciones, optimizando progresivamente los datos (6).

$$w(h)^{(t+1)} = w(h)^t - \eta \times \frac{\partial(MSE)}{\partial w_h}$$

Ecuación 12: Actualización de pesos (ANN)

Donde:

En cada iteración del entrenamiento, $w(h)^t$ representa el peo actual del modelo, mientras que $w(h)^{(t+1)}$ es el peso actualizado tras aplicar el ajuste. Este ajuste se realiza restando el producto del gradiente $\frac{\partial(MSE)}{\partial w_h}$, que indica como varia el error con respecto al peso, por la tasa de aprendizaje η, que controla la magnitud de cambio.

4.4 Objetivos de Analisis

4.4.1 Objetivo principal

Desarrollar un modelo predictivo que clasifique a los pacientes según su estado cognitivo (CN, MCI, o AD) utilizando datos clínicos y biomarcadores.

4.4.2 Objetivos secundarios

- Determinar la relevancia de los biomarcadores P-tau181 y NfL en el diagnóstico temprano de la enfermedad
- Comparar el rendimiento de distintos modelos de Machine Learning para ver cual ofrece mejor equilibrio entre precisión, sensibilidad y especificidad
- Identificar las variables más relevantes y el peso de los coeficientes a través de la Regresion Logistica Multinomial y el Random Forest Multinomial.
- Aplicar técnicas de procesamiento adecuadas, como el balanceo de clases mediante downsampling y la normalización de algunas variables para mejorar el rendimiento del modelo

- Desarrollar una solución de negocio una vez llevado a cabo la parte de modelización para poder dar soluciones “reales” al sector de la medicina.

5. INGENIERIA DEL DATO

5.1 Datos Seleccionados

En este caso, se ha seleccionado el estudio Alzheimer's Disease Neuroimaging Initiative (ADNI), el cual se encuentra disponible a través de IDA (Image & Data Archive) de la Universidad del Sur de California. Estos datos no están publicados para cualquier usuario. Se debe mandar una solicitud con una justificación con el uso que se va a hacer de estos y tras una revisión por el centro se le dará una respuesta. Se han utilizado finalmente 6 informes; 2 de biomarcadores y 4 de datos clínicos generales.

Cada informe trata unos datos clínicos o resultados de biomarcador. En los biomarcadores, tanto para NFL (Neurofilament Light Chain) como Ptau-181(tau fosforilada en el sitio 181), las fechas de prueba son las mismas, por lo que para cierto paciente que se haga una de las dos pruebas en x fecha, esa fecha será la misma para el otro biomarcador (Examdate Ptau181 = Examdate NFL y viceversa). Sin embargo, en los informes de los datos clínicos, hay algunas fechas de los pacientes que no coinciden (Informe del diagnóstico) o tienen alguna visita médica más en un informe que otro. Esta situación es de máxima importancia que sea explicada.

Cuando combinamos los diferentes informes, nos encontramos con que las fechas de las pruebas no coinciden exactamente entre cada uno de los informes. Por ejemplo, un paciente con un RID = 2 puede tener varias fechas de examen o de visitas diferentes en cada informe, lo que significa que, al juntar los datos, no todas las fechas se solapan. Un informe puede contener fechas como 17/08/2005, 22/09/2010 y 19/09/2011, mientras que otro informe puede tener 4 fechas adicionales, pero sin coincidencias exactas en las fechas.

Este desajuste de fechas obliga a generar más líneas de datos al combinar informes. NO se están creando duplicados, sino registros adicionales para poder capturar toda la información relevante. Es decir, al juntar los informes, aunque los identificadores de paciente (RID) sean los mismos, la falta de coincidencia en las fechas de prueba o visita causa que

para un mismo RID tengamos múltiples registros por cada fecha de examen o visita con el fin de asegurar que cada información específica, como el diagnóstico, las pruebas de NFL o los resultados de Ptau181, se incluya correctamente para cada fecha distinta.

En nuestro caso, dado el volumen de los datos, he definido una clave primaria compuesta por cuatro variables. Un RID se puede repetir todas las veces que quiera ya que un paciente puede tener múltiples exámenes o visitas en diferentes fechas. Sin embargo, lo que realmente distingue a cada fila son las variables: PLASMA _NFL, EXAMDATE(PTAU181), PLASMAPTAU181 y VISDATE. Al utilizar estas cuatro variables como clave primaria compuesta, nos aseguramos de que cada combinación de RID + PLASMA_NFL + EXAMDATE(PTAU181) + PLASMAPTAU181 + VISDATE sea única en el conjunto de datos. Esto va a evitar la creación de registros duplicados y garantiza que cada observación refleje correctamente la información de un paciente en una fecha. Les pongo un ejemplo sacado del conjunto de datos.

Tabla 1: Vista de las primeras filas del conjunto de datos

RID	**…….**	**PLASMA_NFL**	**EXAMDATE (PTAU181)**	**VID**	**PLASMAPTAU181**	**VISDATE**
2		253	22/09/2010	9	11939	17/08/2005
2		253	22/09/2010	9	11939	22/09/2010
2		253	22/09/2010	9	11939	19/09/2011
2		253	19/09/2011	10	12936	17/08/2005
2		253	19/09/2011	10	12936	22/09/2010
2		253	19/09/2011	10	12936	19/09/2011
2		253	26/09/2012	11	13563	17/08/2005
2		253	26/09/2012	11	13563	22/09/2010
2		253	26/09/2012	11	13563	19/09/2011
2		253	09/09/2013	12	15506	17/08/2005
2		253	09/09/2013	12	15506	22/09/2010
2		253	09/09/2013	12	15506	19/09/2011
2		309	22/09/2010	9	11939	17/08/2005
2		309	22/09/2010	9	11939	22/09/2010
2		309	22/09/2010	9	11939	19/09/2011
2		309	19/09/2011	10	12936	17/08/2005
2		309	19/09/2011	10	12936	22/09/2010
……		…….	……..	……	……..	…….
4278		237	24/10/2013	6	19738	29/09/2011

Como se ve en *Tabla 1: Vista de las primeras filas del conjunto de datos*, el RID 2 parece estar constantemente repitiéndose una y otra vez, pero si nos fijamos con detalle en la tabla, nos podemos dar cuenta que cada fila es un registro único el cual una de las variables de la clave primaria siempre cambia, lo que permite generar registros únicos en lugar de duplicados.

Tabla 2: Observaciones de los diferentes informes

Informe	Observaciones
DiagnosticoF	10.417
SubjectDemographicsF	3.759
MedHistoryF	3.083
NFL	3.762
Ptau181	3.758

Una vez escogido las variable y características a usar de los diferentes informes, he realizado un INNER JOINT para poder juntarlos mediante power query. Al aplicar un INNER JOINT, significa que solo estamos conservando los registros donde hay coincidencias entre los diferentes informes. En nuestro caso, esto significa que solamente incluiremos aquellos pacientes que tengan registros en todos los informes combinados

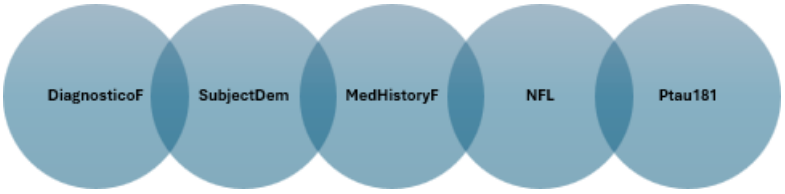

Ilustración 6: Inner Joint *Fuente: Elaboración propia*

Esto da como resultado el informe "FicheroJunto" el cual es el fichero donde se va a desarrollar el análisis.

Tabla 3: FicheroJunto Observaciones

Informe	Observaciones
FicheroJunto	766.807

El fichero contiene 766.807 observaciones, lo que indica que tenemos una gran riqueza de datos de las cuales podemos hacer uso para entrenar a los modelos ya que

contamos con más de 1200 pacientes únicos. Sin embargo, los datos clínicos suelen presentar una estructura compleja y tienden a tener un alto desbalance de clases. Por lo tanto, este volumen también implica la necesidad de un tratamiento de datos cuidadoso para poder evitar todo este tipo de problemas.

5.2 Datos clínicos

Además de los biomarcadores plasmáticos, en este proyecto se han incluido una serie de variables clínicas que aportan información que aportan información relevante sobre el estado cognitivo y funcional de los pacientes. Estas variables ayudan a contextualizar los resultados biológicos y por lo tanto mejorar la precisiones de nuestros modelos.

Tabla 4: N.º de variables, su tipo de datos, formato, periodo e intervalos

	Variable	**Tipo de dato**	**Formato**	**Periodo**	**Intervalo**
1	**RID**	int64	Identificador	2005-2019	[2 – 5296]
2	**EXAMDATE(D)**	datetime64[ns]	Fecha	2005-2019	-
3	**DIAGNOSIS**	int64	Categoría	2005-2019	[1 – 3]
4	**DXMPTR1**[1]	float64	Categoría	2005-2019	[-4 – 1]
5	**DXPARK**	float64	Categoría	2005-2019	[-4 – 1]
6	**VISDATE(N)**	datetime64[ns]	Fecha	2005-2013	-
7	**NXVISUAL**	int64	Categoría	2005-2013	[1 – 2]
8	**NXAUDITO**	int64	Categoría	2005-2013	[1 – 2]
9	**NXTREMOR**	int64	Categoría	2005-2013	[1 – 2]
10	**NXCONSCI**	int64	Categoría	2005-2013	[1 – 2]
11	**NXNERVE**	int64	Categoría	2005-2013	[1 – 2]
12	**NXMOTOR**	int64	Categoría	2005-2013	[1 – 2]
13	**NXFINGER**	int64	Categoría	2005-2013	[1 – 2]
14	**NXHEEL**	float64	Categoría	2005-2013	[1 – 2]
15	**NXSENSOR**	int64	Categoría	2005-2013	[1 – 2]
16	**NXTENDON**	int64	Categoría	2005-2013	[1 – 2]
17	**VISDATE(MEDH)**	datetime64[ns]	Fecha	2005-2013	-
18	**MH3HEAD**	int64	Categoría	2005-2013	[0 – 1]
19	**MH4CARD**	int64	Categoría	2005-2013	[0 – 1]
20	**MH5RESP**	int64	Categoría	2005-2013	[0 – 1]
21	**MH13ALLE**	int64	Categoría	2005-2013	[0 – 1]
22	**MH14ALCH**	int64	Categoría	2005-2013	[0 – 1]
23	**MH16SMOK**	int64	Categoría	2005-2013	[0 – 1]
24	**MH17MALI**	int64	Categoría	2005-2013	[0 – 1]
25	**EXAMDATE(NFL)**	datetime64[ns]	Fecha	2007-2016	-
26	**RECNO**	int64	Categoría	2007-2016	[1 – 2]
27	**PLASMA_NFL**	float64	Numero	2007-2016	[8 – 3056]

28	**EXAMDATE (PTAU181)**	datetime64[ns]	Fecha	2007-2016	-
29	**VID**	int64	Identificador	2007-2016	[2 – 14]
30	**PLASMAPTAU181**	float64	Numero	2007-2016	[10 – 451398]
31	**VISDATE**	datetime64[ns]	Fecha	2005-2015	-
32	**PTGENDER**	float64	Categoría	2005-2015	[-4 – 2]
33	**PTDOB**	datetime64[ns]	Categoría	2005-2015	-
34	**PTEDUCAT**	float64	Numero	2005-2015	[-4 – 20]

Para facilitar la comprensión, a continuación, se proporcionará una descripción detallada. Los identificadores y las fechas no se incluyen.

