Universidad Zaragoza
1542

Estudio de la reproducibilidad e interpretabilidad de los métodos más precisos del TADPOLE Challenge para el diagnóstico y pronóstico de la enfermedad de Alzheimer

Trabajo de Fin de Grado para la titulación de

Grado en Ingeniería Informática

en la Escuela de Ingeniería y Arquitectura

CURSO ACADÉMICO

2019-2020

Dirección

Mónica Hernández Giménez

Elvira Mayordomo Cámara

Francisco Ferraz García

NIA 737312

Departamento de Informática e Ingeniería de Sistemas

Resumen

El 15 de junio de 2017, el EuroPOND Consortium y ADNI lanzaron The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge. Este reto tiene como objetivo identificar qué personas empezarán a mostrar síntomas en un plazo de 1 a 5 años elegidas en un grupo de edad de riesgo de padecer la enfermedad. Para ello, se propuso utilizar un conjunto de mediciones longitudinales realizadas sobre los pacientes prevalecientes de ADNI, con el fin de realizar predicciones de aquellas mediciones futuras más relevantes para el diagnóstico de la enfermedad.

Las predicciones del TADPOLE Challenge se centran sobre el diagnóstico clínico, una clasificación en tres grupos dependientes del nivel de deterioro cerebral por probable enfermedad de Alzheimer (CN, MCI y AD); la puntuación ADAS-Cog13, resultante de un exámen psicológico frecuentemente utilizado en ensayos clínicos; y el volumen de los ventrículos del cerebro, estimado a partir de imágenes por resonancia magnética (MRI). Gracias al TADPOLE Challenge se han desarrollado una serie de métodos de aprendizaje automático que han proporcionado unos resultados muy precisos en dos de las tres mediciones propuestas: el diagnóstico clínico y el volumen de los ventrículos. Por el contrario, las mejores estimaciones de la puntuación ADAS-Cog13 fueron poco mejores que una estimación aleatoria.

El objetivo de este Trabajo de Final de Grado es reproducir los resultados de los tres mejores métodos del TADPOLE Challenge en la predicción del diagnóstico clínico, de la forma más fiel posible dada la escasa información disponible de los mismos. Además, utilizaremos métodos de Inteligencia Artificial Interpretable para comprender por qué estos algoritmos obtienen los mejores resultados en esta tarea, para obtener información relevante para su mejora, y para establecer su fiabilidad y plantear su posible uso en la práctica clínica. Adicionalmente, se probará la efectividad de los métodos desarrollados para el pronóstico de la puntuación de ADAS-Cog13 y la predicción del volumen de los ventrículos.

En particular, se han implementado dos sistemas: un Gradient Booster y un Random Forest, y se ha utilizado un sistema de Support Vector Machines diseñado por los autores para el reto. Mediante el aumento de los datos originales y la optimización de los hiperparámetros, se ha conseguido reproducir e incluso superar los resultados de los métodos ganadores del reto tanto en la predicción del diagnóstico clínico como del volúmen de los ventrículos, con una precisión del 96% y del 91%, respectivamente. En comparación con las métricas del reto, hemos obtenido un mAUC (área bajo la curva característica operativa del receptor) de 97.6 en el problema de diagnóstico, superando el mAUC de 93.1 obtenido por el método ganador, mientras que para el volúmen de los ventrículos obtenemos un MAE (error absoluto medio) de 0.27, superando el 0.45 de referencia.

Mediante el uso de dos algoritmos del estado del arte en interpretabilidad (SHAP y LIME) se ha demostrado la fiabilidad de los modelos, comparando los atributos que usan para obtener el diagnóstico con los utilizados en la práctica clínica, y se han señalado los motivos por los que los sistemas podrían fallar, proponiendo soluciones para aumentar la capacidad de generalización de los modelos.

Índice de Contenido

1 Introducción

1.1 Motivación y Contexto

La enfermedad de Alzheimer se caracteriza por la muerte de las células nerviosas en ciertas áreas del cerebro, más concretamente en el lóbulo frontal, la corteza entorrinal y el hipocampo. Entre otros efectos, se observa un deterioro de la producción de acetilcolina. Cuando la enfermedad se presenta en su fase avanzada, pueden observarse bajo microscopio placas de ciertos fragmentos de la proteína amiloide sobre los tejidos. Esto lleva en el corto plazo a la pérdida de memoria, la capacidad de habla, de pensamiento o de toma de decisiones. En el largo plazo, puede suponer la pérdida de funciones biológicas esenciales, llevando incluso a la muerte.

