Universidad Zaragoza

1542

Trabajo Fin de Grado

Modelos de supervivencia e inteligencia artificial explicable en la predicción del riesgo de conversión de deterioro cognitivo leve a Alzhéimer

Survival models and explainable artificial intelligence in the prediction of the risk of conversion from mild cognitive impairment to Alzheimer's disease

Autor/es

Diego Domingo Ralla

Director/es

Mónica Hernández Giménez

Elvira Mayordomo Cámara

Graduado en Ingeniería Informática

Especialidad en Computación

Escuela de Ingeniería y Arquitectura

2024

Resumen

La enfermedad de Alzhéimer (AD) es la forma más común de demencia a nivel mundial. Se trata de una afección crónica para la que no existen tratamientos curativos. Su diagnóstico es un proceso complejo que consiste en la evaluación de un especialista médico, apoyado en diversas pruebas que pueden incluir escáneres de imagen cerebral, evaluaciones neuropsicológicas o la identificación de biomarcadores genéticos y en el líquido cefalorraquídeo. El deterioro cognitivo leve (MCI) es una etapa temprana del declive de las facultades cognitivas del individuo que, en muchos casos, puede progresar a AD. De esta manera, la detección temprana y el seguimiento de los pacientes con MCI son aspectos cruciales para la identificación de pacientes con un alto riesgo de conversión.

Recientemente, métodos de análisis de supervivencia han mostrado buenas prestaciones en la predicción de conversión de MCI a AD. Sin embargo, este tipo de modelos tiene una naturaleza de caja negra, es decir, su funcionamiento interno es poco transparente, lo que dificulta la interpretación de sus resultados. Para resolver este problema surgen métodos de inteligencia artifical explicable (XAI) como SHapley Additive exPlanations (SHAP), que permite descomponer las predicciones del modelo en la contribución de cada feature, proporcionando una explicación clara y cuantitativa del impacto de cada variable en el resultado del modelo.

En este trabajo se aporta una visión general completa de la eficacia de Random Survival Forests (RSF) y Gradient Boosting Survival Analysis (GBSA) en el problema de la predicción de conversión de MCI a AD en un periodo de tiempo de hasta 5 años. Se han utilizado datos de pacientes con MCI del Alzheimer's Disease Neuroimaging Initiative (ADNI) en diferentes puntos temporales de su seguimiento (baseline, mes 12, mes 24, y una concatenación de baseline y mes 12). Asimismo, se ha elaborado un método simple pero efectivo para la obtención de métricas complementarias similares a las típicamente utilizadas en los modelos de aprendizaje automático tradicionales, y se ha utilizado SHAP para abordar la explicabilidad de los modelos y analizar los factores que más influyen sobre las predicciones realizadas.

En términos generales, los resultados muestran que RSF obtiene mejores prestaciones que GBSA, alcanzando un c-index máximo de 0.875 con datos del mes 12. Entre las features con más impacto sobre las predicciones se han identificado tests de evaluación neuropsicológica como mPACCtrailsB o LDELTOTAL, la escala clínica FAQ, el volumen del giro temporal medio (MidTemp) o FDG, un biomarcador del metabolismo de glucosa en el cerebro. De esta manera, se observa que las predicciones realizadas son el resultado de la contribución acumulada de features de diversos tipos que miden diferentes aspectos del deterioro cognitivo.

En conclusión, en este trabajo se comprueba que la utilización de modelos de supevivencia apoyados por métodos de explicabilidad como SHAP puede resultar de gran utilidad en el ámbito clínico para la predicción del riesgo de conversión de MCI a AD y en la elaboración de tratamientos adaptados a las necesidades de cada paciente.

Índice

Glosario

AD	Enfermedad de Alzhéimer
ADAS	Alzheimer's Disease Assessment Scale
ADNI	Alzheimer's Disease Neuroimaging Initiative
CDRSB	Clinical Dementia Rating Sum of Boxes
CN	Cognitivo normal
CPH	Modelo de riesgos proporcionales de Cox
CSF	Líquido cefalorraquídeo
FAQ	Functional Activities Questionnaire
FDG	Fluorodesoxiglucosa
GBSA	Gradient Boosting Survival Analysis
HCI	Índice de convergencia hipometabólica
ICV	Volumen intracraneal
MCI	Deterioro cognitivo leve
MMSE	Mini-Mental State Examination
mPACC	Modified Preclinical Alzheimer's Cognitive Composite
MRI	Imagen por resonancia magnética
NMDA	Ácido N-metil-D-aspartato
OMS	Organización Mundial de la Salud
PET	Tomografía por emisión de positrones
pMCI	MCI progresivo
RAVLT	Rey Auditory Verbal Learning Test
RF	Random Forests
RSF	Random Survival Forests
SHAP	SHapley Additive exPlanations
sMCI	MCI estable
XAI	Inteligencia artificial explicable

1. Introducción

1.1. Contexto

Según la Organización Mundial de la Salud (OMS) [1], más de 55 millones de personas en todo el mundo padecen demencia. Este término engloba varias enfermedades que afectan a las capacidades cognitivas, y se caracteriza por un empeoramiento progresivo. La enfermedad de Alzhéimer (Alzheimer's disease, AD) es la forma más común de demencia, representando entre el 60 y el 70 % de los casos. Actualmente no existe un tratamiento para curar el AD. Sin embargo, algunos fármacos pueden ayudar a controlar sus síntomas. Por ejemplo, se ha demostrado que los inhibidores de la colinesterasa como el donepezilo, y los antagonistas de los receptores NMDA como la memantina para casos más severos, tienen cierta efectividad en la mejora de los síntomas y la calidad de vida de algunos pacientes [2].

El deterioro cognitivo leve (mild cognitive impairment, MCI) es una condición que se sitúa entre el envejecimiento cognitivo normal y el AD. Se trata de una condición muy heterogénea, donde los síntomas se caracterizan por su subjetividad, y se estima que los pacientes con MCI tienen un riesgo del 33.6 % de desarrollar demencia [3]. Por ello, es de vital importancia desarrollar métodos de predicción temprana tanto para la conversión de MCI a AD como para la propia enfermedad de Alzhéimer.

El diagnóstico de AD es un proceso complejo que generalmente comienza con una evaluación clínica realizada por un médico especialista, quien revisa los antecedentes médicos del paciente y realiza diferentes pruebas cognitivas, como el Mini-Mental State Examination (MMSE) o la Alzheimer's Disease Assessment Scale (ADAS). También se llevan a cabo pruebas de imagen cerebral, como resonancias magnéticas (magnetic resonance imaging, MRI) y tomografías por emisión de positrones (positron emission tomography, PET), que permiten detectar cambios estructurales en el cerebro asociados con el AD. Además, existen biomarcadores específicos en el líquido cefalorraquídeo (cerebrospinal fluid, CSF) y genéticos también asociados con el AD, como el alelo APOE-ϵ4, que se ha relacionado con un 20-25 % de los casos de Alzhéimer [4].

No obstante, todos estos métodos y biomarcadores solo permiten obtener un diagnóstico probable, basado en los síntomas y en la interpretación subjetiva de los resultados por parte del especialista médico. Además, la complejidad del diagnóstico de AD aumenta en etapas tempranas de la enfermedad, pues los síntomas pueden no coincidir completamente con los de un envejecimiento cognitivo normal ni con los de un inicio de demencia. De hecho, el único método de diagnóstico completamente fiable es una biopsia *post-mortem*, y gracias a este método se ha comprobado que hasta un 20 % de los diagnósticos realizados en vida eran erróneos [5].

Los métodos de aprendizaje automático han alcanzado altos niveles de precisión en el diagnóstico temprano de la enfermedad de Alzhéimer [6], así como en la predicción de la conversión de MCI a AD [7, 8]. Hasta el momento, los algoritmos de aprendizaje automático utilizados típicamente

para estos objetivos son métodos de aprendizaje supervisado, principalmente clasificadores binarios entrenados para distinguir entre pacientes con MCI estable y pacientes con MCI cuyo estado evolucionará hacia la demencia.

Sin embargo, estos algoritmos no son eficaces ante la censura de datos, es decir, cuando hay información incompleta o ausente sobre ciertos eventos en el conjunto de datos. En este caso, el evento sería la conversión de MCI a AD, y la censura de datos puede ocurrir cuando algunos pacientes fallecen o abandonan el estudio antes de que esta conversión tenga lugar, algo muy frecuente en estudios clínicos de este tipo. Para abordar estas limitaciones, se ha explorado la eficacia de modelos basados en el análisis de supervivencia, una técnica estadística que pretende calcular el tiempo transcurrido hasta que ocurre un evento de interés en una población, teniendo en cuenta para ello la censura de datos.

Algunos de los métodos más utilizados en este contexto son el modelo de riesgos proporcionales de Cox (Cox Proportional Hazards, CPH) y los Random Forests (RF). Aunque en otros estudios se han logrado buenos resultados con enfoques basados en CPH y RF, es importante destacar que estos modelos tienen una naturaleza de caja negra, lo que dificulta la explicación e interpretación de sus predicciones.

Para intentar abrir estas cajas negras e incrementar la comprensión de las decisiones tomadas por los modelos, surge el campo de la inteligencia artificial explicable (eXplainable Artificial Intelligence, XAI). Un método notable en este campo es SHAP (SHapley Additive exPlanations), que permite calcular la contribución de cada feature a las predicciones del modelo tanto de manera global como local (en el caso de un paciente concreto). Se puede utilizar SHAP mediante el paquete *shap* de Python (https://github.com/shap/shap), compatible con cualquier modelo de aprendizaje automático.

Además, cabe destacar que, dada la naturaleza del análisis de supervivencia, estos métodos no se evalúan mediante las métricas típicas de los sistemas de aprendizaje automático tradicionales. Los modelos de supervivencia calculan la probabilidad de que ocurra un evento concreto en un tiempo determinado, lo que permite obtener curvas de supervivencia que muestran dichas probabilidades en función del tiempo. Por ello, la comparación del rendimiento de modelos de supervivencia con el de modelos tradicionales de aprendizaje supervisado resulta compleja, ya que se trata de enfoques conceptualmente muy diferentes y con resultados que se interpretan de manera distinta.

1.2. Objetivos

Los objetivos de este Trabajo de Fin de Grado (TFG) son:

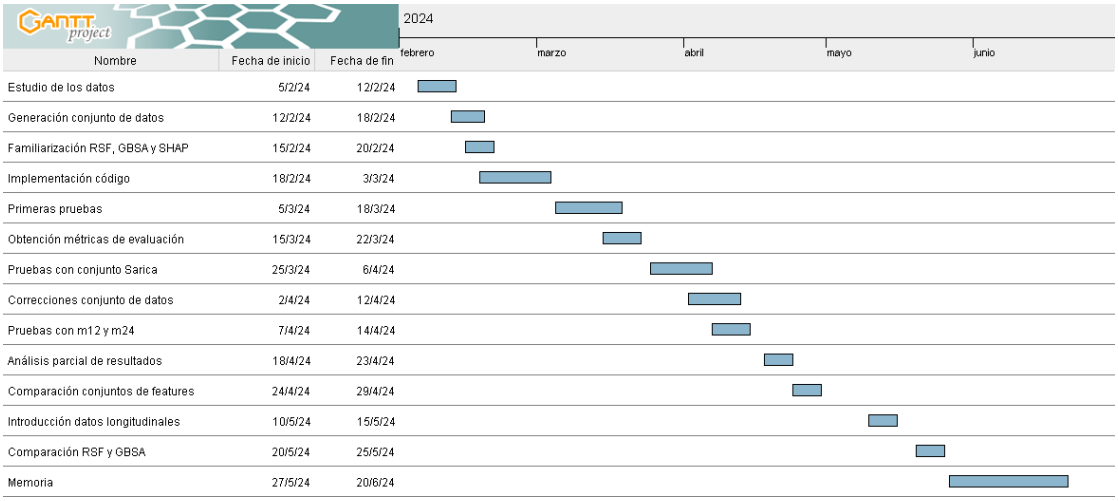
1. Realizar un análisis del estado del arte de modelos de supervivencia con aplicación al diagnóstico y predicción de Alzhéimer y de la evolución de MCI a AD en diferentes periodos de tiempo.

2. Desarrollar sistemas de generación de ficheros de entrada con información longitudinal.
3. Estudiar métodos para obtener métricas equivalentes a las utilizadas en los sistemas tradicionales de aprendizaje automático para una mejor caracterización de las prestaciones de los modelos.
4. Realizar una interpretación de los resultados de explicabilidad utilizando SHAP.
5. Relacionar los resultados en contexto con el conocimiento clínico.