Tabla 5: Descripción de las variables

Variable	**Explicación**	**Categoría**
DIAGNOSIS	Clasificación del diagnóstico de un paciente en tres categorías	1: CN, 2: MCI, 3: DEMENTIA
DXMPTR1	Queja subjetiva de memoria	0: No, 1: Si
DXPARK	Síntomas de Parkinson presentes	0: No, 1: Si
NXVISUAL	Evaluación de la capacidad visual del paciente	1: Ausente, 2: Presente
NXAUDITO	Evaluación de la capacidad auditiva del paciente	1: Ausente, 2: Presente
NXTREMOR	Evaluación de temblores del paciente	1: Ausente, 2: Presente
NXCONSCI	Evaluación de la conciencia del paciente	1: Normal, 2: Anormal
NXNERVE	Nervios craneales	1: Normal, 2: Anormal
NXMOTOR	Evaluación de la función motora del paciente	1: Normal, 2: Anormal
NXFINGER	Evaluación de prueba “Finger-to-Nose”	1: Normal, 2: Anormal
NXHEEL	Evaluación de prueba “Heel-to-Shin”	1: Normal, 2: Anormal
NXSENSOR	Evaluación sensorial del paciente	1: Normal, 2: Anormal
NXTENDON	Evaluación de los reflejos tendinosos profundos	1: Normal, 2: Anormal
MH3HEAD	historial médico relacionado con la evaluación de la cabeza, ojos, oídos, nariz y garganta	0: No, 1: Si

MH4CARD	presencia o ausencia de problemas cardíacos o circulatorios.	0: No, 1: Si
MH5RESP	Problemas respiratorios	0: No, 1: Si
MH13ALLE	Información sobre alergias o sensibilidades	0: No, 1: Si
MH14ALCH	Consumo de alcohol	0: No, 1: Si
MH16SMOK	Tabaquismo	0: No, 1: Si
MH17MALI	Si el paciente tiene antecedentes de enfermedades malignas (cáncer)	0: No, 1: Si
PLASMA_NFL	Nivel de proteína NFL en sangre	
PLASMAPTAU181	Nivel de proteína pTau181 en sangre	
PTGENDER	Genero del paciente	1: Masculino, 2: Femenino
PTDOB	Año de nacimiento del paciente	

Aunque los archivos originales no especificaban las unidades de medida de los biomarcadores, si indicaban que las concentraciones fueron obtenidas mediante la técnica ultrasensible SIMOA (Single Molecule Array)[4]. Basándonos en esta metodología y en la magnitud de los datos, se pudo inferir que los valores de P-tau181 (en torno a los miles) corresponden a picogramos por milímetro (pg/mL), mientras que los valores menores de NfL (alrededor de los 400) eran consistentes con femtogramos por milímetro (fg/mL).

Tabla 6: Unidades Biomarcadores

Variable	Unidades
PLASMAPTAU181	fg/mL
PLASMA_NFL	pg/mL

5.3 Estudio estadístico

Este estudio estadístico nos permitirá resumir el conjunto de datos y proporcionar un valor representativo de la concentración de los valores. Dado que la mayoría de las variables

[4] Es una técnica de análisis ultrasensible que permite la detección y cuantificación de proteínas en concentraciones extremadamente bajas, incluso a nivel de una sola molécula.

categóricas son binarias (1 o 2, 0 o 1) como vimos en *Tabla 5: Descripción de las variables* calcular la media, mediana o moda no aporta información relevante. Por lo tanto, me concentrare en las variables numéricas continuas (como los biomarcadores, los años de educación) excluyendo el diagnostico que lo utilizaremos para segmentar los datos y extraer información más relevante para nuestro análisis.

5.3.1 Tendencia Central

Tabla 7: Tendencia Central Media

Variable	**Media**
PLASMA_NFL	479.151500
PLASMAPTAU181	18504.822410
DIAGNOSIS	1.733668
PTEDUCAT	15.551019

Como nuestra variable objetivo es el DIAGNOSIS, ya que determina el estado en el que se encuentra un paciente, puede ser relevante calcular la media de PLASMA_NFL, PLASMAPTUA181 y PTEDUCAT dentro de cada categoría de DIAGNOSIS (1,2 o 3).

Tabla 8: Media Por Diagnosis

DIAGNOSIS	**PLASMA_NFL**	**PLASMAPTAU181**	**PTEDUCAT**
1	434.28	18777.10	15.97
2	488.29	17492.87	15.30
3	580.58	21102.31	15.15

Según la *Tabla 8: Media Por Diagnosis* tanto el PLASMA_NFL y el PLASMAPTAU181 aumentan conforme avanza el diagnostico, con los pacientes en diagnostico 3 (etapa más avanzada) mostrando los valores más altos, lo que sugiere una mayor acumulación asociada al daño neuronal y la progresión del Alzheimer. Por otro lado, los años de educación son prácticamente los mismos entre los diferentes diagnósticos, indicando que el nivel educativo no varía considerablemente entre los grupos. Sin embargo, podría estar relacionado con la reserva cognitiva. Además, podemos analizar como varían los biomarcadores en función del diagnóstico y los hábitos del paciente, como el consumo del alcohol y el tabaquismo

Tabla 9: Media Por Diagnosis MH16SMOK etc.

DIAGNOSIS	MH16SMOK	MH14ALCH	PLASMA_NFL	PLASMAPTAU181
1	0	0	466.37	17104.49
		1	345.39	31935.19
	1	0	395.22	20641.30
		1	378.50	22876.53
2	0	0	469.02	17734.74
		1	357.48	15931.80
	1	0	523.23	17096.73
		1	380.03	18384.91
3	0	0	589.61	22993.13
		1	367.39	31370.85
	1	0	575.02	15518.22
		1	490.45	38148.01

Los datos provenientes de la tabla fortalecen la tendencia de que a medida los pacientes van pasando de una fase a otra, sus niveles de PLASMA_NFL y PLASMAPTAU181 aumentan. Esto es especialmente notable cuando se asocia con el consumo de alcohol (M14ALCH), lo que sugiere que tanto el avance de la enfermedad como el consumo de alcohol podría estar relacionado con niveles más elevados de estos indicadores. También, podemos ver como el uso de tabaquismo podría estar relacionado con un mayor nivel de los biomarcadores, lo que refuerza la idea de ciertos hábitos pueden influir en la progresión de la enfermedad y en los niveles de los biomarcadores.

5.3.2 Dispersión

Tabla 10: Dispersión

Variable	Desviación estándar
PLASMA_NFL	320.447707
PLASMAPTAU181	19149.489236
DIAGNOSIS	0.684974
PTEDUCAT	3.606689

La desviación estándar revela una alta dispersión en los biomarcadores, lo que sugiere una variabilidad entre los pacientes posiblemente asociado al avance de la enfermedad. Sin embargo, en la variable DIAGNOSIS muestra una baja desviación estándar

(0.68), lo que indica la concentración en una sola clase y lo cual significa que puede haber un posible desbalanceo en dicha clase.

5.3.3 Frecuencias

Tabla 12: Frecuencia Absoluta

Variable	Frecuencia
PLASMA_NFL	482 (5644)
PLASMAPTAU181	15634 (2298)
DIAGNOSIS	2 (352638)
PTEDUCAT	16 (147800)

Tabla 11: Frecuencia Absoluta Diagnosis

DIAGNOSIS	Frecuencia
1	309197
2	352638
3	104972

Según la *Tabla 11: Frecuencia Absoluta Diagnosis*, podemos ver tanto la superioridad en las variables PTEDUCAT como en el DIAGNOSIS en la etapa 2, con 352,638 observaciones, que representa aproximadamente la mitad de las observaciones dentro del fichero. Estos son claros indicios de un posible desbalanceo de clases en ambas categorías. Los datos clínicos suelen presentar un alto desbalanceo de clases. En el caso del ADNI (fuente de datos), la mayoría de los pacientes pertenecen a clases de control o a etapas iniciales de la enfermedad, mientras que los casos más graves o los que están en etapas avanzadas de Alzheimer son menos frecuentes.

Tabla 13: Frecuencia Relativa Diagnosis

DIAGNOSIS	Frecuencia relativa
1	0.459878
2	0.403227
3	0.136895

He puesto el ejemplo de la variable DIAGNOSIS ya que es de la que más importancia tiene en nuestro proyecto. Según la Tabla 23: Frecuencia relativa de DIAGNOSIS, se puede observar un desbalanceo claro en la etapa 3 del diagnóstico, con solo un 13% de las muestras. Este resultado es objetivamente bajo, pero son más de 100 000 observaciones. Hay que recordar que disponemos de un dataset con un volumen de datos muy alto. En nuestro contexto, el foco principal debe estar en equilibrar la variable DIAGNOSIS, ya que es la

etiqueta que nuestro modelo debe aprender para poder predecir el riesgo de un paciente (diferenciar entre CN, MCI y AD).

5.4 Análisis descriptivo

5.4.1 Distribución de los biomarcadores

Previo a la manipulación de nuestro conjunto de datos, se realizó un análisis exploratorio para comprender la estructura del dataset y respaldar el estudio estadístico. Se presentan visualizaciones sobre las distribuciones de variables clave. Además, se muestra el desbalance de clases original, que posteriormente se trató en la transformación de datos.

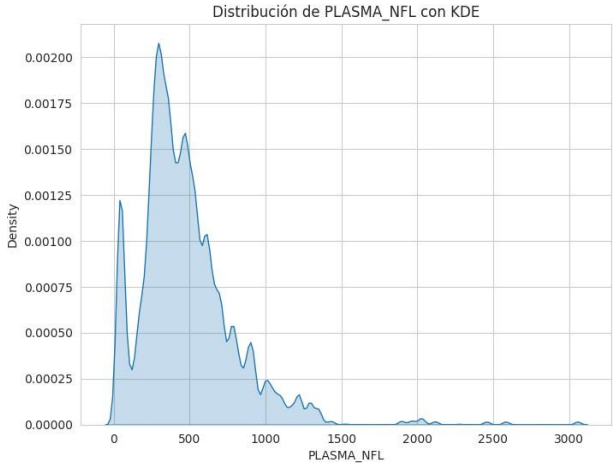

Ilustración 8: Distribución NfL *Fuente: Elaboración propia*

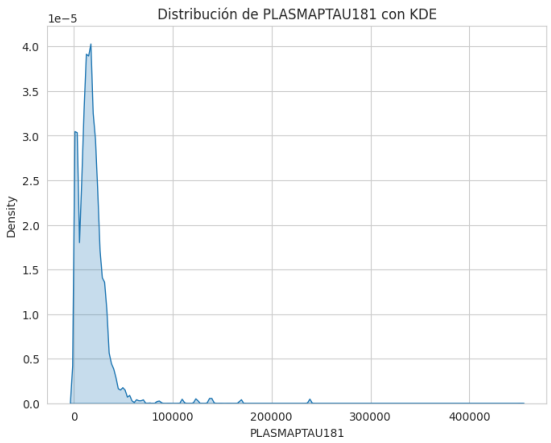

Ilustración 7: Distribución NfL *Fuente: Elaboración propia*

La distribución de PLASMA_NFL muestra una clara asimetría positiva, con la mayoría de los valores concentrados entre 250 y 500, lo cual concuerda con la media enseñada previamente (479.15). Sin embargo, la cola larga que alcanza hasta valores como 2000 o 3000 indica la presencia de posibles outliers. De forma similar, PLASMAPTUA181 también presenta una asimetría positiva, con la mayoría de los pacientes alrededor de 20.000, en línea con el estudio estadístico. No obstante, al igual que en el caso de PLASMA_NFL, tenemos la presencia de algunos valores atípicos extremos alrededor de los 100.000 y llegando más lejos. A primera vista, esto puede sugerir posibles inconsistencias dentro del dataset. Por ello, es clave analizar los outliers posteriormente para decidir si deben mantenerse o excluirse del análisis. En la sección de anexos se puede encontrar la

distribución de ambos biomarcadores en función de su diagnóstico como *Ilustración 19: Dispersión de Ptau181* e *Ilustración 20: Dispersión de NfL.Ilustración 21: Boxplot NfL*

5.4.2 Distribución de las variables categóricas

Con el objetivo de profundizar en las variables categóricas y analizar su comportamiento dentro de cada grupo diagnóstico, se han elaborado dos histogramas que permiten visualizar su distribución y detectar los posibles desequilibrios entre clases.