En el año 2015 se realizó la estimación de que existían en el mundo un total de 46,8 millones de casos de Alzheimer (Alzheimer's Disease, AD), y a día de hoy se cree que este número puede haber superado los 50 millones. Según *Alzheimer's Disease International*, esta cifra puede llegar a duplicarse cada 20 años, llegando a los 75 millones en 2030. Siendo ésta la forma más común de demencia (entre el 50% y 75%), la enfermedad de Alzheimer está siendo estudiada desde una gran variedad de disciplinas, pero todavía no se conocen ni las causas concretas que la producen ni los medios para curarla.

Existen distintos tratamientos que pueden paliar parcialmente los síntomas de la enfermedad, como los inhibidores de colinesterasa [1]. Se ha demostrado que dichos tratamientos son más efectivos en los estadios tempranos de la enfermedad. Por ello, es esencial poseer herramientas de ayuda al diagnóstico temprano y de predicción de la evolución de los diferentes factores de la enfermedad en la población de riesgo para mejorar tanto como sea posible su calidad de vida, tanto en su estado cognitivo como en su degeneración anatómica.

El diagnóstico de AD nunca es concluyente sobre un paciente en vida. El único método de diagnóstico fiable consiste en realizar una biopsia post-mortem. Por tanto, los diagnósticos realizados en la práctica clínica, apoyados por toda la tecnología disponible, son tan sólo de probable AD. Los indicios de la presencia de la enfermedad en una persona se obtienen mediante sintomatologías de diversos tipos. Llamamos a estas sintomatologías biomarcadores. Aunque ningún biomarcador puede por sí mismo facilitar el diagnóstico de la enfermedad, existe evidencia de que la combinación de varios de ellos puede ayudar enormemente en la realización de un diagnóstico más preciso.

Existen dos clases generales de biomarcadores: mediciones de la proteína beta-amiloide y mediciones del daño a las células nerviosas. Las primeras pueden obtenerse usando el Líquido Cefalorraquídeo (CerebroSpinal Fluid, CSF) o una Tomografía por Emisión de Positrones (Positron Emission Tomography, PET). La segunda categoría se puede medir mediante la cuantificación de la

fracción de proteína tau en el CSF, mediante un escáner tau-PET, midiendo el metabolismo cerebral mediante fluorodesoxiglucosa (fluoro-deoxyglucose, FDG) o calculando el deterioro cerebral mediante Imagen por Resonancia Magnética (Magnetic Resonance Imaging, MRI). Todos estos marcadores han demostrado ser efecivos para la predicción y diagnóstico de la probable AD [2]. Del mismo modo, los tests cognitivos como el Mini-Mental State Examination (MMSE) o el Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) también presentan evidencia de su elevada capacidad de diagnóstico [3, 4]. Además, se puede tener en cuenta en el diagnóstico el factor genético mediante el análisis del gen APOE4, para el que existen combinaciones de nucleótidos que aumentan o reducen el riesgo de padecer la enfermedad, u otros genes como el TOMM40.

La iniciativa Alzheimer's Disease Neuroimaging Initiative (ADNI) fue diseñada en 2004 como un estudio multi-centro para el desarrollo de biomarcadores de todo tipo para la detección temprana y el seguimiento de la enfermedad de Alzheimer. En sus 16 años de historia, esta iniciativa ha pasado por 4 fases sobre un grupo de unos 800 pacientes: 200 clínicamente sanos (clinically normal, CN), 400 con deterioro cognitivo leve (mild cognitive impairment, MCI) y 200 enfermos de AD. De esta primera cohorte, algunos pacientes no se incluyeron en los subsecuentes estudios por diferentes motivos, y algunos pacientes nuevos fueron añadidos. Las diferentes fases de ADNI son las siguientes:

1. ADNI-1, con el objetivo de desarrollar biomarcadores como resultado de medidas en estudios clínicos.
2. ADNI-GO, con el objetivo de analizar los biomarcadores en etapas tempranas de la enfermedad.
3. ADNI-2, con el objetivo de producir biomarcadores como predictores del deterioro cognitivo y como medidas resultantes.
4. ADNI-3, con el objetivo de estudiar el uso de tau-PET y técnicas de imagen funcional en ensayos clínicos.