En este trabajo se analiza y compara el rendimiento de Random Survival Forests (RSF) y Gradient Boosting Survival Analysis (GBSA), dos modelos de supervivencia basados en Random Forests y Gradient Booster, respectivamente. Por otra parte, se propone un método sencillo pero efectivo para calcular las métricas propias de los modelos tradicionales de aprendizaje automático, concretamente accuracy, sensitivity, specificity y f1-score, además del c-index propio de los modelos de supervivencia. Por último, se muestran los resultados de aplicar el método SHAP para observar las explicaciones globales y locales de ambos métodos y analizar la coherencia de estas explicaciones en comparación con el conocimiento clínico existente.

Este trabajo ha sido inspirado por el estudio realizado por Alessia Sarica et al., "Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer's disease", Brain Informatics, 2023 [9], en el que se analiza la capacidad de RSF en comparación con CPH para la predicción del evento de conversión de MCI a AD, así como el uso de varios métodos de explicabilidad para una mejor interpretación de los resultados.

El desarrollo del trabajo se ha realizado entre febrero y junio de 2024, con una primera aproximación al problema en enero. En la Figura 1 se presenta un diagrama de Gantt en el que se contabiliza el tiempo dedicado a cada tarea.



Nombre	Fecha de inicio	Fecha de fin
Estudio de los datos	5/2/24	12/2/24
Generación conjunto de datos	12/2/24	18/2/24
Familiarización RSF, GBSA y SHAP	15/2/24	20/2/24
Implementación código	18/2/24	3/3/24
Primeras pruebas	5/3/24	18/3/24
Obtención métricas de evaluación	15/3/24	22/3/24
Pruebas con conjunto Sarica	25/3/24	6/4/24
Correcciones conjunto de datos	2/4/24	12/4/24
Pruebas con m12 y m24	7/4/24	14/4/24
Análisis parcial de resultados	18/4/24	23/4/24
Comparación conjuntos de features	24/4/24	29/4/24
Introducción datos longitudinales	10/5/24	15/5/24
Comparación RSF y GBSA	20/5/24	25/5/24
Memoria	27/5/24	20/6/24

Figura 1: Diagrama de Gantt del proyecto

2. Métodos

2.1. Preprocesado de los datos

Todos los datos utilizados para el entrenamiento de los métodos implementados en este trabajo se han obtenido de la base de datos de Alzheimer's Disease Neuroimaging Initiative (ADNI), que incluye información de más de 2000 pacientes a lo largo de varias visitas periódicas. ADNI es un proyecto de investigación colaborativo iniciado en 2004 con el objetivo de analizar la progresión de la enfermedad de Alzhéimer y mejorar los métodos de diagnóstico y tratamiento. Más concretamente, se ha trabajado con el conjunto de datos *ADNIMERGE.csv*, descargado el 28 de septiembre de 2023, que contiene información demográfica, clínica, cognitiva y datos de pruebas MRI y PET.

Para realizar un primer procesado de los datos se ha utilizado la herramienta KNIME v5.2.1 (https://www.knime.com/downloads). A continuación se enumeran las transformaciones y filtros aplicados al conjunto de datos inicial:

- Se consideran únicamente los pacientes diagnosticados con MCI en baseline.
- Eliminación de visitas sin un diagnóstico.
- Eliminación de visitas con valores faltantes (*missing values*) en todas las features anatómicas, es decir, relacionadas con el volumen de estructuras cerebrales.
- Normalización de las features anatómicas dividiendo por el volumen intracraneal (ICV).
- Eliminación de visitas con valores indefinidos en las features ABETA, TAU y PTAU.
- Eliminación de pacientes *reversers*, es decir, pacientes que, a lo largo de las distintas visitas realizadas, revierten su diagnóstico, bien sea de AD a MCI o de MCI a CN (cognitivo normal).
- Eliminación de pacientes con menos de 3 visitas anuales en los 5 primeros años.

Cabe destacar que se han elaborado varios conjuntos de datos distintos. El conjunto bl considera solo las visitas de baseline, mientras que los conjuntos m12 y m24 toman como visita de referencia las visitas en los meses 12 y 24, respectivamente. De esta manera se busca adelantar dicha referencia y acercarla al evento de interés. Finalmente, el conjunto bl+m12 emplea datos longitudinales, considerando al mismo tiempo tanto las visitas de baseline como la visita en el mes 12.

Para la construcción de las etiquetas para el entrenamiento de los modelos se ha considerado cada paciente como una tupla compuesta por dos elementos: un booleano que indica si el paciente evoluciona o no a AD en los 5 primeros años, y un número entero que indica el tiempo transcurrido en meses desde la visita de referencia hasta que se observa dicho evento. En caso de no haberse observado, corresponde al número de meses que consta que el paciente se ha mantenido estable,

ya que el hecho de no haber observado evolución a AD puede deberse a censura de datos o simplemente a que el paciente no desarrolla la enfermedad en ese tiempo. Si el paciente no evoluciona a AD se le considera estable o sMCI (*stable* MCI), mientras que si desarrolla demencia se le considera progresivo o pMCI (*progressive* MCI).

2.2. Imputación

Uno de los problemas de las bases de datos como la de ADNI, construidas mediante una serie de visitas clínicas de un conjunto amplio de pacientes, es que hay una gran presencia de datos faltantes. Sin embargo, el análisis de supervivencia no se puede realizar en pacientes a los que les faltan datos, además de que los resultados no serían representativos.

Debido a la gran cantidad de restricciones y filtros aplicados sobre el conjunto de datos inicial, eliminar todos los pacientes a los que les falte algún valor no es factible, pues el tamaño final de la muestra sería bastante reducido y no lo suficientemente representativo para garantizar unos resultados coherentes. Por tanto, se ha decidido optar por la imputación de los datos faltantes. Para ello se ha utilizado el algoritmo MissForest [10] del paquete de Python *missingpy* v0.2.0. (https://github.com/epsilon-machine/missingpy). Cabe destacar que esta librería requiere de una versión de *scikit-learn* no superior a la 0.20.1.

Con el objetivo de evitar contaminar el modelo con los datos de test (*data leakage*), se ha separado inicialmente el conjunto de test mediante k-fold, y se ha realizado la imputación del conjunto de entrenamiento y el de test por separado en cada uno de los folds.

2.3. Modelos de supervivencia

El análisis de supervivencia es una técnica estadística cuyo objetivo es analizar el tiempo que transcurre hasta que se observa un evento de interés. En el contexto de este trabajo, dicho evento es la progresión de un paciente con MCI a AD. Sin embargo, no todos los pacientes experimentan el evento en cuestión durante el periodo de tiempo observado. Esto puede suceder por 3 motivos:

a) El paciente no ha experimentado (aún) el evento o no lo va a experimentar.

b) El paciente ha fallecido.

c) El paciente ha abandonado el estudio, por lo que deja de haber un seguimiento.

Una vez establecidos los fundamentos teóricos del análisis de supervivencia, se procede a profundizar en la teoría estadística subyacente en estas técnicas que permite calcular el tiempo de conversión de MCI a AD, así como las probabilidades de supervivencia de los pacientes y el riesgo de conversión. En el contexto del problema que se trata en este trabajo, las probabilidades de supervivencia se podrían interpretar como la probabilidad de cada paciente de que su diagnóstico se mantenga estable en cada momento.

El tiempo hasta que sucede el evento de interés se puede definir como una variable aleatoria positiva T, y dada su función de densidad $f(t)$, la función de distribución acumulada es:

$$F(t) = P(T < t) = \int_{-\infty}^{t} f(u)du$$

La probabilidad de supervivencia $S(t)$ de que el evento de interés no ocurra antes de un tiempo T se define como:

$$S(t) = 1 - F(t) = P(T > t)$$

La función de riesgo $h(t)$ indica la probabilidad de que un evento ocurra en el intervalo $[t + dt)$. Así, la función de riesgo acumulada es:

$$H(t) = \int_{0}^{t} h(u)du$$

La puntuación de riesgo de una muestra x se calcula como la suma del riesgo acumulado en cada punto temporal j, siendo J la cantidad total de partes en que se divide el intervalo de tiempo observado en los datos:

$$r(x) = \sum_{j=1}^{J} H(t_j, x)$$

Los dos modelos estudiados en este trabajo son métodos *ensemble*. Este término se utiliza para denominar métodos en los que se combinan múltiples modelos de aprendizaje automático con el fin de mejorar el rendimiento y la robustez del sistema.

Random Survival Forests (RSF) [11] constituye un método *ensemble* para el análisis de datos de supervivencia. Los RSF extienden la metodología de los RF, creando árboles de decisión independientes y realizando la media de sus predicciones.

Gradient Boosting Survival Analysis (GBSA) es una técnica avanzada de análisis de datos de supervivencia que se basa en el algoritmo de Gradient Boosting [12, 13], una metodología *ensemble* en la que se combinan las predicciones de varios modelos simples, y donde la contribución de cada modelo mejora (o potencia, *"boosts"*) el modelo completo. A diferencia de RSF, un modelo GBSA opera mediante la construcción secuencial de árboles de regresión, donde cada uno se enfoca en mejorar la precisión en áreas donde los anteriores han fallado.

El análisis de supervivencia con ambos modelos se ha llevado a cabo con el paquete de Python *scikit-survival* v0.22.2 (https://github.com/sebp/scikit-survival), que añade compatibilidad con *scikit-learn* e incluye los algoritmos RSF y GBSA. Cabe destacar que *scikit-survival* requiere una versión de *scikit-learn* igual o superior a 1.3.2.

Para evitar que el entrenamiento, evaluación y explicabilidad de los modelos fueran dependientes de la imputación, así como por problemas de compatibilidad con las versiones de las librerías de

Python, se han realizado estas tareas en dos programas separados. De esta manera, el programa de imputación genera una serie de ficheros *csv* con los datos de entrenamiento y de test imputados para cada fold, y el programa de entrenamiento, evaluación y explicabilidad toma estos ficheros como entrada.

2.3.1. Métricas de evaluación

En los modelos de supervivencia no existen las métricas estándar de los modelos de clasificación tradicionales. La naturaleza de estos modelos no es predecir si un evento va a ocurrir o no, sino más bien cuándo ocurrirá dicho evento. De esta manera, en estos modelos se calcula la función de supervivencia de cada individuo para cada uno de los diferentes tiempos estudiados. En el caso concreto de este problema, se calculan unas curvas de supervivencia que representan la probabilidad de cada uno de los pacientes de que su diagnóstico se mantenga estable, tomando varios puntos temporales en cada una de las siguientes visitas anuales desde la visita de baseline hasta la visita en el mes 60.

Una de las métricas más comúnmente utilizadas en modelos de supervivencia es el c-index, que mide la capacidad del modelo para discriminar correctamente entre individuos que experimentan el evento y los que no, teniendo en cuenta también los tiempos de supervivencia de cada uno. De esta forma, un c-index de 0.5 representa que el rendimiento del modelo no es mejor que el azar, mientras que valores cercanos a 1.0 indican una alta capacidad de predicción.

No obstante, para interpretar las prestaciones de los modelos analizados de manera más similar a la de los modelos de aprendizaje automático tradicionales, se ha desarrollado además un método simple para obtener también unas métricas semejantes a las de estos. Más concretamente, se han calculado con este método las métricas de accuracy, sensitivity, specificity y f1-score.

Para ello se ha definido un umbral para las probabilidades de supervivencia calculadas por los modelos en un punto temporal específico (concretamente 5 años desde baseline). Por encima de este umbral se considera que la predicción del modelo es que el paciente va a continuar estable después de 60 meses (y por tanto no desarrolla AD), y por debajo del umbral se considera que la predicción es que el paciente experimentará una evolución de MCI a AD.

Se trata de un enfoque muy similar al de modelos de clasificación binaria, con la definición de un umbral para convertir las probabilidades del modelo en la decisión de pertenencia a una u otra clase. Sin embargo, cabe destacar que las métricas obtenidas a través de este método no reflejan la capacidad del modelo para trabajar con datos longitudinales, sino que solo tiene en cuenta el estado predicho de los pacientes pasados 5 años desde baseline.

Después de varias pruebas y experimentación, se ha determinado que el umbral que produce unos resultados más equilibrados y consistentes es de 0.5, aunque es preciso recordar que un mismo umbral puede no ser el óptimo para cada situación.

Así, la evaluación de los modelos se ha realizado teniendo en cuenta dos enfoques complementarios. Por una parte se han utilizado las curvas de supervivencia y el c-index, de manera que se considera el funcionamiento del modelo en términos de análisis de supervivencia. Por otra parte, se han utilizado las métricas de accuracy, sensitivity, specificity y f1-score para evaluar el modelo de manera más similar a un clasificador binario tradicional.