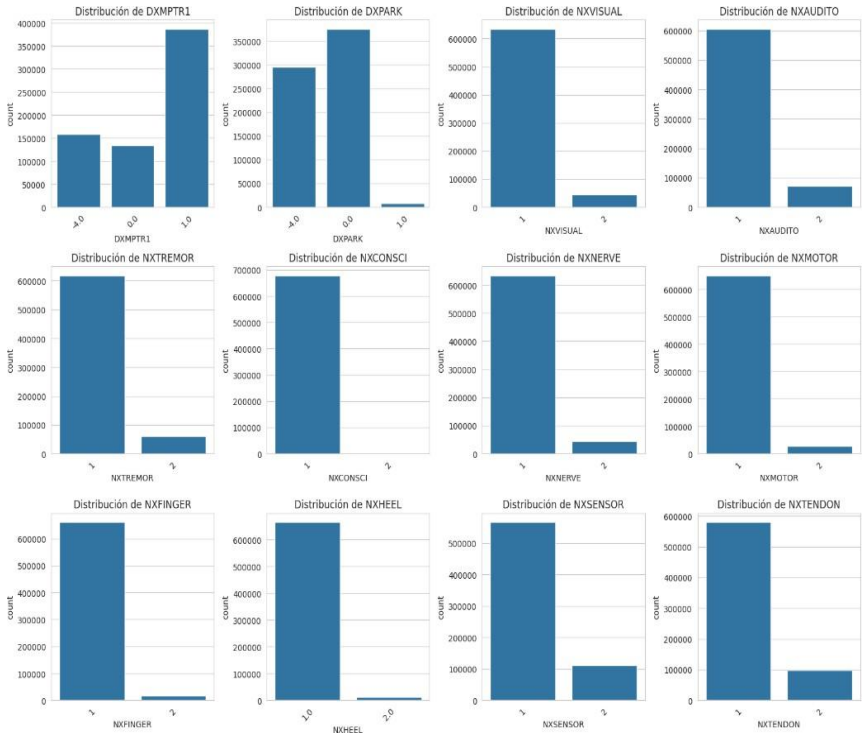

Ilustración 9: Desbalanceo de Clases 1 *Fuente: Elaboración propia*

Los gráficos de barras muestran un claro desbalanceo en las variables categóricas analizadas, tanto en las pruebas neurológicas como en las condiciones médicas y hábitos. En la *Ilustración 9: Desbalanceo de Clases 1*, la mayoría de los pacientes no presentan síntomas, posiblemente debido a la baja representación de pacientes con Alzheimer (13.7%). Por el otro lado, en las variables medicas las cuales están representadas en el anexo como *Ilustración 25: Desbalanceo de clases no neurológicas*, aunque el desbalanceo es menos pronunciado sigue presente.

5.4 Transformación de datos

A pesar de haber hecho un análisis estadístico y descriptivo, resulta fundamental abordar de forma específica la presencia de outliers, valores nulos y desbalanceo entre clases. Dado que trabajamos con un fichero con un gran volumen y en que el objetivo del proyecto es resolver un problema de clasificación, es crucial tratar correctamente estos aspectos. De lo contrario, los modelos podrían entrenarse de manera de forma sesgada, comprometiendo la veracidad del análisis y fiabilidad de los resultados.

5.5.1 Outliers

En el apartado anterior, analizamos la distribución de nuestras variables tanto de forma visual como descriptiva. Sin embargo, en algunas de estas graficas hemos podido identificar posibles outliers (puntos atípicos). Los outliers no debe pasar desapercibido, ya que nos pueden proporcionar información valiosa sobre la estructura de nuestros datos. Por lo tanto, es importante entenderlos y analizar su impacto en el análisis. En este caso, al solo tener dos variables con valores numéricos continuos: PLASMAPTAU181 y PLASMA_NFL, vamos a aplicar la detección de los outliers en estas. PTEDUCAT al no ser una variable binaria, esta también estará incluida en el estudio.

5.5.1.1 Método del Rango intercuartílico (IQR)

El IQR (Interquartile Range) define outliers como valores fuera de la siguiente formula:

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

Ecuación 13: IQR

Tabla 14: Outliers

Variable	Outliers	Porcentaje
PTEDUCAT	10910	1.42%
PLASMAPTAU181	27025	3.52%
PLASMA_NFL	28147	3.67%

Observando la *Tabla 14: Outliers* los biomarcadores presentan el 3,52% y 3,62% outliers, respectivamente. En nuestro contexto, estos valores atípicos no necesariamente

implican errores en los datos, sino que pueden reflejar la heterogeneidad propia de la progresión del Alzheimer. Debemos recordar que según Tabla 17: Variables y desviación estándar, ambos biomarcadores mostraban una dispersión muy grande en sus valores. Esto puede corresponder a pacientes en etapas avanzadas de la enfermedad, donde la neurodegeneración es más severa o casos puntuales que son interesantes para tener en cuenta.

5.5.2 Valores Nulos

Estamos ante un dataset bastante delicado, donde es fundamental analizar la presencia de los valores nulos, ya que un manejo incorrecto puede distorsionar los resultados y afectar a la validez del estudio. Antes de aplicar cualquier técnica de imputación, debemos evaluar la cantidad y distribución de los valores faltantes para determinar si es más conveniente eliminarlos, imputarlos con medidas estadísticas o utilizar métodos más avanzados.

Tabla 15: Valores Nulos

	Variable	**Valores**	**Porcentaje**
1	DXPTR1	234414	30.57%
2	PTGENDER	77652	10.12%
3	PTDOB	87810	11.45%
4	PTEDUCAT	77676	10.12

Según la naturaleza de nuestros datos, no se va a aplicar una interpolación en el conjunto de datos. Principalmente, debido a que los datos médicos no siguen patrones lineales predecibles. Para ello, he optado por diferentes técnicas.

Según la *Tabla 15: Valores Nulos*, DXPTR1 es la variable con el mayor porcentaje de valores nulos (30.57%). Esta representa un indicador relacionado con la memoria del paciente. Dado que existe una relación clínica entre el diagnóstico y la queja subjetiva de la memoria, he optado por una imputación condicional basada en el DIAGNOSIS de cada paciente. En el caso de los pacientes con DIAGNOSIS = 1 (Cognitivamente Normal), los valores nulos se reemplazaron por 0, asumiendo que estos pacientes no presentan problemas

de memoria. Para aquellos con DIAGNOSIS = 2 (Problemas Cognitivos Leves) y DIAGNOSIS = 3 (Demencia o Alzheimer), los valores nulos fueron imputados con 1, ya que en estos casos es más probable que exista algún nivel de deterioro cognitivo que afecte la memoria.

El problema que estamos teniendo con los valores nulos en las tres variables restantes es que estos provienen de filas que previamente han sido completadas. Por lo tanto, tenemos una fila exactamente igual que la previa más tres celdas (PTGENDER, PTDOB, PTEDUCAT) que no tienen ningún valor. Les pongo un ejemplo extraído del fichero "FicheroJunto".

Tabla 16: Valores Nulos PTEDUCAT, PTGENDER, PTDOB

RID	**……..**	**PLASMAPTAU181**	**VISDATE**	**PTGENDER**	**PTDOB**	**PTEDUCAT**
21		9674	28/09/2005	2	01/02/1933	18
21		9674	07/10/2010	2	01/02/1933	18
21		9674	07/10/2010	NaN	NaN	NaN
21		9674	10/10/2011	2	01/02/1933	18
21		3745	28/09/2005	2	01/02/1933	18
21		3745	07/10/2010	2	01/02/1933	18
21		3745	07/10/2010	NaN	NaN	NaN
21		3745	10/10/2011	2	01/02/1933	18
……		…….	……..	……	……..	…….
21		11188	10/10/2011	2	01/02/1933	18

En la *Tabla 27: PTEDUCAT, PTGENDER, PTDOB*, si se añade el sexo, fecha de nacimiento y años de educación al paciente en las filas donde hay valores nulos utilizando un forward fill se convertiría en una fila duplicada ya que son características personales que no van a cambiar durante el tiempo, por lo que dichas filas serian exactamente iguales a sus anteriores, de manera que no nos aportaría información relevante. Consecuentemente, para evitar la creación de registros duplicados y valores nulos, se procederá a eliminar las filas con valores nulos en esas tres variables ya que todos los nulos existentes de esas tres variables en nuestro conjunto de datos son del mismo caso que se muestra en la tabla.

5.5.3 Desbalanceo de clases

Como se mostró en la sección la sección de histogramas de variables categóricas, ciertos grupos diagnósticos estaban sobrerrepresentados, lo que podía sesgar el aprendizaje del modelo. Para solucionar este problema, decidí realizar un downsampling en lugar de aplicar técnicas de sobremuestreo como SMOTE. Esta decisión se tomó debido a que el dataset inicial contaba con 766,807 observaciones, y realizar un downsampling permitió reducir el desbalance sin consumir grandes cantidades de memoria. Como resultado, se han mantenido 1166 participantes únicos, lo cual es más que suficiente para entrenar el modelo. El dataset quedó reducido a 272,553 observaciones, preservando la diversidad y variabilidad de las clases.

Tabla 17: Observaciones Downsampling

Informe	Observaciones
FicheroJunto	766,807
FicheroJuntoPT	272,553

Posteriormente, se balanceo la clase objetivo (DIAGNOSIS), ajustando la clase mayoritaria para igualarla a la clase minoritaria, en este caso la clase 3 (AD). Para ello, se empleó el método “StratifiedShuffleSplit”, que asegura que la proporción de las clases se mantenga constante en los conjuntos de entrenamiento y prueba

Tabla 18: Distribución Diagnosis Downsampling

Diagnosis	Observaciones	Frecuencia
1	98,689	36,20%
2	90,851	33,33%
3	83,013	30,45%

No obstante, el downsampling no solo se aplicó a la variable objetivo, sino que también en las variables predictoras serian usadas como características de nuestros modelos.