Actualmente la base de datos de ADNI la componen 1920 pacientes, con distinto número de visitas por individuo.

Gracias a los esfuerzos desarrollados por ADNI, se ha generado una base de datos extensiva con una gran cantidad de biomarcadores. Esto ha permitido que en los últimos años una gran cantidad de estudios científicos hayan utilizado esta información para el desarrollo de métodos de diagnóstico y predicción de la evolución de la enfermedad, así como para realizar comparativas de los resultados del estudio sobre ciertos marcadores con la literatura disponible hasta el momento. En concreto, se recogen en su página web un total de 1800 publicaciones.

Dadas las evidentes bondades de esta iniciativa, el grupo EuroPOND Consortium, un equipo de científicos dedicado al desarrollo de técnicas de ciencia de datos para la neurología clínica y la neurociencia computacional, creó en junio de 2017 The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge, o Reto por la Predicción de la Evolución Longitudinal de la Enfermedad de Alzheimer. Este es un reto abierto con premios de hasta £5.000 a cualquier grupo universitario, empresarial, colegial o independiente que pudiese superar al resto de participantes en la predicción de la evolución de AD según tres métricas distintas: el diagnóstico clínico, la puntuación ADAS-Cog13 y el volumen de los ventrículos sobre los pacientes de la base de datos de ADNI, utilizando tanto métodos clínicos como computacionales. Los resultados del reto fueron publicados en la página web de TADPOLE (https://tadpole.grand-challenge.org/) en julio de 2019, y posteriormente un artículo escrito por los directores fue publicado analizando de manera general los resultados del mismo [5].

Con la aparición de esta iniciativa, ha surgido la oportunidad de realizar una investigación en profundidad sobre los métodos que han proporcionado los mejores resultados, con objeto de descubrir los motivos que los hacen destacar sobre el resto. El artículo hace mediciones extensivas sobre el total de los métodos basándose en una serie de formularios rellenados por los participantes a la hora de presentar sus resultados, pero al no disponer ni de los algoritmos ni de herramientas para poder analizar de forma genérica todos ellos, no es posible determinar qué atributos o marcadores concretos son los determinantes de cara al diagnóstico de un paciente de forma certera. Este no es un problema exclusivo de TADPOLE, sino que la interpretabilidad de los métodos de inteligencia artificial es materia de estudio constante que está empezando a dar sus primeros frutos, con una serie de métodos novedosos que permiten obtener tanto explicaciones locales como generales de métodos predictivos que en otras circunstancias se hubieran considerado "cajas negras".

Creemos que utilizar estos métodos de interpretabilidad sobre los modelos de TADPOLE puede proporcionar información relevante sobre la importancia de los biomarcadores de la base de datos de ADNI, además de permitir una justificación lógica desde el punto de vista clínico para establecer el grado de confianza en el diagnóstico o pronóstico realizado por un modelo de aprendizaje automático. Además, los resultados pueden generalizarse a otros ámbitos de la inteligencia artificial con aplicación a ciencias de la salud y avanzar hacia la creación de herramientas profesionales que los utilicen.

1.2 Objetivos

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

1. Revisar el estado del arte en el diagnóstico y la predicción de la enfermedad de Alzheimer.
2. Estudiar los objetivos del TADPOLE challenge.
3. Recopilar los datos de ADNI necesarios para la realización del estudio.
4. Implementar los métodos más precisos de TADPOLE challenge en cada categoría.
5. Estudiar la reproducibilidad de los resultados reportados en la web y la publicación del challenge.
6. Análizar los resultados obtenidos y realizar una comparativa con otros del estado del arte utilizando diferentes datos o técnicas.
7. Estudiar el problema de la interpretabilidad de los métodos para la mejora del análisis de los resultados.

Con el desarrollo de este trabajo se han conseguido implementar dos algoritmos equivalentes a los dos mejores del TADPOLE Challenge para la predicción del diagnóstico clínico y el volumen de los ventrículos. Además se ha realizado el estudio de su interpretabilidad para poder demostrar su fiabilidad desde un punto de vista clínico y proponer mejoras al sistema. Complementariamente, ha sido posible ejecutar el tercer mejor método de TADPOLE gracias a la disponibilidad de los códigos, aunque no se ha podido realizar el estudio de interpretabilidad debido a problemas de incompatibilidad entre las librerías utilizadas.

1.3 Estructura del Documento

En el primer capítulo de esta memoria se presenta el contexto de este TFG, se enumeran los objetivos planteados y se explica la estructura del documento.