2.4. Explicabilidad

SHAP (SHapley Additive exPlanations) [14] es un framework diseñado para la interpretación de predicciones de modelos de aprendizaje automático. SHAP se basa en la teoría de juegos, asignando a cada feature un valor Shapley que representa su contribución media a las predicciones del modelo (explicabilidad global), o bien la contribución a una predicción concreta (explicabilidad local).

En este trabajo se ha utilizado SHAP para explicar la influencia de cada feature en las predicciones de cada uno de los modelos analizados. Para ello se ha utilizado el paquete de Python *shap* v0.44.1, y más concretamente se ha utilizado la interfaz *shap.Explainer*.

Dado que la imputación de los datos se ha realizado en cada fold separando el conjunto de entrenamiento del de test, para poder obtener una visión general de la explicabilidad del modelo ha sido necesario combinar los valores SHAP calculados en cada fold en una única matriz. Esto ha implicado además reindexar la matriz de features original utilizando los índices correspondientes a los datos de test en cada fold.

2.4.1. Ejemplo de explicabilidad con SHAP

Para facilitar la interpretación de los resultados de explicabilidad que se van a exponer más adelante, a continuación se incluyen algunos ejemplos ilustrativos sobre la aplicación de SHAP. En estos ejemplos (obtenidos de https://www.kaggle.com/meliao/shap-on-titanic-why-is-rose-alive-but-jack-not) se entrena un modelo de clasificación para predecir la supervivencia de los pasajeros del Titanic en función de features como su género, clase del billete o edad, y se utiliza SHAP para comprobar la relevancia de cada una de estas features.

La Figura 2 muestra una gráfica *beeswarm* de explicabilidad global. En este tipo de gráfica, las features se ordenan de arriba a abajo en orden decreciente de importancia. Cada punto corresponde a un pasajero, y su posición a lo largo del eje x representa el valor del impacto que tiene esa feature sobre las predicciones del modelo. Asimismo, el color de los puntos indica el valor de las features para cada muestra, y permite relacionar fácilmente con qué clase se relacionan unos valores u otros de las features.

En el caso del problema del ejemplo, el impacto de las features se puede entender como la influencia sobre la probabilidad de supervivencia. De esta manera, los valores SHAP negativos contribuyen a probabilidades bajas de supervivencia, mientras que los valores positivos contri-

buyen a probabilidades altas. Observando la gráfica se puede apreciar que el modelo asocia una mayor probabilidad de supervivencia a valores altos de género (femenino), valores bajos de clase (primera clase) y valores bajos de edad (es decir, niños). Esto es coherente con el conocimiento de la tragedia del Titanic, y demuestra que el método de aprendizaje automático es capaz de reconocer las decisiones humanas que influyeron sobre la supervivencia de los pasajeros.

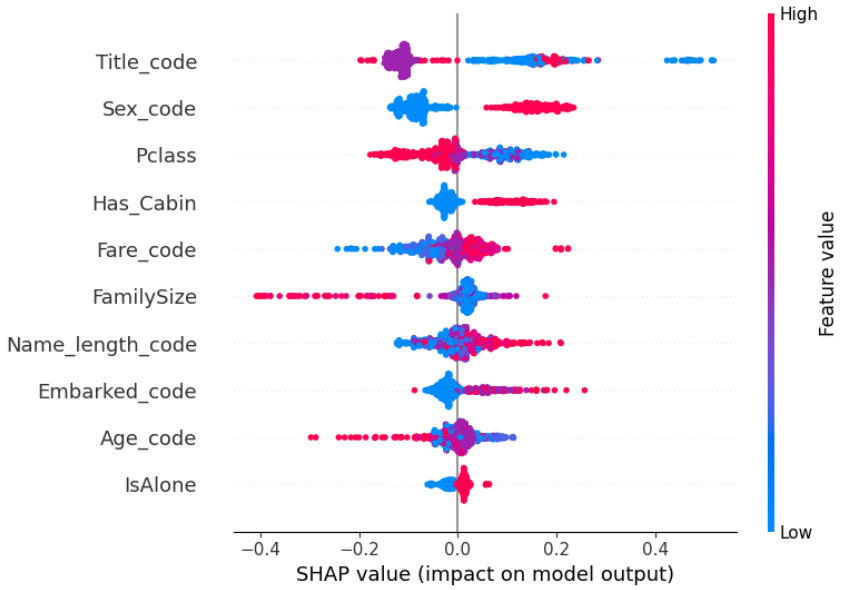
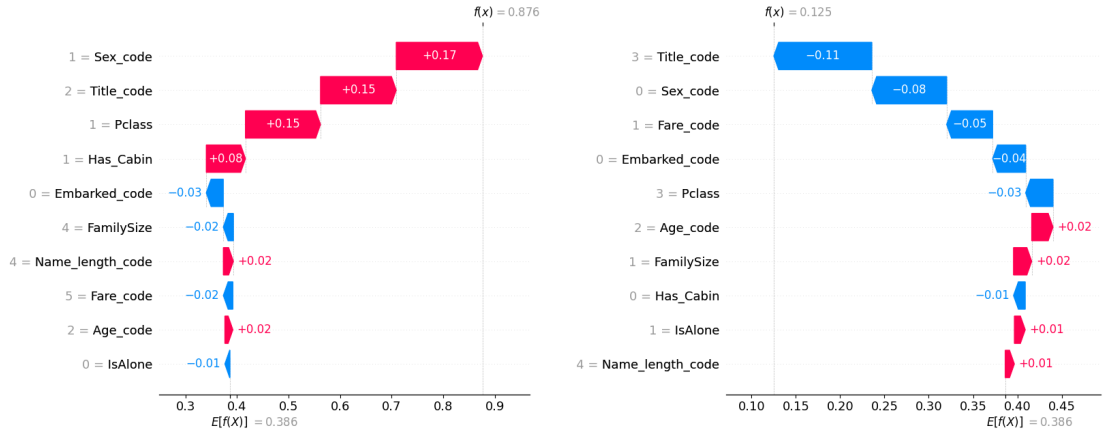


Figura 2: Gráfica SHAP de explicabilidad global representando el impacto de los valores de las features de los pasajeros del Titanic en la probabilidad de supervivencia

SHAP también permite obtener explicaciones para predicciones individuales. De esta manera, para el segundo ejemplo se ha calculado la probabilidad de supervivencia de Rose y Jack, protagonistas de la película Titanic, y se ha comprobado la influencia de sus features sobre las predicciones realizadas por el modelo, que se ajustan con el desenlace de la película.

La Figura 3 muestra gráficas de cascada con las explicaciones locales de Rose y Jack, respectivamente. Como ocurría con la explicabilidad global, las features están ordenadas en orden decreciente de importancia. En la parte inferior se indica el valor base, que es la predicción promedio $E[f(x)]$ realizada por el modelo. A lo largo de la gráfica se muestra cómo se acumulan las contribuciones (positivas o negativas) de las distintas features, comenzando desde este valor base hasta la predicción final, indicada en la parte superior como $f(x)$. Asimismo, a la izquierda de cada feature se indica su valor para esa muestra concreta.

Se puede ver que el modelo predice que la probabilidad de supervivencia de Rose es muy alta, y las features que contribuyen más a esta predicción son que es mujer (además de su título *Miss.*) y que tiene cabina en primera clase. En cambio, la probabilidad de supervivencia de Jack es muy baja, principalmente porque es hombre (y tiene título *Mr.*) y es de tercera clase.



(a) Rose (b) Jack

Figura 3: Gráficas SHAP de explicabilidad local representando la contribución de las features a las probabilidades de supervivencia de los protagonistas de la película Titanic

En el problema tratado en este trabajo se utilizan gráficas *beeswarm* y de cascada de manera similar para establecer la relevancia de las features sobre el riesgo de conversión a AD. Los valores negativos representan pacientes o valores que el modelo interpreta como identificativos de los pacientes sMCI, mientras que los valores positivos identificarían a los pacientes pMCI.

3. Resultados

3.1. Características de los conjuntos de datos utilizados

Se han realizado pruebas con 4 conjuntos de features diferentes. Cada uno de ellos proviene de estudios longitudinales del estado del arte [9, 15-17] realizados con el conjunto de datos de ADNIMERGE, así como también se ha considerado el conjunto de features completo. En la Tabla A.1 del Apéndice A se puede ver una lista con las features que componen cada conjunto.

Para cada conjunto de features se han realizado pruebas con diferentes subconjuntos de pacientes en función de la visita tomada como referencia (bl, m12, m24 o bl+m12). Con el fin de evitar incluir un número muy elevado de pruebas, se ha decidido incluir en este apartado y analizar los resultados obtenidos únicamente con el conjunto de features obtenido del estudio de Sarica et al. [9]. Más comentarios al respecto y los resultados obtenidos para el resto de conjuntos de features se incluyen en el Apéndice A.

En la Tabla 1 se muestra la distribución de los pacientes para cada uno de los conjuntos considerados en función de la visita tomada como referencia. Cabe destacar que en cada caso se toman únicamente los pacientes que tenían MCI en baseline y que todavía tienen MCI en la visita elegida como referencia. De esta manera se explica que el número de pacientes se vaya reduciendo cuando se acerca la visita de referencia al punto de interés. También es destacable que cada vez hay un menor porcentaje de pacientes que terminan evolucionando a AD. Esto puede ser porque una parte importante de los pacientes que tenían más riesgo de empeorar ya lo han hecho cuando llega el mes 24, y por tanto los pacientes que todavía tienen MCI en este punto tienen un menor riesgo de desarrollar demencia.

	Baseline	**Mes 12**	**Mes 24**	**Baseline + m12**
sMCI	321 (65 %)	333 (73 %)	272 (81 %)	280 (71 %)
pMCI	173 (35 %)	121 (27 %)	64 (19 %)	117 (29 %)
Total	494	454	336	397

Tabla 1: Distribución de los pacientes en función de la visita tomada como referencia. El diagnóstico es el realizado en el mes 60 (o el último disponible si no existe)

En la Tabla 2 se indican los porcentajes de valores faltantes para cada una de las features utilizadas. Como se ha comentado anteriormente, estos valores se han imputado mediante el algoritmo MissForest utilizando k-fold. Destaca especialmente la gran cantidad de valores faltantes en DIGITSCOR en los 3 conjuntos, así como en FDG para m12 y en las features CSF (ABETA, TAU y PTAU) para m12 y m24.

	Baseline	**Mes 12**	**Mes 24**
AGE	0.00 %	0.00 %	0.00 %
PTGENDER	0.00 %	0.00 %	0.00 %
PTEDUCAT	0.00 %	0.00 %	0.00 %
APOE4	0.40 %	0.00 %	0.30 %
FDG	23.28 %	78.85 %	44.05 %
ABETA	36.64 %	82.60 %	64.29 %
TAU	36.64 %	82.60 %	64.29 %
PTAU	36.64 %	82.60 %	64.29 %
CDRSB	0.00 %	1.10 %	0.89 %
ADAS11	0.40 %	0.00 %	0.60 %
ADAS13	0.61 %	0.44 %	1.19 %
MMSE	0.00 %	0.22 %	0.60 %
RAVLT_immediate	0.00 %	0.22 %	0.89 %
RAVLT_learning	0.00 %	0.22 %	0.89 %
RAVLT_forgetting	0.00 %	0.44 %	0.89 %
RAVLT_perc_forgetting	0.00 %	0.88 %	1.19 %
LDELTOTAL	0.00 %	0.00 %	1.19 %
DIGITSCOR	54.05 %	59.69 %	63.39 %
TRABSCOR	0.81 %	0.44 %	1.79 %
FAQ	1.01 %	0.88 %	0.89 %
Ventricles	2.63 %	5.29 %	8.04 %
Hippocampus	13.97 %	13.44 %	11.61 %
WholeBrain	0.40 %	1.76 %	2.98 %
Entorhinal	14.37 %	15.42 %	22.32 %
Fusiform	14.37 %	15.42 %	22.32 %
MidTemp	14.37 %	15.42 %	22.32 %
ICV	0.00 %	0.00 %	0.00 %
mPACCdigit	0.00 %	0.00 %	0.60 %
mPACCtrailsB	0.00 %	0.00 %	0.60 %

Tabla 2: Porcentaje de valores faltantes en los datos en función de la visita tomada como referencia

En la Tabla 3 se muestra un resumen con los datos demográficos, clínicos, cognitivos, biomarcadores y MRI de los pacientes sMCI y pMCI, incluyendo la media y desviación típica de todas las features empleadas en el conjunto de Sarica, así como la distribución por género. Estos valores aportan una visión general del conjunto de datos, así como también proporcionan información valiosa que permite comprender de una mejor manera los resultados de explicabilidad local, que se analizarán más adelante.