Para ello, el criterio de downsampling no se aplicó de manera uniforme, sino que se ajustó en función del grado de desbalance presente en cada diagnóstico y variable categórica

relevante. En ciertos casos, como en la variable DXMPTR1 dentro del grupo de pacientes con DIAGNOSIS = 2, se redujo hasta un 40% de las observaciones para evitar que algunas categorías estuvieran sobrerrepresentadas. Sin embargo, en otros casos, donde la diferencia no era tan clara, se aplicaron reducciones más pequeñas, como del 10%. Para realizar este ajuste, se eligieron aleatoriamente las observaciones a modificar, asegurando que el muestreo mantuviera la coherencia en la distribución de los datos. Esto ayudo a corregir la disparidad en la representación de las categorías sin introducir sesgos artificiales ni afectar la estructura global del dataset.

A continuación, les muestro una visualización del antes y después del balanceo de la clase DXMPTR1 como ejemplo.

Tabla 19: Distribución DXMPTR1

DIAGNOSIS	**DXMPTR1**	**FRECUENCIA**
1	-4.0	53.17%
	0.0	46.82%
2	1.0	97.86%
	0.0	1.94%
3	1.0	81.45%
	-4.0	18.54%

Tabla 20: Distribución Post Downsampling

DIAGNOSIS	**DXMPTR1**	**FRECUENCIA**
1	-4.0	52,72%
	0.0	42,53%
	1.0	4,72%
2	-4.0	0.20%
	1.0	31,32%
	0.0	68,48%
3	-4.0	18,29%
	1.0	81,71%

6. PLAN DE DESARROLLO DEL PROYECTO

6.1 Metodologia

A continuación, se detalla ordenadamente la metodología que se sigue para la realización del proyecto:

1. En primer lugar, se escogerán los modelos predictivos que cumplan nuestros objetivos planteados y que tengan la capacidad para clasificar con biomarcadores y variables clínicas. Para este proyecto hemos escogido los siguientes: Regresion Logistica Multinomial, Random Forest Multinomial y Red Neuronal Artificial.
2. A continuación, se determinará el entorno de trabajo. En este caso se ha decidido desarrollar en el lenguaje de programación Python, utilizando el entorno de programación Jupyter Notebook.
3. Una vez configurado el entorno, se llevará a cabo un preprocesamiento exhaustivo del dataset original, que incluye:
    a. Unificación y limpieza del dataset: Se hará uso de un Inner Joint para consolidar las tablas en una, y posteriormente trataremos valores nulos, outliers etc.
    b. Tratamiento del desbalanceo de clase: Dada la gran superioridad de la clase CN frente a MCI y AD, se realiza un downsampling para evitar sesgos durante el entrenamiento del modelo.
    c. Normalización de variables continuas y PTEDUCAT: Se normalizarán los biomarcadores y la clase PTEDUCAT para que puedan estar a la misma escala que las demás variables categóricas.
4. Desarrollo y entrenamiento de los modelos: Entrenamos los modelos previamente planteados en el primer punto ajustando cada uno a sus técnicas específicas para optimizar el rendimiento.

5. Tras haber hecho la predicción, hacemos una evaluación comparativa del rendimiento de los modelos utilizando las diferentes métricas y elegimos la mejor predicción
6. Finalmente, se analiza el impacto de los resultados obtenidos en el ámbito del diagnóstico temprano de Alzheimer y procedemos ofrecer recomendaciones para futuros investigadores.

Con la finalidad de presentar de forma más comprensible la metodología mostrada, se facilita un flujograma principal que se muestra en el anexo como *Ilustración 27: Flujograma metodología*.

6.2 Herramientas Seleccionadas

6.2.1 Excel

Se ha utilizado Excel para unificar los informes debido a su facilidad para combinar y visualizar grandes volúmenes de datos de forma natural y controlada para la posterior manipulación en Python.

6.2.2 Python

Para la creación y entrenamiento de todos los modelos predictivos, se ha utilizado el lenguaje de programación Python junto con diversas librerías:

- Pandas: Se ha utilizado para la manipulación y limpieza de datos, facilitando la gestión de grandes volúmenes de información con estructuras como Dataframe.
- Numpy: Nos ha permitido hacer todas las operaciones matemáticas y manipulaciones con vectores, matrices etc.
- Matplotlib: Se ha empleado para la creación de visualizaciones que ayudan a interpretar los datos y los resultados del modelo.
- Scikit-learn: Es la librería encargada para la creación y evaluación de los modelos de machine learning.

- Tensorflow.keras: Facilita la construcción y entrenamiento de redes neuronales artificiales (ANN) con un interfaz de alto nivel para deep learning

6.3 Normalización

Se aplico una normalización Z – Score a los biomarcadores debido a sus diferencias en las unidades como se vio en *Tabla 6: Unidades Biomarcadores* y a PTEDUCAT para escalar las características del rango [0,1], lo cual evita que las variables con rangos de valores mayores dominen el proceso de entrenamiento

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Ecuación 14: Normalización

Donde x es el valor original, $\min(x)$ es el valor mínimo de la variable y $\max(x)$ es el valor máximo. Esta transformación garantiza que todas las variables tengan un rango similar evitando posibles sesgos en la parte de modelización.

7. ANALISIS DEL DATO: MODELOS PREDICTIVOS

7.1 Introducción a los modelos elaborados

Para desarrollar los modelos predictivos para la clasificación del Alzheimer, se mantendrán todas las variables vistas en la *Tabla 4: N.º de variables, su tipo de datos, formato, periodo e intervalos* como nuestras principales variables de interés. Este estudio centrara su predicción en la clasificación de pacientes en tres etapas clínicas: cognitivamente normal (CN), deterioro cognitivo leve (MCI) y enfermedad de Alzheimer (AD). El conjunto de datos utilizado abarca varios años e incluye registros clínicos y biológicos de más de 1.000 individuos. Para evaluar y comparar los resultados generados, todos los modelos serán entrenados y validados utilizando una división de datos del 80% para entrenamiento y 20% para prueba.

7.2 Dimensión Temporal

Se decidió no incluir las fechas de las pruebas (VISDATE, EXAMDATE, etc.) en el modelado, ya que como hemos visto previamente, los datos temporales no siguen un patrón consistente. Algunos pacientes tienen más registros que otros en la misma prueba y diferente

día, lo que genera incoherencia. Por lo tanto, al no ajustarse a una serie temporal estructurada, solo se utilizarán las fechas para análisis descriptivo y no en los modelos predictivos.

7.3 Criterios de desempeño

Las métricas seleccionadas permiten medir tanto la precisión general del modelo como su capacidad para diferenciar entre clases.

Tabla 21: Métricas a utilizar

	Variable	**Valores**	**Porcentaje**
1	Accuracy	El accuracy mide el porcentaje de predicciones correctas sobre el total de observaciones. Es una métrica que nos da una perspectiva general sobre el rendimiento del modelo.	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ *Ecuación 15: Accuracy*
2	Precision	La precisión mide la proporción de las personas positivas correctas, es decir, el porcentaje de los pacientes que el modelo clasifica en una categoría determinada realmente pertenece a esa clase.	$Precison = \frac{TP}{TP + FP}$ *Ecuación 16: Accuracy*
3	Recall	el Recall mide la capacidad del modelo para identificar correctamente los casos positivos dentro de esa clase. Es decir, evalúa cuantos de los pacientes que realmente pertenecen a una clase especifica son detectados correctamente por el modelo.	$Recall = \frac{TP}{TP + FN}$ *Ecuación 17: Recall*
4	F1 - Score	Este es la media armónica entre la precisión y recall. Es especialmente útil en casos como el nuestro dónde tanto los falsos positivos como los falsos negativos tienen un impacto significativo. Busca un balance entre ambas métricas, evitando que el modelo que tenga un buen desempeño en una de las métricas, pero un bajo rendimiento en la otra se sobrevalore	$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ *Ecuación 18: F1-Score*

Donde TP Son los casos el modelo predice correctamente una clase positiva. En nuestro caso, por ejemplo, es cuando un paciente con Alzheimer es clasificado correctamente con Alzheimer. Por el otro lado, TN son los casos en los que el modelo precie correctamente una clase negativa. Estos van a ser los pacientes CN o MCI que fueron clasificados

correctamente como CN o MCI (es decir, cualquier clase que no sea AD y que tampoco haya sido clasificada como AD).

Por el otro lado, FP ocurre cuando el modelo predice una clase positiva incorrectamente. Un falso positivo sería un paciente sin Alzheimer (CN o MCI) que el modelo clasifica erróneamente como AD.

7.3.1 Matriz de confusión

La matriz de confusión es otra herramienta visual que nos muestra el número de predicciones correctas e incorrectas para cada clase, lo que facilita la identificación de patones de error, como falsos positivos y falsos negativos. Nos permite ver cuántos pacientes de cada categoría (CN, MCI y AD) han sido clasificados correctamente y cuantos han sido mal clasificados.

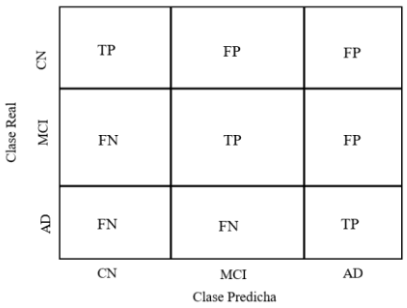

Ilustración 10: Matriz de Confusión
Fuente: Elaboración propia

7.3.2 ROC – AUC

Es una curva que muestra la relación entre la tasa de positivos verdaderos (recall) y la tasa de falsos negativos, ofreciéndonos visualizar cómo se comporta el modelo a medida que se ajustan sus umbrales de decisión

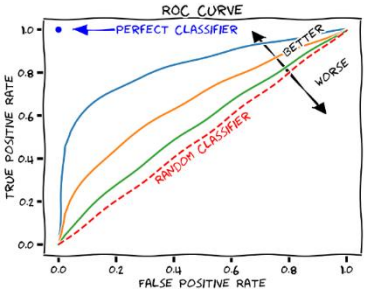

Ilustración 11: AUC - ROC *Fuente: Medium (2021)*

7.4 Analisis del dato: Modelos Predictivos

Por las características del proyecto, la variable que tomaremos como referencia será "DIAGNOSIS", la cual será nuestra variable dependiente $\hat{y}$. El resto de las variables, menos los identificadores y las fechas serán utilizadas como las variables independientes x (características).

7.4.1 Resultados (RL)

Para cada modelo, se ofrecerá una tabla con los hiperparametros utilizados y una justificación de su elección:

Tabla 22: Hiperparametros RL

Hiperparametros	Valor	Justificación
penalty	'L2'	Se aplico una regularización de tipo Ridge para minimizar la magnitud de los coeficientes, reduciendo la varianza y sobreajuste sin eliminar variables relevantes.
solver	'saga'	Es un optimizador basado en descenso estocástico. Es bastante adecuado para grandes volúmenes de datos, como en nuestro caso y permite la combinación de L2 con clasificación multinomial.
multi_class	'multinomial'	Dado que buscamos predecir directamente CN, MCI o AD sin recurrir a clasificadores binarios, es esencial usar este enfoque multinomial.
max_iter	1000	Se aumenta frente al valor por defecto (100) para asegurar convergencia en un espacio paramétrico amplio.