En el segundo capítulo de la memoria se hace un análisis de TADPOLE Challenge, los datos utilizados y los mejores métodos presentados.

En el tercer capítulo, se hace una introducción teórica a los modelos de aprendizaje automático implementados, y se analiza el método de generación de modelos y datasets utilizado.

En el cuarto capítulo, se realiza una introducción teórica a los métodos de interpretabilidad de modelos de caja negra seleccionados para este TFG.

En el quinto capítulo, se presentan los resultados principales de los experimentos realizados, se analizan los resultados en reproducibilidad y se genera una valoración de los modelos mediante interpretabilidad.

En el sexto capítulo, se desarrollan las conclusiones más relevantes nacidas de este trabajo y se establecen las líneas de trabajo futuro.

Finalmente en los anexos se incluyen diversos análisis y comentarios adicionales que se han realizado, de menor relevancia y extensión excesiva para los objetivos de este trabajo.

2 TADPOLE Challenge

2.1 Introducción y Objetivos

El objetivo de TADPOLE Challenge es el siguiente: Dado un conjunto de datos extraído de las poblaciones de ADNI 1, 2 y GO realizar al menos una de las siguientes tres predicciones utilizando los datos futuros procedentes de ADNI-3.

1. Estado Clínico (o Diagnóstico). Clasificar al individuo en Cognitivamente Normal (Cognitively Normal, CN), con Deterioro Cognitivo Leve (Mild Cognitive Impairment, MCI) o enfermo de Alzheimer (Alzheimer's Disease, AD).
2. ADAS-Cog13. Predecir la puntuación obtenida por el paciente en este test cognitivo.
3. Volumen de los Ventrículos (VV). Predecir el volumen de esta estructura cerebral dividido por volumen intracraneal del mismo modo que se obtiene para ADNI, mediante la herramienta FreeSurfer.

El momento en el que los datos de ADNI-3 son adquiridos para cada individuo no es conocido, por lo que se presentarán predicciones de los individuos para cada mes dentro del período establecido y en el momento de la evaluación de resultados se contará sólo con las predicciones del momento coincidente. Las predicciones se presentan como estimaciones de probabilidad para cada una de las tres clases de diagnóstico (CN, MCI y AD) y se presentará un intervalo de confianza para las predicciones de ADAS-Cog y VV.

De inicio a fin, el reto tardó 6 meses en completarse. En este tiempo, los participantes pudieron generar sus estimaciones con cualquier algoritmo predictivo, red neuronal o método manual. Durante el desarrollo de los algoritmos utilizados en el reto, los datos sobre ADNI-3 no eran públicos, constituyendo este un verdadero test-set para probar los modelos.

2.2 Datos Utilizados

Los datos de TADPOLE, resumidos en la Figura 1, proceden de la base de datos de ADNI, y están compuestos por individuos prevalecientes en todas sus fases para los que se requieren las predicciones. Los participantes tienen la libertad de modificar los datos o ampliarlos del modo que prefieran con el objetivo de mejorar los resultados. Cualquier conocimiento o material adicional puede ser añadido a los mismos. En cualquier caso, se desarrollaron tres conjuntos de datos estándar para el reto:

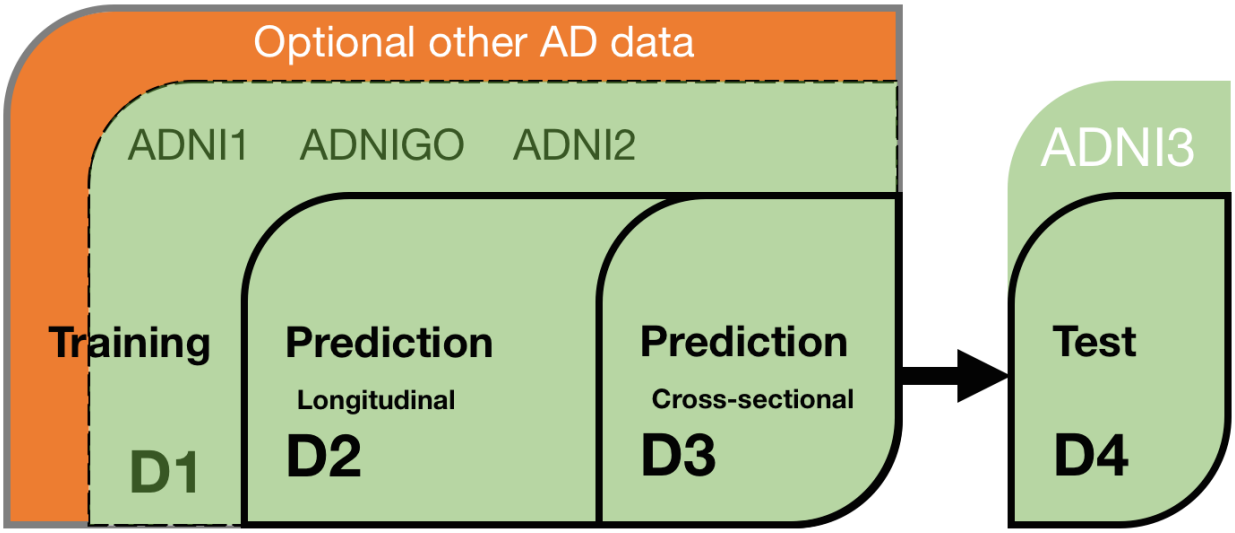


Figure 1: Resumen visual de los datos estándar de TADPOLE

1. $D1$: Datos longitudinales extensivos para el entrenamiento.
2. $D2$: Datos longitudinales extensivos para la validación de predicciones.
3. $D3$: Datos limitados para la validación de predicciones. Se trata de un subconjunto de atributos en los mismos pacientes de $D2$.

Tras la finalización del reto se presentó $D4$, el conjunto de datos de ADNI-3 que los modelos entrenados debían intentar predecir, es decir, el test-set o conjunto de evaluación. Todos estos datos están disponibles en la base de datos de ADNI para su descarga.

$D1$ contiene 1667 pacientes y un total de 8841 visitas, mientras que $D2$ contiene 896 pacientes y 5177 visitas, donde 883 pacientes prevalecen desde $D1$ y sólo 13 son nuevas incorporaciones. $D3$ tiene los mismos participantes que $D2$, pero un conjunto de atributos más limitado de los mismos, a una visita por paciente. En concreto, $D1$ y $D2$ presentan 1886 atributos por visita, mientras que $D3$ presenta sólo 382. Esta diferencia en la cantidad de atributos se debe a que con $D3$ se pretende proporcionar una cantidad de información similar a la disponible usualmente a la hora de hacer un diagnóstico inicial en la visita de baseline, separando así los datos de entrenamiento en ”extensivos” o ”longitudinales”. Otros aspectos de la demografía de los participantes del reto pueden observarse en la Figura 2. En el Anexo I: TADPOLE Challenge puede encontrarse una explicación del significado de los atributos más relevantes para el estudio realizado en este TFG.

2.3 Mejores Métodos

Aunque algunos participantes del reto propusieron métodos basados en los utilizados en la práctica clínica, métodos estadísticos o modelos de progresión de la enfermedad, los tres mejores métodos

Demographics

		D1	D2	D3	D4
Overall number of subjects		1667	896	896	219
Controls†	Number (% all subjects)	508 (30.5%)	369 (41.2%)	299 (33.4%)	94 (42.9%)
	Visits per subject	8.3 ± 4.5	8.5 ± 4.9	1.0 ± 0.0	1.0 ± 0.2
	Age	74.3 ± 5.8	73.6 ± 5.7	72.3 ± 6.2	78.4 ± 7.0
	Gender (% male)	48.6%	47.2%	43.5%	47.9%
	MMSE	29.1 ± 1.1	29.0 ± 1.2	28.9 ± 1.4	29.1 ± 1.1
	Converters*	18	9	-	-
MCI†	Number (% all subjects)	841 (50.4%)	458 (51.1%)	269 (30.0%)	90 (41.1%)
	Visits per subject	8.2 ± 3.7	9.1 ± 3.6	1.0 ± 0.0	1.1 ± 0.3
	Age	73.0 ± 7.5	71.6 ± 7.2	71.9 ± 7.1	79.4 ± 7.0
	Gender (% male)	59.3%	56.3%	58.0%	64.4%
	MMSE	27.6 ± 1.8	28.0 ± 1.7	27.6 ± 2.2	28.1 ± 2.1
	Converters*	117	37	-	9
AD†	Number (% all subjects)	318 (19.1%)	69 (7.7%)	136 (15.2%)	29 (13.2%)
	Visits per subject	4.9 ± 1.6	5.2 ± 2.6	1.0 ± 0.0	1.1 ± 0.3
	Age	74.8 ± 7.7	75.1 ± 8.4	72.8 ± 7.1	82.2 ± 7.6
	Gender (% male)	55.3%	68.1%	55.9%	51.7%
	MMSE	23.3 ± 2.0	23.1 ± 2.0	20.5 ± 5.9	19.4 ± 7.2
	Converters*	-	-	-	9