La tabla incluida corresponde únicamente al conjunto de datos en baseline por presentar un mayor tamaño y distribución equilibrada de pacientes. De esta manera, se considera más representativo para esta tarea.

	sMCI	pMCI
AGE	72.68 ± 7.71	74.21 ± 6.74
PTGENDER (M/F)	213/108	104/69
PTEDUCAT	16.14 ± 2.73	15.92 ± 2.85
APOE4	0.50 ± 0.64	0.85 ± 0.70
FDG	1.26 ± 0.13	1.14 ± 0.12
ABETA	946.97 ± 351.11	690.56 ± 271.47
TAU	251.28 ± 112.62	337.66 ± 124.20
PTAU	24.02 ± 12.38	33.91 ± 14.12
CDRSB	1.22 ± 0.72	1.84 ± 0.93
ADAS11	8.64 ± 3.42	12.62 ± 4.45
ADAS13	13.77 ± 5.24	20.50 ± 6.10
MMSE	28.07 ± 1.69	26.79 ± 1.82
RAVLT_immediate	37.28 ± 10.09	30.01 ± 8.36
RAVLT_learning	4.85 ± 2.53	3.00 ± 2.30
RAVLT_forgetting	4.46 ± 2.48	4.92 ± 2.15
RAVLT_perc_forgetting	51.56 ± 30.50	73.10 ± 29.59
LDELTOTAL	6.99 ± 2.95	3.77 ± 3.22
DIGITSCOR	39.28 ± 10.02	36.15 ± 11.23
TRABSCOR	98.76 ± 48.54	131.83 ± 76.42
FAQ	1.66 ± 2.99	5.31 ± 3.22
Ventricles	0.0251 ± 0.0124	0.0276 ± 0.0133
Hippocampus	0.0047 ± 0.0007	0.0041 ± 0.0007
WholeBrain	0.6863 ± 0.0498	0.6479 ± 0.0448
Entorhinal	0.0025 ± 0.0005	0.0021 ± 0.0005
Fusiform	0.0134 ± 0.0015	0.0119 ± 0.0016
MidTemp	0.0134 ± 0.0015	0.0119 ± 0.0016
ICV	1550235.20 ± 154192.60	1550252.60 ± 165378.99
mPACCdigit	-4.61 ± 3.21	-8.49 ± 3.46
mPACCtrailsB	-4.15 ± 3.09	-8.24 ± 3.22

Tabla 3: Datos de los pacientes de baseline separados en sMCI y pMCI. Se muestra la media y desviación típica con 2 decimales, salvo para features MRI, que al estar normalizadas requieren mayor precisión, y PTGENDER, donde se indica la distribución por género

3.2. Rendimiento de los modelos

Tanto en el caso de RSF como el de GBSA se han entrenado 3 modelos, cada uno de ellos tomando como punto de partida puntos temporales diferentes (bl, m12 y m24), y un cuarto modelo con datos longitudinales (concatenación de features de bl+m12). En todos los casos se ha tomado como referencia para el cálculo de métricas la visita en el mes 60 (de manera que para m12 y m24 se considera el riesgo de conversión al cabo de 4 y 3 años, respectivamente).

Para RSF se han utilizado los hiperparámetros recomendados por el creador de la biblioteca en su documentación, mientras que en el caso de GBSA se han ajustado manualmente para utilizar una configuración análoga a la de RSF, como se puede ver en la Tabla 4.

	Hiperparámetro	Valor por defecto	Valor utilizado
RSF	n_estimators	100	1000
	min_samples_split	6	10
	min_samples_leaf	3	15
GBSA	n_estimators	100	1000
	learning_rate	0.1	0.01

Tabla 4: Hiperparámetros elegidos para RSF y GBSA

En la Tabla 5 se incluyen las métricas obtenidas en cada una de las pruebas con RSF y GBSA. Para facilitar la visualización de los datos se muestran únicamente la media y desviación típica de todos los folds, y para cada métrica se indica en negrita el mejor valor obtenido.

		c-index	accuracy	sensitivity	specificity	f1-score
RSF	Baseline	0.844 ± 0.031	0.777 ± 0.028	0.738 ± 0.061	0.802 ± 0.058	0.697 ± 0.037
	Mes 12	**0.875 ± 0.032**	0.813 ± 0.021	**0.760 ± 0.070**	0.831 ± 0.040	0.682 ± 0.035
	Mes 24	0.872 ± 0.035	**0.836 ± 0.030**	0.653 ± 0.150	0.888 ± 0.042	0.591 ± 0.058
	Baseline + mes 12	0.869 ± 0.027	0.811 ± 0.039	0.746 ± 0.078	0.839 ± 0.074	0.701 ± 0.036
GBSA	Baseline	0.843 ± 0.025	0.788 ± 0.031	0.745 ± 0.077	0.816 ± 0.061	**0.708 ± 0.052**
	Mes 12	0.860 ± 0.050	0.815 ± 0.048	0.727 ± 0.064	0.845 ± 0.059	0.678 ± 0.057
	Mes 24	0.833 ± 0.032	0.804 ± 0.028	0.456 ± 0.100	**0.889 ± 0.049**	0.458 ± 0.116
	Baseline + mes 12	0.843 ± 0.033	0.791 ± 0.051	0.713 ± 0.095	0.824 ± 0.067	0.670 ± 0.056

Tabla 5: Media y desviación típica de las métricas obtenidas para cada visita con RSF y GBSA

En cuanto a RSF, parece que al acercar el punto de partida al evento de interés (es decir, al mes 60) se obtiene un c-index más alto (0.84 en bl, 0.87 en m12 y m24). Esto significa que el modelo tiene una mejor capacidad para predecir el riesgo de conversión según avanza el tiempo. Esto probablemente se debe a que, con el paso de los meses, los pacientes con la enfermedad más avanzada muestran más cambios en su estado y nuevos biomarcadores que ayudan a reconocer ese riesgo. En el resto de métricas, salvo el f1-score, se puede apreciar también un mejor rendimiento del modelo de m12 respecto de bl, lo cual se puede atribuir al mismo motivo.

En cambio, en m24 aumentan accuracy y sensitivity respecto de bl y m12, mientras que hay una pérdida considerable de sensitivity y f1-score (0.65 y 0.59 respectivamente frente a 0.76 y 0.68 en m12). Esto probablemente se debe a la disminución del número de pacientes entre el conjunto de m24 y el de m12, así como también al hecho de que el porcentaje de pacientes pMCI se reduce hasta un 19 % (ver Tabla 1). El hecho de que el modelo mejore en accuracy y specificity pero empeore tanto en sensitivity y f1-score puede deberse a que el modelo adquiera cierta tendencia a predecir que los pacientes se mantienen estables (clase mayoritaria), incurriendo de esta manera en un mayor número de falsos negativos.

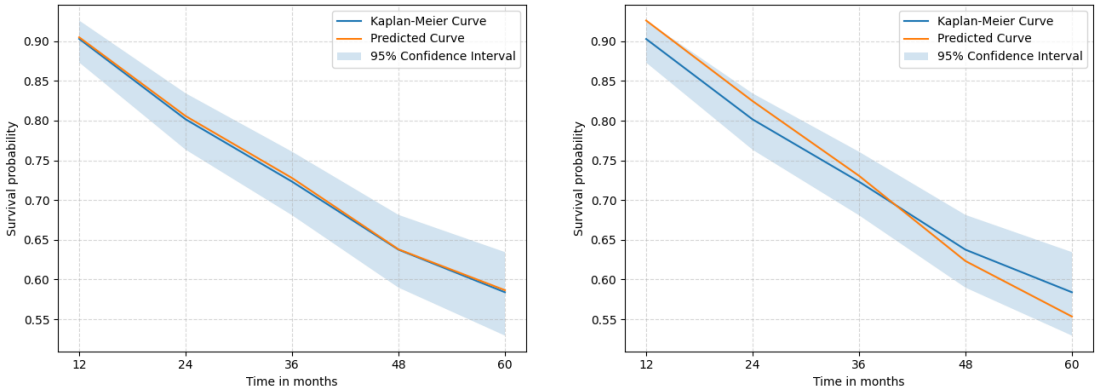
Por otra parte, al utilizar datos longitudinales (incluyendo datos de bl y m12 al mismo tiempo) se obtiene un modelo equilibrado, con un c-index similar al de m12 aunque algo inferior (en torno a 0.87 en ambos casos). En cuanto a las métricas de clasificación, el rendimiento es muy similar

al de m12, con mejoras en la specificity (0.84 frente a 0.83 en m12) y el f1-score (0.70 frente a 0.68 en m12).

En los modelos de GBSA se observa una tendencia similar a la de RSF, aunque en m24 se reduce todavía más su sensitivity y su f1-score (0.46 en ambos casos respecto de 0.73 y 0.68 en m12). Asimismo, al incluir datos longitudinales también se obtienen unos resultados similares a m12, aunque en este caso inferiores y con más diferencia entre ellos respecto a lo que ocurría en RSF.

Globalmente se puede observar que, en líneas generales, RSF rinde mejor que GBSA, especialmente con datos longitudinales y en m24, si bien es cierto que los resultados de GBSA en baseline son algo superiores. Cabe destacar que en cada modelo se ha experimentado con diferentes umbrales de decisión para el cálculo de las métricas de clasificación, especialmente para los modelos de m24. Aunque se ha conseguido en algunos casos mejorar la sensitivity y f1-score, la variabilidad entre folds aumentaba considerablemente, por lo que se ha mantenido el umbral en 0.5. De esta manera se garantizan unos resultados más estables que no dependan tanto de la población concreta de pacientes con la que se esté entrenando el modelo.

Además, se han dibujado las curvas de supervivencia de RSF y GBSA (ver Figura 4) en el conjunto de baseline y se han comparado con la curva de Kaplan-Meier [18], un método que permite estimar la función de supervivencia real de los datos. De esta manera se pueden comparar las predicciones de ambos modelos con una estimación de la evolución real de los pacientes del conjunto de test.



(a) Random Survival Forests

(b) Gradient Booster

Figura 4: Comparación de las curvas de supervivencia predichas por los modelos RSF y GBSA en baseline con la curva Kaplan-Meier con intervalo de confianza del 95 %

Para ello, se han calculado las probabilidades de supervivencia medias de todos los pacientes en cada fold. Esto se ha hecho únicamente con el conjunto de bl con el fin de que los resultados sean más representativos y se incluya todo el intervalo temporal de seguimiento de los pacientes.

Observando la Figura 4, se puede apreciar que la precisión de RSF es bastante alta, pues la curva de supervivencia predicha es muy cercana a la Kaplan-Meier. Además, el rendimiento parece no disminuir con el transcurso de los meses. En cambio, el error de GBSA es mayor, especialmente en los meses 12 y 60. Es preciso notar además que, hasta el mes 36, el modelo GBSA tiende a ser más optimista en el cálculo del riesgo de conversión, mientras que a partir del mes 48 ocurre lo contrario. Si bien es cierto que este comportamiento pesimista a partir del mes 48 se podría considerar deseable, sobre todo si se tiene en cuenta que en una enfermedad tan poco predecible como el AD es muy difícil hacer diagnósticos acertados, no se puede desestimar el hecho de que RSF tiene un error notablemente menor respecto de la curva Kaplan-Meier.

3.3. Estudio de la explicabilidad

3.3.1. Explicabilidad global

Para el estudio de la explicabilidad global se han obtenido las explicaciones agrupadas de todos los folds de los modelos RSF y GBSA para cada uno de los conjuntos utilizados (bl, m12, m24 y bl+m12). De esta manera, se van a comprobar cuáles son las features que más influyen sobre las predicciones de los modelos, así como si estas son consistentes al cambiar la visita de referencia utilizada y, en caso contrario, analizar el motivo.

En la Figura 5 se muestra la explicabilidad global de los modelos entrenados con las visitas de baseline. En ella se puede ver que ambos modelos coinciden en 4 de las 5 features más importantes (mPACCtrailsB, FAQ, LDELTOTAL y MidTemp), siendo la única diferencia el orden de las dos primeras. En el resto del ranking también guardan ciertas similitudes.