Una vez ya observado los hiperparametros que se han utilizado, a continuación, les proporciono los resultados obtenidos en el proceso de modelización:

Tabla 23: Accuracy RL

Conjunto de datos	Accuracy
Train	81,09%
Test	81,11%

Tabla 24: Informe de Clasificacion Train RL

Clase	Precision	Recall	F1 – Score
1	0,87	0,92	0,89
2	0,79	0,74	0,76
3	0,78	0,79	0,79

Los resultados del modelo en el conjunto de prueba son los siguientes:

Tabla 25: Informe de Clasificación Test RL

Clase	Precision	Recall	F1 – Score
1	0,87	0,92	0,89
2	0,79	0,74	0,76
3	0,78	0,79	0,79

Los resultados obtenidos para la Regresión Logistica Multinomial son sólidos como modelo bases, con un accuracy del 81,09% en entrenamiento y 81,11% en prueba, lo que indica una buena capacidad de generalización y ausencia de sobreajuste. En la clase 1 (CN – Cognitivamente Normal), el modelo alcanza un F1 – Score de 0,89, con un buen equilibrio entre la precisión y el recall.

Por el otro lado, para la clase (MCI – Deterioro Cognitivo Leve), la precisión y el recall bajan hasta un 79% y 74%, lo que significa que genera un 26% de falsos negativos. Por último, en la clase 3 (AD – Alzheimer), la precisión es ligeramente más baja con un 78% pero el recall es de un 79%, cifras similares a MCI. Esta cercanía en los valores entre clases más complejas (MCI y AD) puede deberse a la similitud de características entre ambas, lo que dificulta una separación clara por parte del modelo. Esto lo podemos ver de forma más visual en la matriz de confusión.

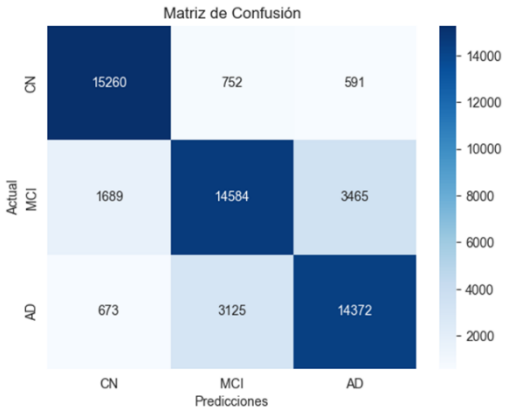

Ilustración 12: Matriz de Confusión RL
Fuente: Elaboración propia

La matriz de confusión *Ilustración 12: Matriz de Confusión RL* respalda el análisis hecho previamente, evidenciando que la mayor fuente de error en la clasificación del modelo se da entre los pacientes con MCI y AD. Específicamente, las 3,121 predicciones que el modelo realizo como MCI, los cuales realmente padecen Alzheimer, mientras que de las 3,466 predicciones como AD, muchas pertenecen a la categoría MCI. Esto nos confirma que la principal confusión ocurre entre estas dos clases, lo cual es lógico ya que comparten características clínicas y biomarcadores similares. Además, para respaldar el análisis de forma visual, he generado la curva AUC- ROC la cual se encuentra en el apartado de Anexos como *Ilustración 24: AUC – ROC RL*.

7.4.2 Clasificación de Coeficientes RL

Una de las principales características de la Regresion Logistica Multinomial como hemos hablado anteriormente es la interpretabilidad de sus coeficientes, lo que nos permite analizar el impacto de cada variable en la clasificación de los pacientes. A través de estos coeficientes, podemos determinar que característica tienen mayor influencia en la probabilidad de que un paciente pertenezca a una de las tres clases (CN, MCI o AD)

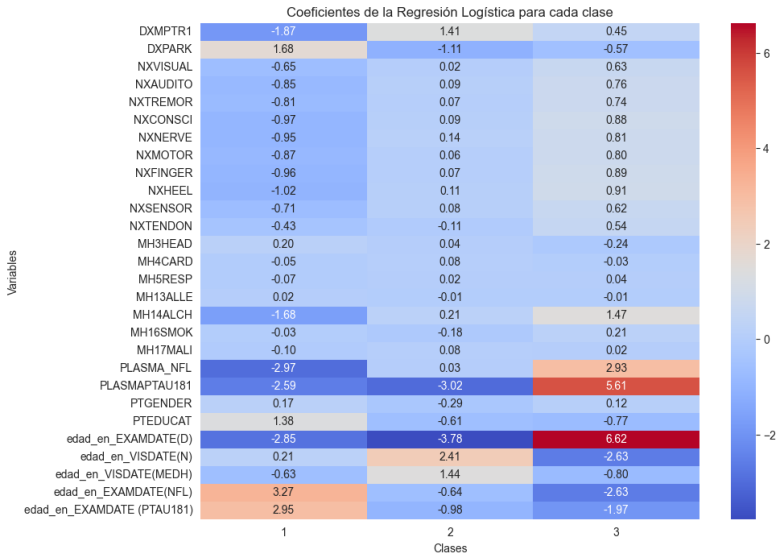

Ilustración 13: Coeficientes de la RL para cada clase *Fuente: Elaboración propia*

Como podemos ver en la *Ilustración 13: Coeficientes de la RL para cada clase*, la relevancia de ambos biomarcadores es notable, muestran coeficientes negativos en la clase CN con valores como -2.97 o -2.59 y muy positivos en la clase AD con valor más alto en el biomarcador P-tau181, lo cual es consistente con lo observado en análisis previo. La edad que tomamos como referencia es la “edad_en_EXAMDATE(D)” la cual tiene coeficientes negativos en CN y positivos en AD, reforzando la relación entre la edad avanzada y progresión de la enfermedad. Las variables neurológicas como NXMOTOR, NXFINGER y NXVISUAL reflejan una tendencia similar, su deterioro reduce la probabilidad de pertenecer a la clase CN y aumenta a la probabilidad de pertenecer en MCI o AD, lo cual es completamente coherente con el avance del deterioro cognitivo.

7.5 Random Forest

7.5.1 Hiperparametros

Tabla 26: Hiperparametros RF

Hiperparametros	Valor	Justificación
n_estimators	300	Se utilizaron 300 árboles para asegurar una estabilidad en el dataset ya que contamos con una alta variabilidad clínica.
max_depth	None	Al no limitar la profundidad, los árboles pueden crecer lo necesario para identificar combinaciones complejas entre los biomarcadores y los datos clínicos. El riesgo de sobreajuste decrece con otras restricciones como el número de arboles
min_samples_split	10	Este valor obliga a que cada nodo se divida solo si hay al menos 10 muestras, lo cual evita decisiones clínicas basadas en subgrupos muy pequeños que podrían no ser representativos.
min_samples_leaf	5	Se estableció un mínimo de 5 muestras por hoja para mejorar la generalización y evitar reglas de decisión extremas.
max_features	'sqrt'	Seleccionar $\sqrt{n}$ variables por split aumenta la diversidad entre árboles y mejora la detección de patrones clínicos relevantes evitando redundancias.
class_weight	'balance'	Dado al desbalance inicial en las clases, esta opción asegura que no se subestimen clases con proporción más baja y evitar sesgos.
validacion cruzada	k = 5	La validación cruzada permite al modelo evaluar con diferentes particiones del dataset, lo que mejora la generalización y reduce el sobreajuste.

7.5.2 Resultados (RF)

Tabla 27: Accuracy RF

Conjunto de datos	Accuracy
Train	87,86%
Test	87,37%

Tabla 28: Informe de Clasificación RF

Clase	Precision	Recall	F1 – Score
1	0,90	0,92	0,91
2	0,87	0,79	0,83

3	0,85	0,92	0,89

Tabla 29: Cross - Validation RF

Conjunto de datos	Accuracy
Cross – Validation	0.8759 ± 0.0015

Los resultados del Random Forest Multinomial muestran un accuracy del 87,86% en el conjunto de entrenamiento, mientras que en el de prueba un 87,37%, lo que a primeras muestra una buena generalización sin sobreajuste. Esto lo podemos respaldar con la *Ilustración 23:* . En la validación cruzada, los resultados son consistentes con los de la *Tabla 27: Accuracy RF* al evaluar su desempeño en múltiples particiones de los datos, esto sugiere que el modelo no está sobreajustado, es decir, que no ha aprendido patrones específicos el conjunto de entrenamiento que no se aplican a casos nuevos.

Con respecto a la *Tabla 28: Informe de Clasificación RF* la clase 1 (CN), con 90% de precisión y 92% recall, se logra un buen balance, identificando correctamente los pacientes sanos y minimizando los falsos positivos y negativos. En la clase 2 (MCI), la precisión aumenta una mejora significativamente con respecto al 79% de la RL, aunque el recall del 79% indica que aún hay un 21% de casos de MCI que el modelo no logra detectar correctamente. Por último, la clase 3 con un recall de 92% y la precisión de 85% evidencian un progreso significativo en la identificación de pacientes con Alzheimer, aunque todavía persiste una confusión con los pacientes MCI, donde un 15% de las predicciones de Alzheimer podrían corresponder erróneamente a MCI. Esto refleja nuevamente las dificultades entre las características de ambos diagnósticos, lo que subraya la importancia de seguir afinando el modelo para mejorar dicha diferenciación. Para poder ver más claro el desafío entre las predicciones que hace nuestro modelo he desarrollado otra matriz de confusión *Ilustración 14: Matriz de Confusion RF* y una curva AUC – ROC la cual se muestra en el apartado de Anexos como *Ilustración 26: AUC – ROC RF*.

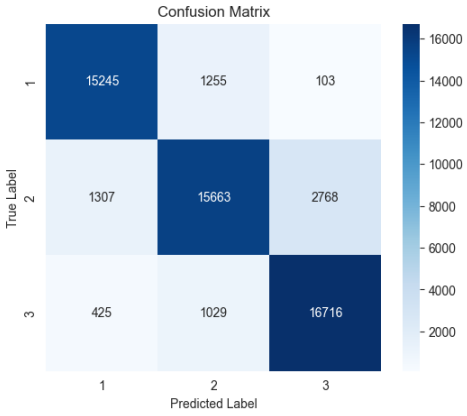

Ilustración 14: Matriz de Confusion RF
Fuente: Elaboración propia

Al igual que la Regresión Logística Multinomial, el Random Forest Multinomial tiene una característica única que nos aporta un valor especial a nuestro caso de estudio: permite identificar las características más relevantes utilizadas para clasificar a un paciente en una de las clases:

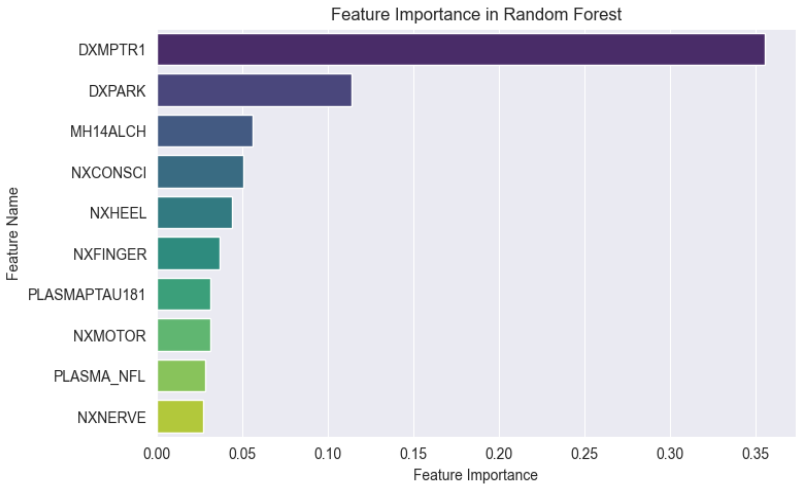

Ilustración 15: Importancia de las características RF Fuente: Elaboración propia

La queja subjetiva de memoria (DXMPTR1) se destaca significamente como un factor crucial. Para ponerles en contexto, la queja subjetiva se refiere a situaciones en las que los pacientes reportan dificultades como olvidar conversaciones recientes, perder objetos importantes con más frecuencia o luchar para recordar detalles que antes eran fácil de evocar. Estas quejas no son simples despistes cotidianos. Este indicador refleja el reconocimiento por parte de los pacientes de fallo en la memoria, lo que claramente es una señal temprana de deterioro cognitivo. Aunque es una medida subjetiva, su relevancia radica en que es la única "prueba" que recoge respuestas que no son evidentes a través de pruebas más formales. Este tipo de autoconciencia sobre la memoria a menudo precede el deterioro cognitivo más avanzado.

En cuanto a los biomarcadores, si es verdad que no están entre las tres primeras, pero son tanto el séptimo como noveno puesto lo cual lo evidencia como componentes esenciales del modelo al proporcionar evidencia biológica directa del daño neuronal. Respectivo al PLASMAPTUA181, su presencia en niveles elevados indica como hemos visto previamente procesos neurodegenerativos tempranos, lo que lo convierte en un marcador clave para

identificar casos en etapas iniciales. Por el otro lado, el biomarcador PLASMA_NFL tal vez tenga menos relevancia ya que el PLASMAPTAU181 se relaciona directamente con la acumulación de ovillos neurofibrales, que es una característica distintiva de la enfermedad mientras que el NfL genera un daño neuronal más generalizado. Este biomarcador complementa al Ptau181 al ofrecer una imagen más amplia del estado de la neurodegeneración.

Por último, otras variables como DXPARK, el consumo del alcohol (MH14ALCH) y las pruebas clínicas NXCONSCI y NXHEEL aportan información complementaria al modelo. La asociación entre Parkinson y Alzheimer destaca porque es rara vez vista que una persona con Alzheimer padezca de Parkinson como vimos en la tabla de coeficientes de la regresión logística, mientras que el consumo de alcohol pone en evidencia el impacto de los hábitos de la vida en el riesgo neurológico.

7.6 Red Neuronal Artificial

7.6.1 Hiperparametros

Se aplico una estructura *feed foward network* con los siguientes parámetros:

Tabla 30: Hiperparametros ANN

Hiperparametros	Valor	Justificación
Número de capas ocultas	3	Tres capas ocultas que permiten al modelo capturar complejidades no lineales en los datos clínicos y biomarcadores. Esto convertiría a nuestra ANN en un Deep Learning Model.
Número de neuronas por capa	64	Fue elegido para equilibrar la capacidad de aprendizaje del modelo y evitar sobreajuste.
Función de activación	ReLU	Se utiliza esta función de activación ya que no tiende a tener problemas de desvanecimiento de gradiente.
Tasa de aprendizaje	0.001	Una tasa de aprendizaje baja ayuda a evitar que el modelo salte a soluciones subóptimas, permitiendo un entrenamiento más preciso.
Optimización	Adam	Este optimizador ajusta la tasa de aprendizaje durante el entrenamiento, favoreciendo la detección de patrones en nuestros datos.
Función de perdida	Categorical Crossentropy	Al ser un modelo de clasificación multinomial, es la indicada y predeterminada para este tipo de problemas.

Épocas	50	Aseguramos un numero de épocas suficientes para que la red converja sin riesgo de sobreajuste.
Tamaño de Batch	32	Este tamaño permite que el modelo se entrene de manera eficiente y converja más rápidamente sin perder generalización. Además, es adecuado para datasets de tamaño medio – grande como en este caso.
Dropout	0.2	Parámetro de regularización que ayuda a evitar sobreajuste evitando la dependencia entre las neuronas durante el entrenamiento.

7.6.2 Resultados (ANN)

Tabla 31: Accuracy ANN

Conjunto de datos	Accuracy
Test	90,0%

Tabla 32: Informe de Clasificación ANN

Clase	Precision	Recall	F1 – Score
1	0,91	0,94	0,93
2	0,90	0,84	0,87
3	0,89	0,92	0,92

Los resultados muestran un rendimiento excelente, superior en comparación con los modelos previos, tanto la regresión logística como el random forest. La precisión del modelo en el conjunto de prueba alcanza un 90%, lo que supone una mejora respecto al 87,37% de RF y al 81,13% de RL.

Al analizar la *Tabla 32: Informe de Clasificación ANN*, se observa que la clase 1 alcanza una precisión del 91% y un recall del 94%, lo que indica que el modelo es altamente eficaz en la identificación de pacientes sin deterioro, minimizando los falsos negativos. Por el otro lado, la clase 2 representa un desafío mayor debido por su similitud con la clase 3 como previamente hemos visto. Sin embargo, muestra una mejora respecto al RF y RL, con precisión de 90%y recall de 84%. Aunque el recall sigue siendo más bajo que la precisión, la ANN logra una clasificación más equilibrada.

Por último, la clase 3 presenta una mejora en precisión respecto al RF (de 0,85 a 0,89) aunque el recall se mantiene constante, indicando que el modelo pasa por alto el 8% de los pacientes con Alzheimer, un número relativamente bajo. Además, se han reducido los falsos positivos a un 11%, lo que significa que el modelo comete errores en un 11% de los casos al diagnosticar incorrectamente a personas sin Alzheimer.

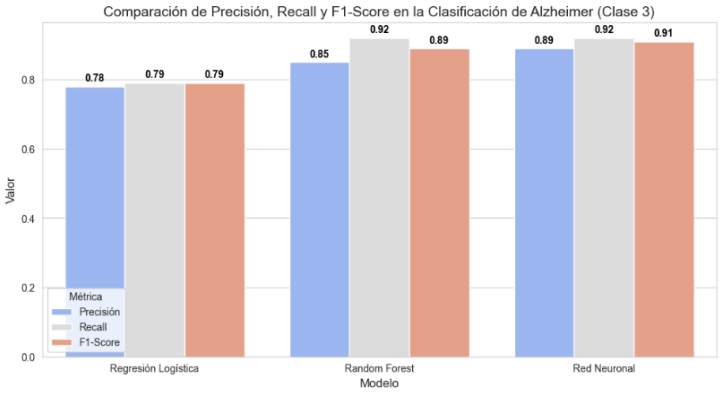

Ilustración 16: Comparativa entre modelos clase AD
Fuente: Elaboración propia

7.6.3 Curva de Error ANN

Por último, mostrara si existe o no un sobreajuste en nuestra red neuronal mostrando tanto la evolución de la perdida como de la precisión.

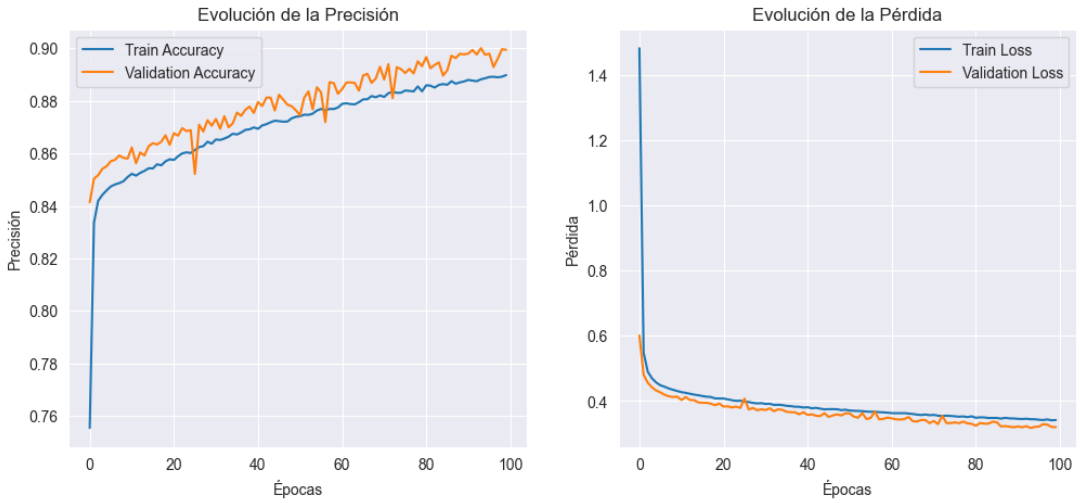

Ilustración 17: Evolución de Precision y de Perdida

Fuente: Elaboración propia

Observando la *Ilustración 17: Evolución de Precision y de Perdida* se observa que la precisión del entrenamiento aumenta de forma constante hasta estabilizarse en el rango de 0,88 y 0,89. La curva de validación sigue un patrón similar, con algunas fluctuaciones menores, pero siempre por encima de la curva de entrenamiento, alcanzando una predicción final más alta. En cuanto a la perdida, se reduce significativamente al inicio del

entrenamiento, estabilizándose en torno al 0,35. La pérdida del conjunto de validación también desciende, pero siempre manteniéndose cerca de la curva de entrenamiento.

Por lo tanto, no parece haber indicio de overfitting ya que las curvas de validación y entrenamiento siguen patrones similares, las pequeñas fluctuaciones son normales porque el modelo puede experimentar variabilidad en las métricas debido a la aleatoriedad presente de los datos, la inicialización de los pesos o el proceso de optimización, sin que implique un problema de generalización.

8 ANALISIS DE NEGOCIO

En esta sección se identificarán los principales aprendizajes y oportunidades que este modelo predictivo puede aportar desde una visión consultiva. El objetivo final es explorar como nuestra solución basada en datos puede generar valor real en nuestro día a día como por ejemplo en entornos clínicos o instituciones científicas, dejando de lado los aspectos puramente técnicos y centrarnos en su aplicabilidad en el ámbito del negocio.

Se podría desarrollar una herramienta web o aplicación integrada en los sistemas hospitalarios que, a partir de unos pocos datos clínicos y biomarcadores del paciente con los que se haya entrenado el modelo, proporcione la probabilidad de pertenecer a cada categoría diagnostica. Esto en ningún momento buscaría remplazar el diagnostico dado por el médico, pero si funcionaria como un sistema de alerta o un apoyo inteligente, ayudando a los profesionales sanitarios a priorizar casos de riesgo. Para que lo vean visualmente, les muestro un ejemplo de la interfaz de la aplicación en la *Ilustración 18: Interfaz.*

Ilustración 18: Interfaz

Fuente: Elaboración propia

8.2 Conclusiones y Recomendaciones

8.2.1 Conclusión

En el camino hacia una medicina más precisa y anticipativa, esta investigación ha abordado la predicción del riesgo de Alzheimer mediante la integración de datos clínicos y biomarcadores, utilizando el machine learning, una rama de la IA para analizar factores

como el desbalance de clases y el preprocesamiento de datos. Los resultados han destacado la importancia de variables como la queja subjetiva, el consumo de alcohol o ambos biomarcadores en la predicción temprana de la enfermedad. Además, se ha demostrado que la Red Neuronal Artificial (ANN) es el modelo más efectivo, superando a otros como la regresión logística y el random forest.