En ambos casos se observa que los modelos dan una gran importancia a features de evaluación neuropsicológica, como los ya mencionados mPACCtrailsB (1º en RSF y 2º en GBSA), que mide los primeros signos de declive cognitivo, y LDELTOTAL (4º en ambos rankings), que evalúa la memoria episódica. Por otra parte, features como Hippocampus, que mide el volumen del hipocampo, una de las estructuras cerebrales más dañadas por el avance de la enfermedad, aparecen muy abajo en el ranking. De esta manera, queda claro que en baseline todavía hay muchos pacientes que, incluso si van a progresar a AD, todavía están en una fase temprana de su deterioro anatómico.

En cuanto a las diferencias entre ambos modelos, se puede comprobar que sobre el modelo GBSA tienen más influencia los biomarcadores del líquido cefalorraquídeo (ABETA, TAU y PTAU) que en RSF, donde aparecen más abajo. También tiene una mayor importancia FDG, la fluorodesoxiglucosa media del metabolismo cerebral (7º en RSF y 3º en GBSA). Asimismo, cabe destacar que en el modelo RSF tiene una gran importancia mPACCdigit, que también mide algunos de los primeros signos de deterioro, mientras que para GBSA no aparece en el ranking.

En los modelos entrenados con la visita en el mes 12 se observa una tendencia similar (ver Figura 6), coincidiendo en 4 de las 5 features más importantes, aunque con ciertos cambios. Entre estas

features destaca ADAS13 (2º en ambos rankings), que mide el estado cognitivo con 13 pruebas de memoria y lenguaje, y FDG, que en baseline no entraba en el top 5 de RSF. Resulta notable también que ADAS11, una versión reducida de ADAS13, aparezca en el puesto 11 del ranking de RSF; sin embargo, no figura en la gráfica de GBSA.

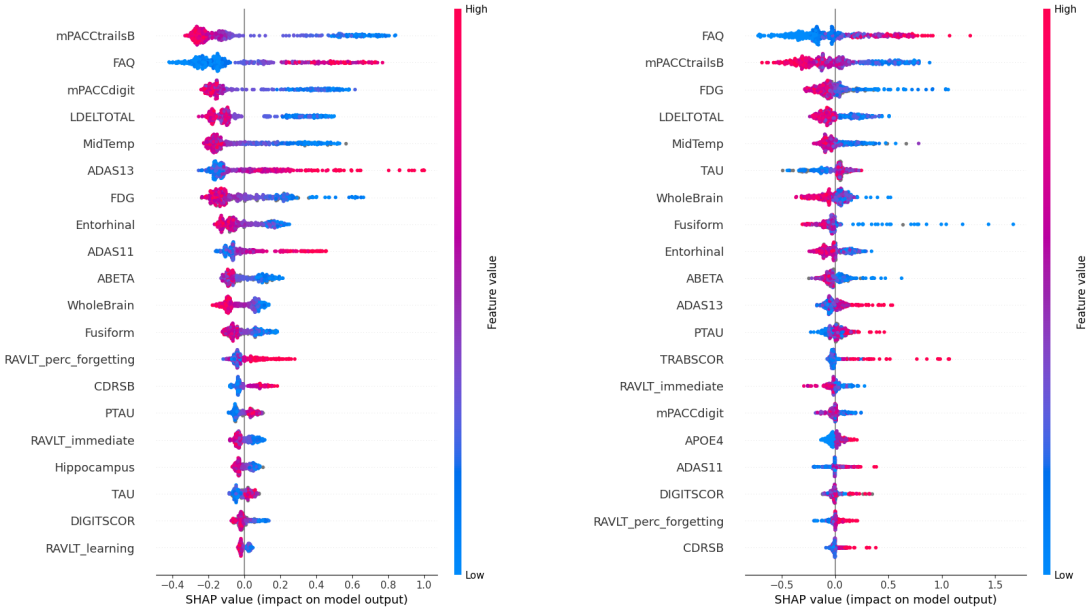
Se mantiene el peso de mPACCtrailsB, LDELTOTAL y MidTemp sobre la salida de los modelos, aunque FAQ pierde importancia en el caso de RSF respecto de baseline. También cabe destacar que en RSF ganan algo de importancia los biomarcadores CSF, mientras que en el caso de GBSA la pierden. Por otra parte, parece que el volumen del hipocampo tiene un mayor peso en el mes 12, especialmente en el modelo GBSA. Aunque sigue sin ser tener un gran impacto sobre las predicciones de los modelos, esta evolución entre baseline y el mes 12 permite ver cómo con el paso de los meses se va deteriorando de manera progresiva esta estructura anatómica.

Por último, al entrenar los modelos con datos longitudinales se tienen en cuenta tanto las features de baseline como las de la visita en el mes 12. Dado que esta última representa implícitamente cierta evolución de la enfermedad respecto de baseline por ser más cercana al evento de interés, es razonable que la mayor parte de las features que aparecen en las gráficas correspondientes (ver Figura 8) sean del mes 12. Por otra parte, algunas de las más importantes entre las features de baseline son FAQ, MidTemp y FDG, que también se encontraban entre las 3 más importantes en los modelos de baseline (ver Figura 5).

Se puede interpretar que la escala clínica FAQ es realmente informativa, hasta el punto de que las puntuaciones obtenidas en baseline tienen un mayor impacto sobre los modelos longitudinales que muchas de las features del mes 12. De hecho, FAQ en baseline aparece entre las 5 features más importantes para el modelo GBSA y entre las 10 más importantes para RSF. Lo mismo sucede con MidTemp y FDG, aunque con menor impacto sobre los modelos que FAQ.

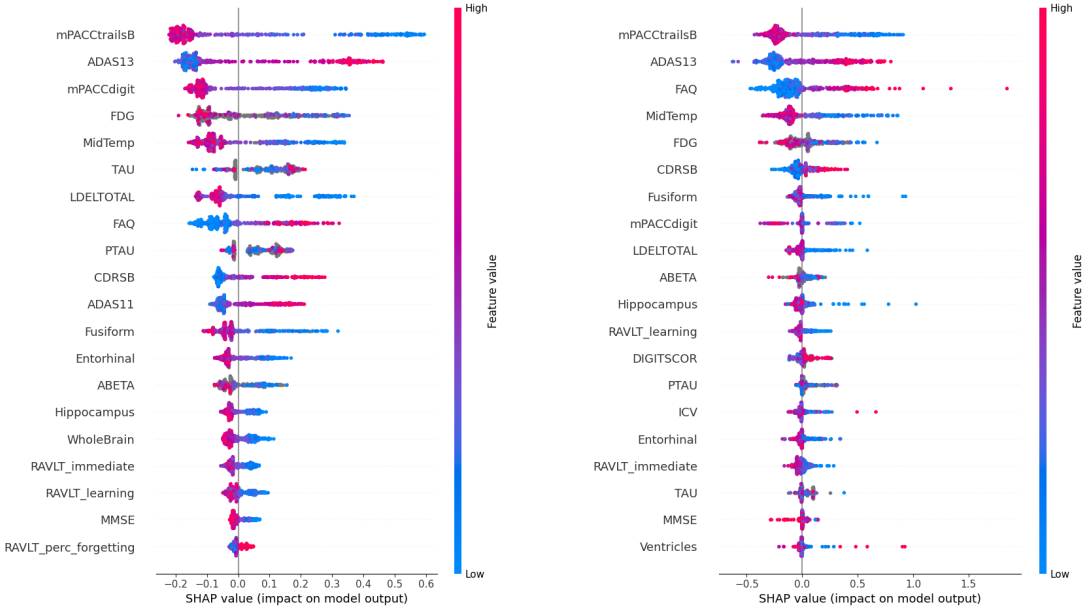
Teniendo todo esto en cuenta, se puede razonar que en ambos modelos hay algunas features importantes que parecen menos independientes del avance de la enfermedad, especialmente mPACCtrailsB, FAQ y MidTemp. Por tanto, se podría considerar que mPACCtrailsB es una prueba realmente útil que puede ayudar a identificar el riesgo de progresión a AD, así como FAQ parece ser una escala clínica no sólo mucho más informativa que CDRSB, sino también más informativa que muchos otros tests cognitivos y pruebas de diagnóstico, especialmente en etapas tempranas de la enfermedad, aunque con el transcurso de los meses la diferencia se reduce. Asimismo, MidTemp es la única feature de MRI que mantiene una importancia muy alta desde el principio, por lo que se podría interpretar que el giro temporal medio puede ser una de las estructuras cerebrales más afectadas inicialmente, mientras que otras como el hipocampo tardan más en deteriorarse.

Por último, se puede concluir que los resultados obtenidos son coherentes con el conocimiento clínico de la enfermedad, y ambos modelos coinciden en las features que más contribuyen a identificar los signos de deterioro asociados con la progresión a la enfermedad de Alzhéimer.



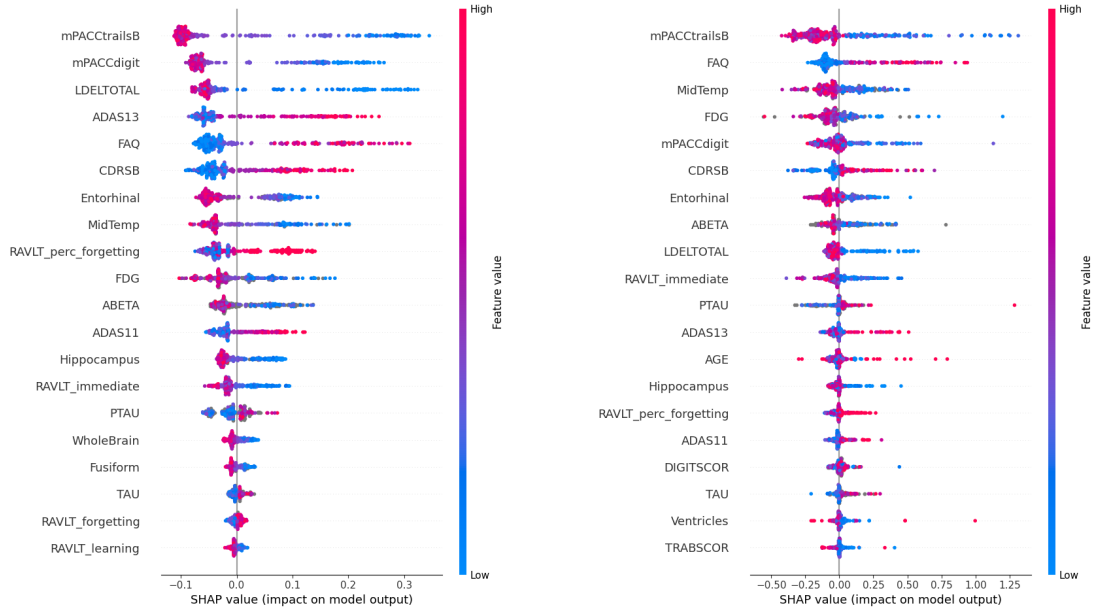
(a) Random Survival Forests (b) Gradient Booster

Figura 5: Gráficas SHAP de explicabilidad global de los modelos RSF y GBSA entrenados con la visita de baseline



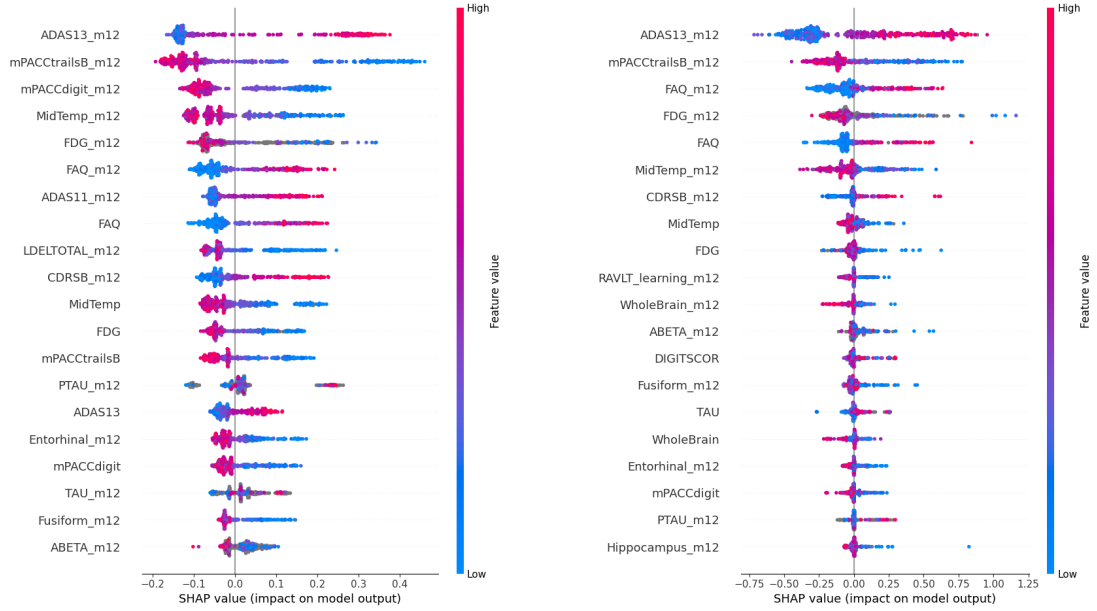
(a) Random Survival Forests (b) Gradient Booster