Desde una perspectiva clínica, la aplicabilidad de este modelo ofrece un gran potencial, tanto en el apoyo al diagnóstico temprano como en la segmentación de pacientes para ensayos clínicos. La creación de herramientas web o aplicaciones podría mejorar la detección en etapas tempranas y la priorización de caso, ofreciendo a los profesionales de la salud una herramienta complementaria que refuerce el diagnóstico médico. Por último, las farmacéuticas podrían beneficiarse de una segmentación más precisa de pacientes en ensayos clínicos, optimizando la selección y reduciendo los costos.

8.2.2 Recomendaciones

Tras reflexionar sobre todo el proceso y los resultados obtenidos, he decidido presentar una serie de recomendaciones y aprendizajes para futuros estudiantes o investigadores que quieran llevar a cabo y mejorar un proyecto como este.

Tabla 33: Recomendaciones

	Recomendación	**Descripción**
1	Consideración de la Dimensión Temporal y Datos Imágenes	Aunque en el trabajo se descartó la dimensión temporal por la inconsistencia en las fechas, futuros estudios podrían explorar su integración mediante el uso de imágenes cerebrales. Esto permitiría visualizar la evolución del deterioro cognitivo y aplicar modelos más complejos como RNN o CNN
2	Uso de Datos Multimodales	La combinación de datos clínicos y biomarcadores ha sido efectiva, pero en futuros estudios podrían integrarse datos genéticos, ofreciendo un perfil más complejo del paciente y aumentando la complejidad del análisis.
3	Planificación y Estandarización en la Integración de Informes	Un reto clave del proyecto fue integrar informes médicos con formato y fechas inconsistentes, lo que obligo a crear una clave primaria compleja. En futuros estudios, se aconseja definir desde el inicio una clave única estandarizada.
4	Estrategia de Manejo de Desbalanceo de Clases	Los datos clínicos tienden a tener un desbalanceo de clases notable, por lo tanto, según el contexto es recomendable hacer un estudio previo de la distribución de cada clase y aplicar downsampling o oversampling.