Figura 6: Gráficas SHAP de explicabilidad global de los modelos RSF y GBSA entrenados con la visita en el mes 12



(a) Random Survival Forests (b) Gradient Booster

Figura 7: Gráficas SHAP de explicabilidad global de los modelos RSF y GBSA entrenados con la visita en el mes 24



(a) Random Survival Forests (b) Gradient Booster

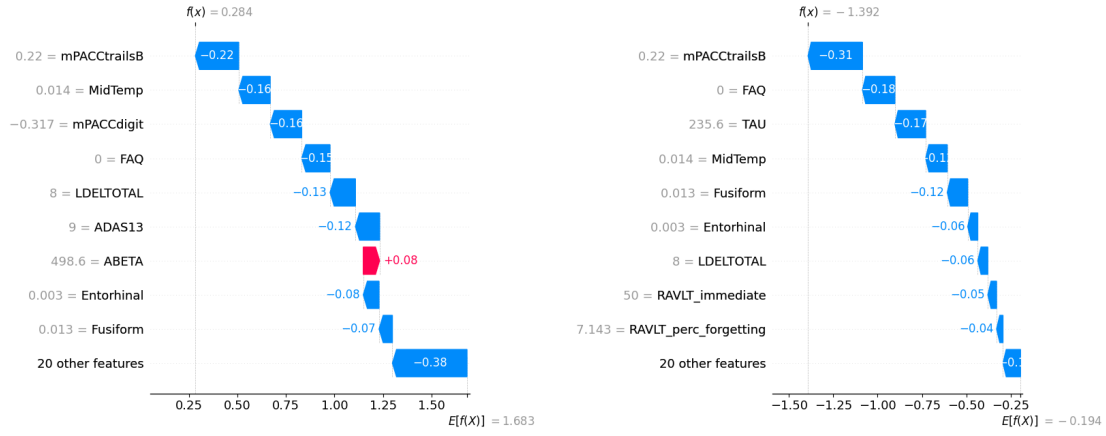
Figura 8: Gráficas SHAP de explicabilidad global de los modelos RSF y GBSA entrenados con datos longitudinales (visitas de baseline y mes 12)

3.3.2. Explicabilidad local

Para el análisis de la explicabilidad local se ha optado por realizar predicciones individuales con ambos modelos entrenados únicamente sobre la visita de baseline, ya que realizar pruebas con todos los conjuntos de datos supondría un gran número de experimentos. Además, resultaría casi imposible realizar las pruebas sobre los mismos individuos, pues la población de pacientes varía cuando se modifica la visita con la que se va a entrenar el modelo, así como también los pacientes concretos incluidos en cada fold.

De esta manera, se han escogido 4 pacientes, denominados sMCI#1, sMCI#2, pMCI#1 y pMCI#2 de ahora en adelante, que se han considerado interesantes por representar casos variados de declive cognitivo. Más concretamente:

- El paciente sMCI#1 no desarrolla AD al cabo de 5 años, y tiene un periodo de seguimiento de 60 meses, por lo que no hay censura de datos. Ambos modelos predicen correctamente la no conversión del paciente.
- El paciente sMCI#2 no desarrolla AD al cabo de 5 años, y se tienen datos del mismo hasta el mes 36, por lo que hay censura de datos. Ambos modelos predicen progresión a AD. Debido a la censura de datos, se puede evaluar si es error del modelo o no analizando las explicaciones locales.
- El paciente pMCI#1 tiene un diagnóstico de AD a partir del mes 48. Ambos modelos predicen incorrectamente que el paciente no desarrolla AD.
- El paciente pMCI#2 tiene un diagnóstico de AD a partir del mes 12. Ambos modelos predicen correctamente la progresión a AD.

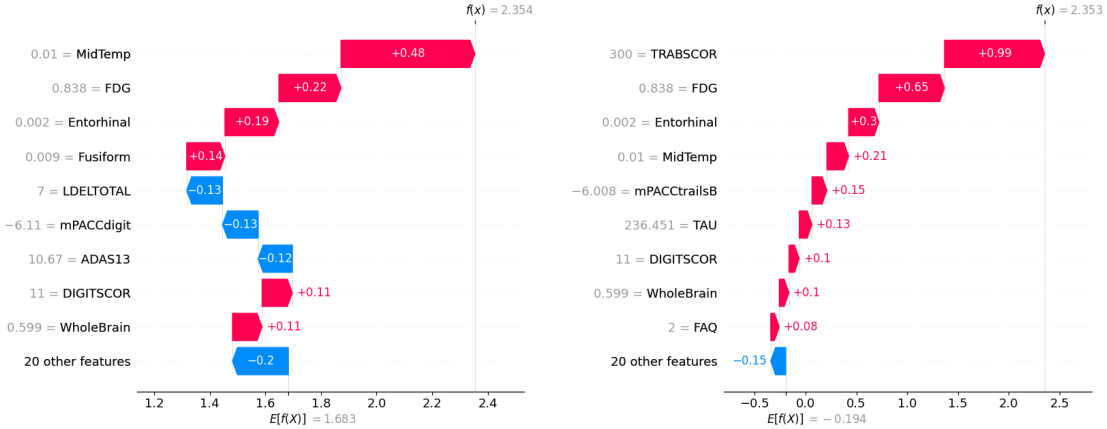


(a) Random Survival Forests (b) Gradient Booster

Figura 9: Gráficas SHAP de explicabilidad local de los modelos RSF y GBSA para el paciente sMCI#1

Tanto RSF como GBSA predicen correctamente que el paciente sMCI#1 no desarrolla AD en los siguientes 5 años. Ambos modelos predicen curvas bastante estables (ver Figura 13), con probabilidades muy altas de que el paciente se mantenga estable pasado el mes 60. Por otra parte, como se puede comprobar en la Figura 9, en RSF no parece haber una feature que contribuya considerablemente más que el resto a dicha predicción, sino que más bien es la contribución acumulada de varias features. En el caso de GBSA también ocurre algo similar, si bien destaca el impacto de mPACCtrailsB, FAQ y TAU.

Como se ha comprobado en el análisis de explicabilidad global, se puede apreciar también aquí que las pruebas mPACC y la escala FAQ aportan mucha información útil para predecir el diagnóstico. Asimismo, un volumen alto en estructuras cerebrales como el giro temporal medio (MidTemp), la corteza entorrinal (Entorhinal) o el giro fusiforme (Fusiform) denota que estas no están afectadas por la enfermedad.



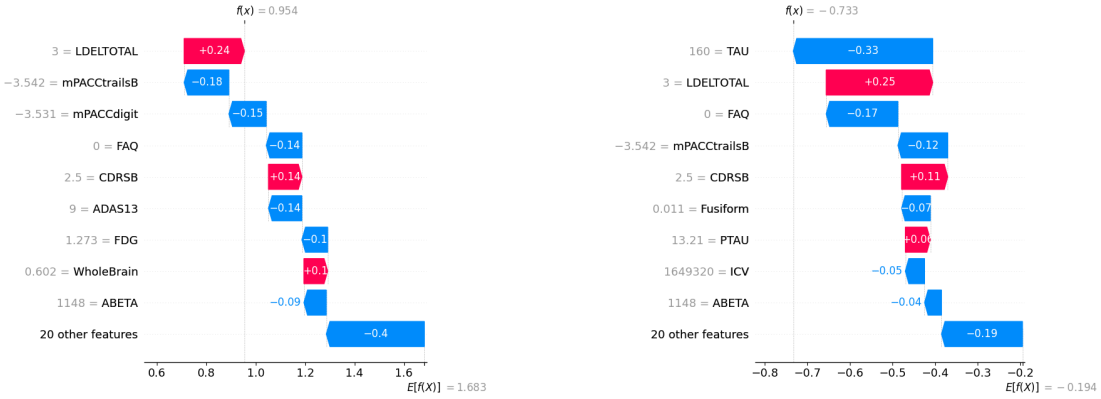
(a) Random Survival Forests

(b) Gradient Booster

Figura 10: Gráficas SHAP de explicabilidad local de los modelos RSF y GBSA para el paciente sMCI#2

Se sabe que el paciente sMCI#2 no tiene AD a los 36 meses, pero al no contar con más datos, no se conoce su diagnóstico a los 5 años. En cambio, ambos modelos predicen que sí habrá conversión. Según los valores SHAP del modelo RSF (ver Figura 10a), esta predicción viene principalmente dada por valores bajos en las features MRI (MidTemp, Entorhinal y Fusiform), que indican un volumen algo por debajo de lo normal en estas estructuras cerebrales. Asimismo, también se asocian valores bajos de FDG con progresión a AD. En cambio, el modelo GBSA (ver Figura 10b) interpreta un riesgo considerablemente más alto de conversión, principalmente debido a un valor de TRABSCOR muy por encima de la media (131.83 de media para pMCI, mientras que este tiene un valor de 300). Destaca también la contribución de las features MRI y FDG, al igual que en RSF. La principal diferencia es que algunas de las features que más contribuyen negativamente para este paciente en RSF no lo hacen en GBSA.

Su curva de supervivencia en RSF (ver Figura 13a) muestra un aumento significativo del riesgo de conversión justo entre el mes 36 y el 48, por lo que no se puede saber con seguridad si realmente se puede considerar como un error del modelo, ya que es posible que el paciente hubiera terminado desarrollando AD después del tercer año como el modelo predice. Sin embargo, en GBSA la curva de supervivencia (ver Figura 13b) muestra un deterioro rápido y pronunciado, donde el paciente desarrollaría demencia a partir del mes 24. En este caso sí se podría afirmar que se trata de un error del modelo.

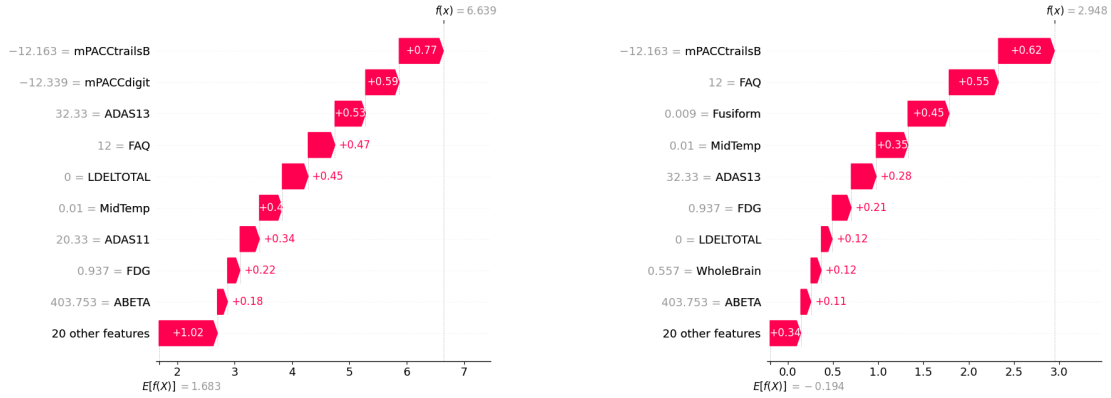


(a) Random Survival Forests (b) Gradient Booster

Figura 11: Gráficas SHAP de explicabilidad local de los modelos RSF y GBSA para el paciente pMCI#1

Tanto RSF como GBSA predicen que el paciente pMCI#1 no desarrolla AD en los 5 años posteriores a la visita de baseline, cuando en realidad lo hace en el mes 48. En el caso de RSF (ver Figura 11a), las features que más contribuyen a la identificación del paciente como no conversor son las pruebas mPACC y ADAS13 y la escala FAQ, como ocurría en el paciente sMCI#1. Por otra parte, en GBSA (ver Figura 11b) destaca también el impacto de TAU. No obstante, hay otras features como la prueba de memoria episódica LDELTOTAL o la escala clínica CDRSB que parecen indicar correctamente la conversión del paciente. De esta manera, se aprecia cierta contradicción, sobre todo entre FAQ y CDRSB, siendo ambas escalas clínicas que pretenden medir el grado de severidad del deterioro cognitivo o la etapa dentro de este.

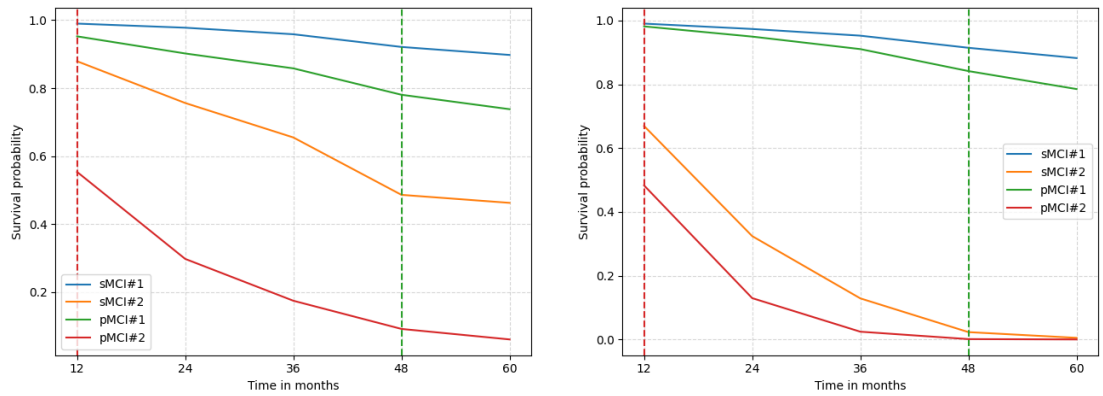
Las curvas de supervivencia (ver Figura 13) no muestran un declive significativo, con solo una disminución algo mayor en ambos modelos de la probabilidad de supervivencia a partir del mes 48, si bien el riesgo al cabo de 5 años sigue siendo bajo. En este caso, parece que el paciente muestra algunos signos de declive (LDELTOTAL, CDRSB), mientras que en pruebas como FAQ muestra un funcionamiento todavía normal que no afecta a su vida diaria. De esta manera, es probable que el deterioro se produjera de manera más repentina, por lo que sería interesante comprobar las predicciones de los modelos si se hubieran entrenado con datos posteriores.



(a) Random Survival Forests (b) Gradient Booster

Figura 12: Gráficas SHAP de explicabilidad local de los modelos RSF y GBSA para el paciente pMCI#2

Por último, ambos modelos predicen correctamente que el paciente pMCI#2 experimenta el evento de conversión dentro de los 5 años de seguimiento. Teniendo en cuenta la prontitud del diagnóstico real del paciente (12 meses), se puede interpretar que el individuo tenía un riesgo muy alto de conversión. Esto se puede ver más claro tanto en los valores SHAP (ver Figura 12) como en la curva de supervivencia (ver Figura 13), donde además de tener probabilidades de supervivencia bajas desde el principio, hay una disminución considerable a partir del mes 12. Prácticamente todas las features parecen contribuir a la predicción, lo que indica el nivel de deterioro del paciente. Entre estas features destacan principalmente las pruebas mPACC y ADAS13, la escala FAQ, y el volumen del giro temporal medio (MidTemp). También destaca el volumen del giro fusiforme en GBSA, si bien no parece tener un gran impacto en el caso de RSF.



(a) Random Survival Forests (b) Gradient Booster

Figura 13: Curvas de supervivencia en RSF y GBSA de los pacientes sMCI#1, sMCI#2, pMCI#1 y pMCI#2

4. Discusión

Se han realizado pruebas muy variadas sobre RSF y GBSA en la predicción del riesgo de conversión a AD en los 5 años siguientes a la visita de baseline. En estas pruebas se han analizado las capacidades de los modelos en cuanto a análisis de supervivencia, así como el rendimiento de ambos al cambiar el enfoque e interpretar las curvas de supervivencia de una manera más similar a un clasificador binario.

Se han considerado los modelos entrenados en diversas situaciones temporales. Para ello, se han elaborado distintos conjuntos de datos cambiando la visita de baseline por visitas más cercanas al evento de interés (m12 y m24), así como también entrenando los modelos con datos longitudinales (bl+m12). De esta manera se analiza el comportamiento de los modelos con datos de varias visitas al mismo tiempo, es decir, teniendo en cuenta la evolución del paciente.

En resumen, se podría decir que ambos modelos se benefician de utilizar datos más cercanos al evento de conversión, pues en ellos el deterioro cognitivo se hace más notable y es algo más sencillo distinguir los pacientes que se van a mantener estables de los que no. Sin embargo, debido a la inevitable disminución de la cantidad de pacientes con el paso de los meses, así como también al creciente desbalanceamiento de clases, los resultados obtenidos se vuelven más inestables al entrenar con datos del mes 24, obteniendo además valores bajos de sensitivity y f1-score. De esta manera, es razonable pensar que esta tendencia se mantendría si se utilizan datos de meses posteriores.

Asimismo, parece que utilizar datos longitudinales produce resultados satisfactorios y mejores que si se utilizan únicamente los datos de baseline, aunque no se puede decir que haya una mejora considerable respecto a utilizar únicamente las visitas del mes 12. Sería interesante continuar investigando al respecto, quizá teniendo en cuenta más visitas del paciente. Sin embargo, también hay que considerar que un número demasiado elevado de features podría producir sobreajuste en el modelo.

Si se compara el rendimiento de ambos modelos se puede concluir que, en general, RSF muestra un mejor comportamiento que GBSA, así como también produce resultados más consistentes y responde mejor al utilizar datos de otras visitas y datos longitudinales. Además, al comparar las curvas predichas por ambos modelos con la curva de Kaplan-Meier para los datos de test (ver Figura 4), las predicciones de RSF se ajustan más en media a los datos reales que las predicciones de GBSA.

En cuanto a la explicabilidad, se han analizado las features que más impacto producen sobre la salida de ambos modelos. Además, se ha comprobado también cómo cambia la explicabilidad en función de los datos utilizados para el entrenamiento. Las features mPACC han resultado ser las más importantes para ambos modelos, independientemente de la visita considerada. También han destacado otras como FAQ, FDG, LDELTOTAL o MidTemp, que se han encontrado siempre entre las más importantes. No se han observado diferencias considerables entre los modelos

en este aspecto, aunque sí algunas diferencias menores en el orden, o algunas features que se podrían considerar de importancia media que varían más dependiendo de la visita considerada, como ADAS13, los biomarcadores CSF (ABETA, TAU y PTAU), algunas de las features MRI (Entorhinal, Fusiform) o CDRSB.

Cabe destacar que se ha observado que GBSA parece verse más afectado por valores atípicos en los datos que RSF, como evidencian las explicaciones locales del paciente sMCI#2 (ver Figura 10b), cuyo valor de TRABSCOR es muy superior a la media de los pacientes pMCI (ver Tabla 3). Es interesante comprobar cómo en RSF ni siquiera aparece esta feature. Este es uno de los principales motivos por los que la curva de supervivencia predicha para este paciente difiere tanto entre ambos modelos.

Como se ha comentado anteriormente, este trabajo está inspirado en un estudio de Sarica et al. [9] donde se utiliza RSF para la predicción del riesgo de conversión de MCI a AD en un periodo de tiempo de 4 años trabajando únicamente con datos de baseline. En dicho estudio se obtiene un c-index de 0.890 sobre el conjunto de test, superior al 0.844 en baseline de este trabajo. Sin embargo, cabe destacar que en el estudio se utiliza la biblioteca *pysurvival* para la implementación de Random Survival Forests, mientras que en este trabajo se escogió *scikit-survival* por ofrecer además una implementación de Gradient Booster.

En cuanto a la explicabilidad, las 5 features más importantes en el estudio de Sarica son FDG, ABETA, HCI, FAQ y RAVLT_immediate, de las cuales FDG y FAQ coinciden con los resultados obtenidos en este trabajo. No obstante, es preciso mencionar que los datos utilizados en este estudio provienen de varios conjuntos de datos además de ADNIMERGE, por lo que es posible que hubiera diferencias en el conjunto final, sobre todo en cuanto a datos faltantes. Asimismo, la feature HCI no se encuentra en ADNIMERGE, motivo por el cual no se ha tenido en cuenta en este trabajo.

Por último, se ha observado una diferencia inesperada en las gráficas SHAP de explicabilidad global. En el análisis realizado en este trabajo se puede observar que, en las features mPACC, los valores bajos contribuyen positivamente a la salida del modelo, mientras que los valores altos contribuyen negativamente. Esta tendencia es coherente con las medias observadas para estas features según el tipo de paciente (ver Tabla 3). Sin embargo, en el análisis realizado por Sarica ocurre lo contrario: los valores altos tienen una contribución positiva, y los valores bajos tienen una contribución negativa. No se ha encontrado una explicación a esta discrepancia, aunque se considera que podría tener relación con que estas features suelen tener valores negativos.

De esta manera, se puede concluir que las principales diferencias observadas en los resultados de ambos estudios vienen derivadas de la utilización de bibliotecas distintas para la implementación de los modelos de supervivencia y de diferentes conjuntos de datos. No obstante, a pesar de estas diferencias, en ambos casos se demuestra la efectividad de RSF en el problema tratado, así como la utilidad de SHAP para una adecuada interpretación de los resultados.

4.1. Relación de la explicabilidad con el conocimiento clínico

Las features mPACC (digit y trailsB) consisten en métricas calculadas teniendo en cuenta las puntuaciones de varios tests cognitivos, donde uno de ellos puede ser de sustitución de dígitos (mPACCdigit, que considera el Digit Symbol Substitution) [19] o de coordinación visomotora (mPACCtrailsB, que considera la parte B del Trail Making Test) [20].

En especial mPACCtrailsB ha resultado ser una de las features con más influencia sobre ambos modelos (ver Figuras 5–8), lo que parece indicar que es una medida que permite reconocer los primeros signos de declive cognitivo a través de la composición de las puntuaciones de varios tests. Además, se trata de la feature que más ha contribuido en ambos modelos a predecir correctamente el diagnóstico de los pacientes sMCI#1 (ver Figura 9) y pMCI#2 (ver Figura 12).

Aunque mPACCdigit también tiene un gran impacto sobre la salida del modelo RSF y contribuye a realizar predicciones correctas, tanto globalmente como en los pacientes sMCI#1 (ver Figura 9a) y pMCI#2 (ver Figura 12a), no es tan importante en GBSA. Además, resulta interesante que mPACCdigit es una de las features que más contribuyen a diferenciar correctamente el paciente sMCI#2 (ver Figura 10a), aunque finalmente el peso de otras features con contribución positiva es mayor. No obstante, es necesario recordar que esta predicción en concreto puede considerarse o no un error del modelo, pues se finaliza el seguimiento del paciente después del mes 36.

FAQ es una escala clínica que mide la capacidad funcional del individuo en actividades de la vida diaria. Según los resultados de explicabilidad obtenidos en ambos modelos, es una de las features que más contribuye a reconocer pacientes que desde el principio comienzan a mostrar problemas funcionales. Además, como se ha comprobado en el análisis con datos longitudinales (ver Figura 8), la puntuación obtenida en FAQ en baseline aporta más información al modelo que muchas features obtenidas un año después, por lo que resulta un dato que contribuye significativamente al diagnóstico temprano de la demencia.

Localmente se ha observado que contribuye a la predicción correcta de los pacientes sMCI#1 (ver Figura 9a) y pMCI#2 (ver Figura 12a), mientras que es uno de los principales motivos por los que ambos modelos predicen incorrectamente el diagnóstico del paciente pMCI#1 (ver Figura 11a). Esto se debe a que el paciente obtiene una puntuación de 0 en FAQ, que indica independencia total en todas las tareas sobre las que se le pregunta [21], y normalmente se asociaría con un envejecimiento cognitivo normal.

La feature FDG, que mide la cantidad media de fluorodesoxiglucosa en los cíngulos angular, temporal y posterior del cerebro, resulta fundamental en el estudio del Alzhéimer. Niveles bajos de FDG reflejan una disminución del metabolismo de glucosa en estas áreas, lo que a su vez se ha asociado con el desarrollo del AD [22]. Esto se ve reflejado además en su importancia sobre las predicciones de ambos modelos, siendo mayor su impacto sobre GBSA (ver Figuras 5–7b), mientras que en RSF tiene más importancia en el mes 12 (ver Figura 6a). Además, se posiciona como la tercera feature de baseline más importante con datos longitudinales (ver Figura 8).