9 ANEXOS

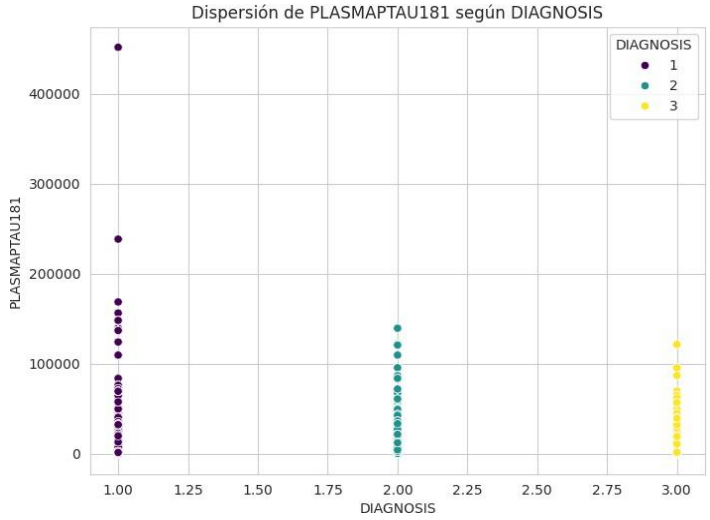

Ilustración 19: Dispersión de Ptau181 *Fuente: Elaboración propia*

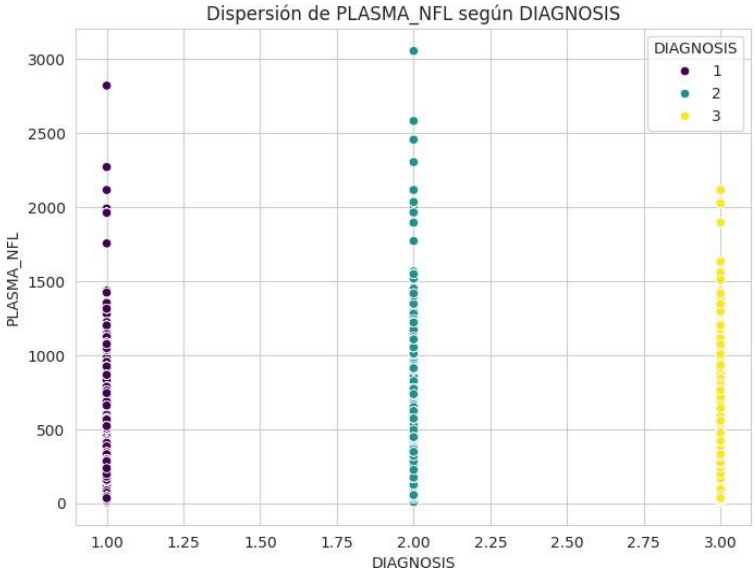

Ilustración 20: Dispersión de NfL *Fuente: Elaboración propia*

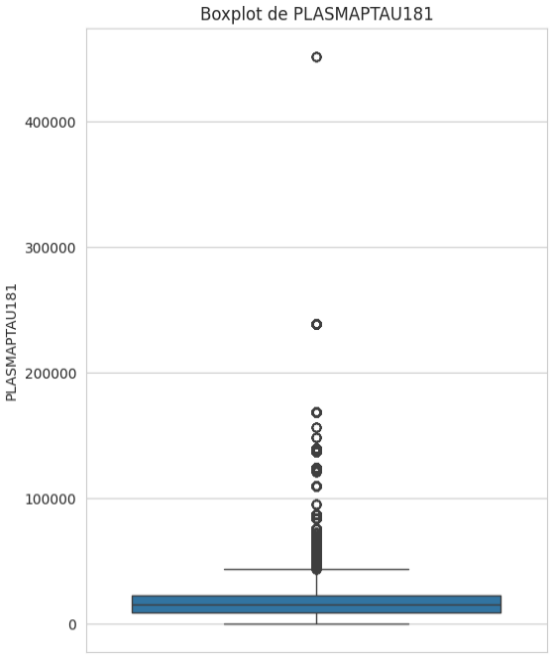

Ilustración 22: Boxplot Ptau181 *Fuente: Elaboración propia*

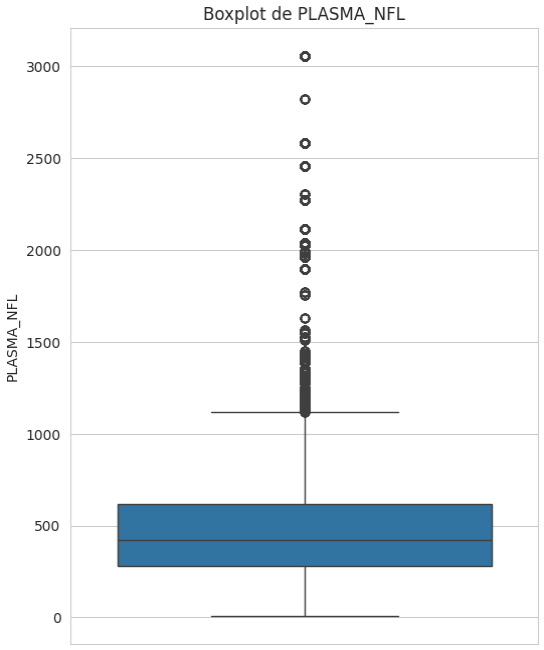

Ilustración 21: Boxplot NfL *Fuente: Elaboración propia*

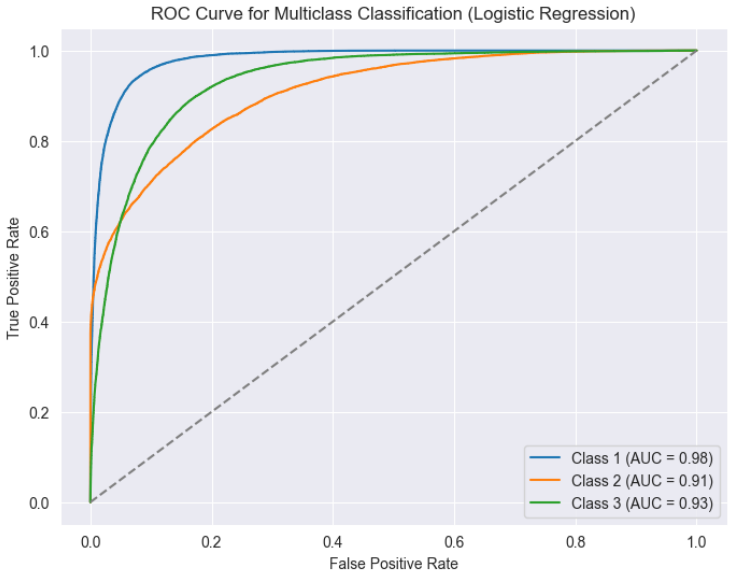

Ilustración 24: AUC – ROC RL *Fuente: Elaboración propia*

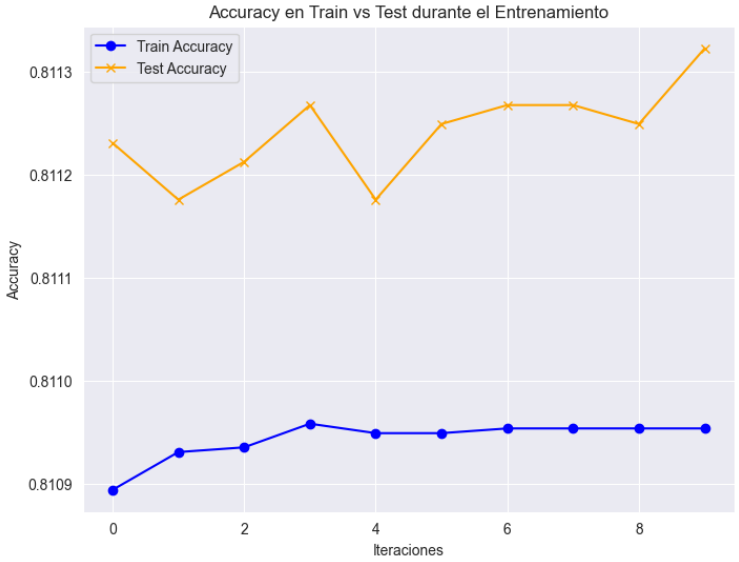

Ilustración 23: Train vs Test Entrenamiento RL *Fuente: Elaboración propia*

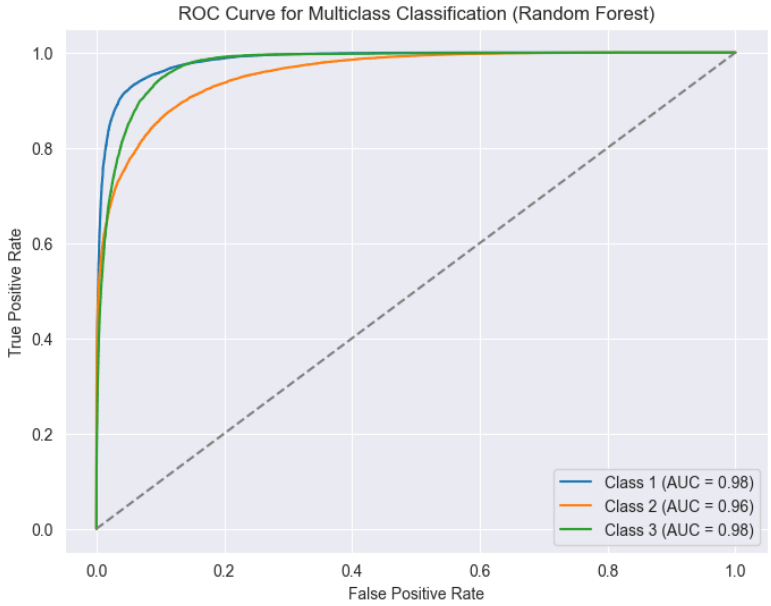

Ilustración 26: AUC – ROC RF *Fuente: Elaboración propia*

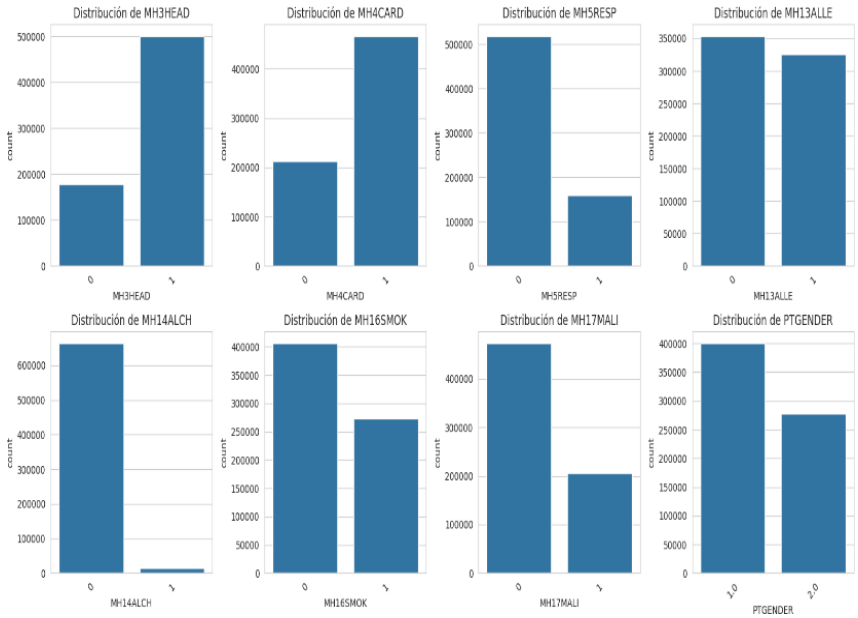

Ilustración 25: Desbalanceo de clases no neurológicas *Fuente: Elaboración propia*

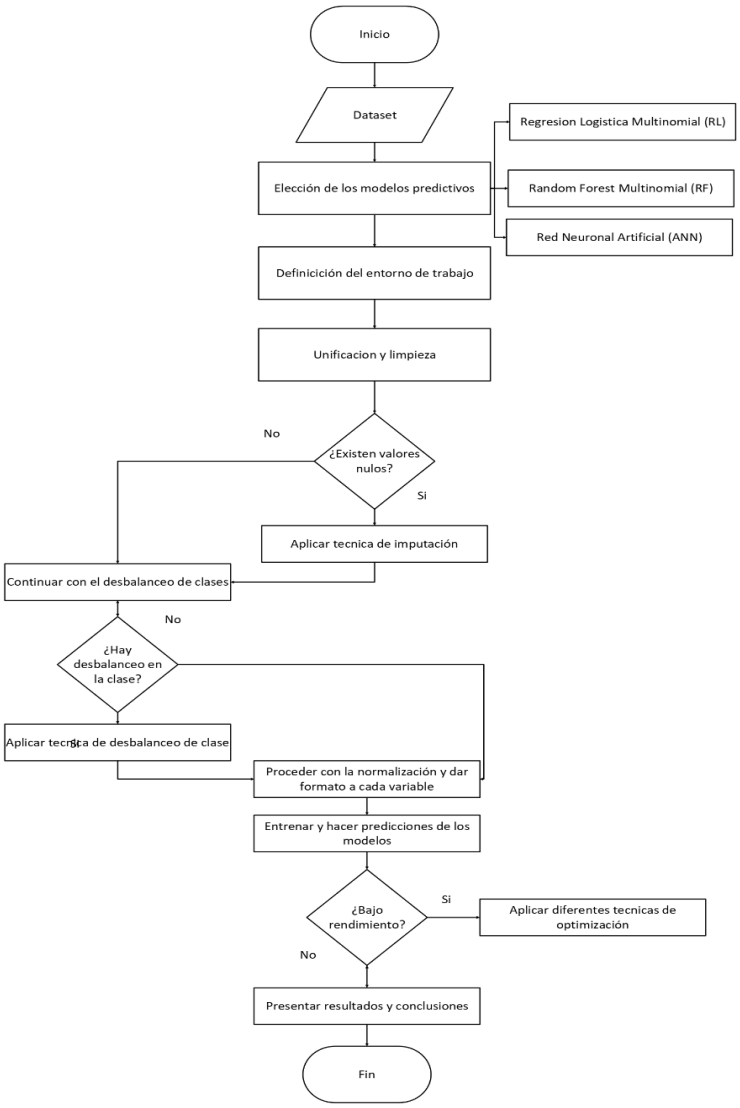

Ilustración 27: Flujograma metodología *Fuente: Elaboración propia*

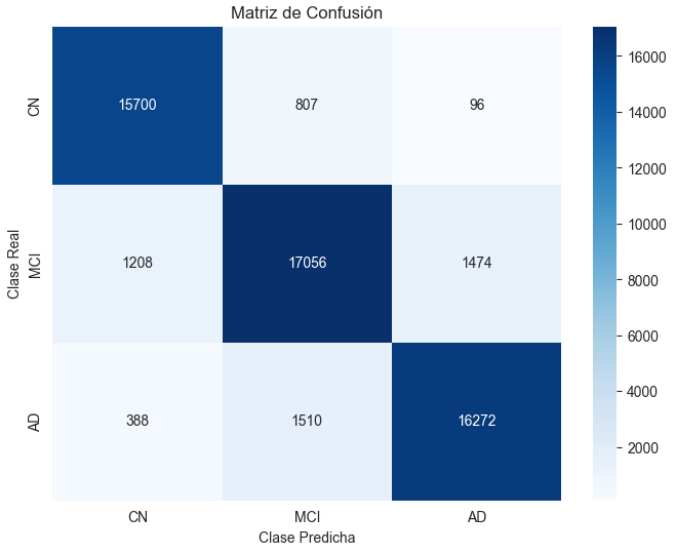

Ilustración 28: Matriz de Confusion ANN *Fuente: Elaboración propia*

10 BIBLIOGRAFIA

Aguilar, L. R. (2018, December 21). *Proteína ligera de los neurofilamentos, ¿cuál es su relación con la neurodegeneración?* TiTi. https://infotiti.com/neurofilamentos/

Alladi, S., Xuereb, J., Bak, T., Nestor, P., Knibb, J., Patterson, K., & Hodges, J. R. (2007). Focal cortical presentations of Alzheimer's disease. *Brain, 130*(10), 2636–2645. https://doi.org/10.1093/brain/awm213

Arranz, E. (2024, January 25). *El envejecimiento alcanza un nuevo máximo histórico en España, del 137,3%...* Fundación Adecco. https://fundacionadecco.org/notas-de-prensa/el-envejecimiento-alcanza-un-nuevo-maximo-historico-en-espana-del-1373...

Blennow, K., & Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: Current status and prospects for the future. *Journal of Internal Medicine, 284*(6), 643–663. https://doi.org/10.1111/joim.12816

Brage, M. (2020). *Análisis de datos categóricos: regresión logística y multinomial* [TFG, Universidad de La Laguna]. https://riull.ull.es/xmlui/bitstream/handle/915/20667/Analisis%20de%20datos%20categoricos%20regresion%20logistica%20y%20multinomial.pdf?sequence=1&isAllowed=y

Breiman, L. (2001). Random Forest. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cadena SER. (2024, September 18). *El Alzheimer comienza veinte o treinta años antes...* https://cadenaser.com/andalucia/2024/09/18/el-alzheimer-comienza-veinte-o-treinta-anos-antes-de-que-aparezcan-los-primeros-sintomas-de-perdida-de-memoria-radio-sevilla.com

Ezzati, A., Zammit, A. R., Harvey, D. J., Habeck, C., Hall, C. B., & Lipton, R. B. (2019). Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *Journal of Alzheimer's Disease, 71*(3), 1027–1036. https://doi.org/10.3233/jad-190262

Figure 7. A simple and deep learning network. (n.d.). ResearchGate. https://www.researchgate.net/figure/A-simple-and-deep-learning-network_fig4_368274660

Jack, C. R., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Jo, T., Nho, K., & Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in Aging Neuroscience, 11*. https://doi.org/10.3389/fnagi.2019.00220

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Mattsson, N., Cullen, N. C., Andreasson, U., Zetterberg, H., & Blennow, K. (2019). Association between longitudinal plasma neurofilament light and neurodegeneration in patients with Alzheimer disease. *JAMA Neurology, 76*(7), 791. https://doi.org/10.1001/jamaneurol.2019.0765

Park, J. H., et al. (2020). Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *npj Digital Medicine, 3*(1). https://doi.org/10.1038/s41746-020-0256-0

Ozechi, S. (2025, March 5). *In-Depth Overview of Linear Regression Modelling*. Towards Data Science. https://towardsdatascience.com/in-depth-overview-of-linear-regression-modelling-a46ac4eb942a/

ReportLinker. (2023, April 4). *Biomarkers Global Market Report 2023*. GlobeNewswire. https://www.globenewswire.com/news-release/2023/04/04/2640841/0/en/Biomarkers-Global-Market-Report-2023.html

Ventura, C., & Olivares-Castiñeira, I. (2021, December 24). *Análisis de los factores de riesgo de la enfermedad del Alzheimer y su detección temprana mediante machine learning*. https://openaccess.uoc.edu/handle/10609/138908

Visor de libros. (n.d.). https://www.educa2.madrid.org/web/argos/la-maquina-del-tiempo/-/book/la-enfermedad-del-alzheimer?_book_viewer_WAR_cms_tools_chapterIndex=314d8095-c65f-4c96-927e-4d03f05f8321

Wang, N., Chen, J., Xiao, H., Wu, L., Jiang, H., & Zhou, Y. (2019). Application of artificial neural network model in diagnosis of Alzheimer's disease. *BMC Neurology*, *19*(1). https://doi.org/10.1186/s12883-019-1377-4

World Health Organization. (2025, March 31). *Demencia*. https://www.who.int/es/news-room/fact-sheets/detail/dementia.com