En ambos modelos FDG contribuye a la predicción de conversión del paciente pMCI#2 (ver Figura 12), aunque hay varias features con mayor impacto. Asimismo, es una de las features más importantes en la predicción de AD en el paciente sMCI#2 (ver Figura 10). Los valores atípicos de FDG se han relacionado con un 72.82 % de los casos de pacientes pMCI [23], por lo que quizá si se hubiera continuado el seguimiento del paciente su diagnóstico hubiera cambiado y la predicción hubiera terminado siendo correcta.

LDELTOTAL es una medida de la memoria episódica que se calcula en base a la cantidad de elementos correctos que el paciente recuerda de una historia breve [24]. La memoria episódica es una de las primeras capacidades cognitivas en verse afectadas por la etapa preclínica de la enfermedad de Alzhéimer [25], lo que explica que LDELTOTAL tenga una gran influencia sobre las predicciones de los modelos. A nivel global, se encuentra entre las features más importantes en RSF, aunque no parece influir tanto al trabajar con datos longitudinales (ver Figura 8a). Por otra parte, su impacto en GBSA es más moderado, a excepción de baseline (ver Figura 5b), donde se encuentra entre las 5 features más importantes.

En cada una de las explicaciones locales, LDELTOTAL se encuentra entre las 5 features que más impacto tienen en la salida de RSF. Además de contribuir a la predicción correcta de los pacientes sMCI#1 (ver Figura 9a) y pMCI#2 (ver Figura 12a), es la feature que más contribuye positivamente a la conversión del paciente pMCI#1 (ver Figura 11a), con un mayor impacto sobre el modelo que los tests mPACC y la puntuación obtenida en la escala FAQ. Como ya se ha comentado anteriormente, sería interesante ver qué predicción realizaría el modelo para este paciente con visitas posteriores, ya que la pérdida de memoria episódica puede ser uno de los primeros síntomas potenciales de conversión a AD, incluso en casos en los que el paciente todavía mantiene su dependencia funcional.

Por último, cabe destacar la importancia de las features MRI, especialmente MidTemp, que mide el volumen del giro temporal medio, aunque también de otras que consistentemente tienen una importancia más moderada, como Entorhinal o Fusiform, que miden el volumen de la corteza entorrinal y el giro fusiforme, respectivamente. Estas estructuras cerebrales son algunas de las más afectadas por la enfermedad, de modo que volúmenes anormalmente bajos en algunas de estas zonas (o en todas ellas) en escáneres MRI pueden indicar un mayor riesgo de conversión.

Localmente MidTemp es una de las features más importantes, especialmente en las predicciones de RSF de que el paciente sMCI#1 (ver Figura 9a) no desarrolla AD, así como de que el paciente sMCI#2 (ver Figura 10a) sí lo hace. En esta predicción concreta contribuyen significativamente tanto MidTemp como Entorhinal y Fusiform, así como también FDG. De nuevo, sin contar con más visitas de este paciente no se puede saber con seguridad si habría experimentado el declive predicho por el modelo después del tercer año (ver Figura 13), aunque no parece algo irrazonable si se tiene en cuenta todo lo anterior.

5. Conclusiones

En este trabajo se ha demostrado la efectividad de métodos de análisis de supervivencia como Random Survival Forests y Gradient Boosting Survival Analysis en la predicción del tiempo de conversión de MCI a AD. Este enfoque surge de la necesidad de superar las limitaciones de los modelos tradicionales, que pueden ser menos eficaces en presencia de censura de datos. Asimismo, se ha comprobado la eficacia de ambos métodos trabajando con datos longitudinales, obteniendo resultados superiores a los obtenidos en baseline. Se puede concluir que se obtienen unas prestaciones más elevadas con RSF, así como unos resultados más consistentes al utilizar otras visitas en el entrenamiento o al introducir datos longitudinales.

No obstante, no se debe olvidar que los resultados obtenidos están sujetos a los hiperparámetros utilizados en cada modelo, que se han elegido manualmente buscando que el rendimiento de los modelos fuera equiparable. Por tanto, es posible que los resultados hubieran sido algo distntos al utilizar técnicas de optimización de hiperparámetros como *grid search*.

También se ha analizado la explicabilidad de ambos modelos, un aspecto crucial para aumentar la fiabilidad de métodos de análisis de supervivencia por su naturaleza de caja negra. De esta manera, se pueden explicar los factores que influyen en las predicciones de cada modelo y corroborar si estos son coherentes con el conocimiento clínico.

Es claro que ambos modelos presentan limitaciones, principalmente derivadas del conjunto de datos, como pueden ser el tamaño del mismo o la utilización de features con un porcentaje elevado de valores faltantes. En estas features, los valores imputados pueden ser menos confiables en comparación con otras features donde todos los valores son observados. Además, a pesar de que los modelos de supervivencia están diseñados para lidiar con datos censurados, es necesario recordar que en este trabajo también se han evaluado los modelos de manera complementaria utilizando un enfoque más similar al de un clasificador binario tradicional, y que el concepto de predicción correcta o incorrecta, tanto para el cálculo de algunas métricas como para algunos de los comentarios incluidos en el análisis de la explicabilidad local, se basa en este enfoque.

Asimismo, en cuanto al análisis de las predicciones individuales realizadas sobre pacientes determinados, es probable que realizar las mismas pruebas entrenando los modelos con datos de visitas más cercanas al evento de interés hubiera aportado información valiosa para una predicción más precisa de las probabilidades de supervivencia, especialmente en casos como el del paciente pMCI#1. Sin embargo, es preciso recordar que el diagnóstico de la enfermedad de Alzhéimer es una tarea compleja, ya que la progresión de la enfermedad puede ser muy diferente en cada paciente, afectando a distintas capacidades cognitivas.

Como trabajo futuro, sería interesante continuar experimentando con datos longitudinales en ambos modelos para comprobar si estos son capaces de encontrar patrones en la evolución real de los pacientes y así realizar mejores predicciones.

En este trabajo y en numerosos estudios recientes se está demostrando el potencial de la inteligencia artificial y los modelos de supervivencia en su aplicación a la práctica clínica. Estas herramientas, combinadas con métodos de explicabilidad como SHAP, pueden ser de gran utilidad en el ámbito clínico como herramienta de apoyo para los especialistas médicos en el diagnóstico no solo de Alzhéimer, sino también de otras enfermedades en las que sea vital un diagnóstico temprano. Asimismo, la integración de modelos de supervivencia con herramientas de explicabilidad puede mejorar la medicina personalizada al proporcionar al especialista médico información que puede servir para ajustar los tratamientos según las necesidades específicas de cada paciente.

Referencias

[1] World Health Organization, *Dementia*, 2024. dirección: https://www.who.int/news-room/fact-sheets/detail/dementia (visitado 28-05-2024).

[2] K. X. Dou, M. S. Tan, C. C. Tan, X. P. Cao, X. H. Hou, Q. H. Guo et al., "Comparative safety and effectiveness of cholinesterase inhibitors and memantine for Alzheimer's disease: a network meta-analysis of 41 randomized controlled trials," *Alzheimer's Research & Therapy*, vol. 10, n.º 1, pág. 126, 2018. DOI: 10.1186/s13195-018-0457-9.

[3] A. J. Mitchell y M. Shiri-Feshki, "Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies," *Acta Psychiatrica Scandinavica*, vol. 119, n.º 4, págs. 252-265, 2009. DOI: 10.1111/j.1600-0447.2008.01326.x.

[4] J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genetics*, vol. 45, n.º 12, págs. 1452-1458, 2013. DOI: 10.1038/ng.2802.

[5] S. Cure, K. Abrams, M. Belger, G. dell'Agnello y M. Happich, "Systematic literature review and meta-analysis of diagnostic test accuracy in Alzheimer's disease and other dementia using autopsy as standard of truth," *Journal of Alzheimer's Disease*, vol. 42, n.º 1, págs. 169-182, 2014. DOI: 10.3233/JAD-131559.

[6] H. Ahmed, H. Soliman, S. El-Sappagh, T. Abuhmed y M. Elmogy, "Early detection of Alzheimer's disease based on laplacian re-decomposition and XGBoosting," *Computer Systems Science & Engineering*, vol. 46, n.º 3, 2023. DOI: 10.32604/csse.2023.036371.

[7] E. E. Bron, M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," *NeuroImage*, vol. 111, págs. 562-579, 2015. DOI: https://doi.org/10.1016/j.neuroimage.2015.01.048.

[8] S. El-Sappagh, H. Saleh, F. Ali, E. Amer y T. Abuhmed, "Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time," *Neural Computing and Applications*, vol. 34, n.º 17, págs. 14 487-14 509, 2022. DOI: 10.1007/s00521-022-07263-9.

[9] A. Sarica, F. Aracri, M. G. Bianco, F. Arcuri, A. Quattrone, A. Quattrone et al., "Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer's disease," *Brain Informatics*, vol. 10, n.º 1, pág. 31, 2023. DOI: 10.1186/s40708-023-00211-w.

[10] D. J. Stekhoven y P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, n.º 1, págs. 112-118, 2012. DOI: 10.1093/bioinformatics/btr597.

[11] H. Ishwaran, U. B. Kogalur, E. H. Blackstone y M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, n.º 3, págs. 841-860, 2008. DOI: 10.1214/08-AOAS169.

[12] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, n.º 5, págs. 1189-1232, 2001. DOI: 10.1214/aos/1013203451.

[13] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, n.º 4, págs. 367-378, 2002. DOI: 10.1016/S0167-9473(01)00065-2.

[14] S. M. Lundberg y S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017. DOI: 10.48550/arXiv.1705.07874.

[15] S. Jeong, W. Jung, J. Sohn y H. I. Suk, "Deep geometrical learning for Alzheimer's disease progression modeling," en *2022 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2022, págs. 211-220. DOI: 10.1109/ICDM54844.2022.00031.

[16] J. Cai, W. Hu, J. Ma, A. Si, S. Chen, L. Gong et al., "Explainable machine learning with pairwise interactions for predicting conversion from mild cognitive impairment to Alzheimer's disease utilizing multi-modalities data," *Brain Sciences*, vol. 13, n.º 11, 2023. DOI: 10.3390/brainsci13111535.

[17] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng y B. T. T. Yeo, "Predicting Alzheimer's disease progression using deep recurrent neural networks," *NeuroImage*, vol. 222, pág. 117 203, 2020. DOI: 10.1016/j.neuroimage.2020.117203.

[18] E. L. Kaplan y P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, n.º 282, págs. 457-481, 1958. DOI: 10.1080/01621459.1958.10501452.

[19] M. C. Donohue, R. A. Sperling, D. P. Salmon, D. M. Rentz, R. Raman, R. G. Thomas et al., "The Preclinical Alzheimer Cognitive Composite: measuring amyloid-related decline," *JAMA Neurology*, vol. 71, n.º 8, págs. 961-970, 2014. DOI: 10.1001/jamaneurol.2014.803.

[20] M. C. Donohue, R. A. Sperling, R. Petersen, C. K. Sun, M. W. Weiner, P. S. Aisen et al., "Association between elevated brain amyloid and subsequent cognitive decline among cognitively normal persons," *JAMA*, vol. 317, n.º 22, págs. 2305-2316, 2017. DOI: 10.1001/jama.2017.6669.

[21] R. I. Pfeffer, T. T. Kurosaki, C. H. Harrah Jr., J. M. Chance y S. Filos, "Measurement of functional activities in older adults in the community," *Journal of Gerontology*, vol. 37, n.º 3, págs. 323-329, 1982. DOI: 10.1093/geronj/37.3.323.

[22] L. Mosconi, "Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 32, n.º 4, págs. 486-510, 2005. DOI: 10.1007/s00259-005-1762-7.

[23] Y. N. Ou, W. Xu, J. Q. Li, Y. Guo, M. Cui, K. L. Chen et al., "FDG-PET as an independent biomarker for Alzheimer's biological diagnosis: a longitudinal study," *Alzheimer's Research & Therapy*, vol. 11, n.º 1, pág. 57, 2019. DOI: 10.1186/s13195-019-0512-1.

[24] P. Battista, C. Salvatore e I. Castiglioni, "Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study," *Behavioural Neurology*, vol. 2017, n.º 1, pág. 1 850 909, 2017. DOI: 10.1155/2017/1850909.

[25] D. Tromp, A. Dufour, S. Lithfous, T. Pebayle y O. Després, "Episodic memory in normal aging and Alzheimer disease: insights from imaging and behavioral studies," *Ageing Research Reviews*, vol. 24, págs. 232-262, 2015. DOI: 10.1016/j.arr.2015.08.006.