Figure 2: Resumen de la demografía de los participantes de TADPOLE y su distribución en los datasets estándar.

presentados al reto respecto de su clasificación en diagnóstico clínico han resultado ser métodos tradicionales de aprendizaje automático supervisado. Se presentan a continuación los nombres de los equipos y las descripciones generales de los algoritmos:

1. **Frog**: Keli Liu, Christina Rabe, Paul Manser. Institución: Genentech, USA.
   SELECCIÓN DE ATRIBUTOS: Automática
   ATRIBUTOS SELECCIONADOS: Además de los presentes en los datos, se realizaron los siguientes aumentos: medida más reciente, tiempo desde la medida más reciente, máximo y mínimo histórico, tiempo desde el máximo y mínimo histórico, cambio más reciente en la medida y tiempo desde el cambio más reciente.
   DATOS FALTANTES: El paquete los gestiona automáticamente.
   MÉTODO DE PREDICCIÓN: Modelos y atributos flexibles seleccionados mediante una máquina de Gradient Boosting, utilizando el paquete XGBoost. Se entrenaron distintos modelos para las siguientes ventanas temporales desde la última visita del paciente: 0-8 meses, 9-15,16-27, 28-39, 40-60, >60.
   RESULTADOS: Puesto = 1, mAUC = 0.931, BCA = 0.849.

2. **Threedays**: Paul Moore, Terry J. Lyons, John Gallacher, Institución: Mathematical Institute, University of Oxford, Department of Psychiatry, University of Oxford, UK.
   SELECCIÓN DE ATRIBUTOS: Manual
   ATRIBUTOS SELECCIONADOS: Edad, meses desde el baseline, género, raza, estado civil,

diagnóstico, tests cognitivos (MMSE, CDRSB, ADAS11, ADAS-Cog13, RAVLT, FAQ) y estado APOE.

DATOS FALTANTES: El paquete los gestiona automáticamente.

MÉTODO DE PREDICCIÓN: Se entrenan dos modelos de Random Forest, uno que parte de un diagnóstico CN y otro que parte de un diagnóstico MCI. Los datos son ordenados por fecha de obtención, asociando los atributos de cada paciente con el horizonte temporal. *Nota: El mismo equipo publicó posteriormente un artículo [6] en el que se seguía la misma metodología pero entrenando un sólo Random Forest, que por simplicidad se toma como referencia para este TFG*.

RESULTADOS: Puesto = 2, mAUC = 0.921, BCA = 0.823.

3. **EMC-EB**: Esther E. Bron, Vikram Venkatraghavan, Stefan Klein, Institución: Erasmus MC, Países Bajos.

   SELECCIÓN DE ATRIBUTOS: Automática - se seleccionaron los atributos con mayores cambios a lo largo del tiempo para pacientes que desarrollaron AD.

   ATRIBUTOS SELECCIONADOS: diagnóstico clínico, tests cognitivos, MRI (Freesurfer cross-sectional), FDG PET, medidas de DTI (FA, MD, RD, AD) y atributos de CSF.

   DATOS FALTANTES: Inserción utilizando interpolación de vecino más cercano basada en las visitas más recientes del paciente. En caso de que no fuera posible, inserción del valor medio de los datos de entrenamiento. Se excluyen de los datos de entrenamiento las visitas sin diagnóstico.

   MÉTODO DE PREDICCIÓN: Support Vector Machine (SVM). Los autores utilizaron un kernel de SVM con función de base radial (Radial Basis Function, RBF), cuyo C-parámetro se estableció a 0.5 y gamma al recíproco del número de atributos. Todos los atributos se normalizaron a media cero y desviación estándar unitaria.

   RESULTADOS: Puesto = 3, mAUC = 0.907, BCA = 0.805.

Como puede observarse, la descripción de los métodos es bastante sucinta y no se dispone de los códigos utilizados para la reproducción de los resultados. Durante el desarrollo de este TFG, ha surgido la iniciativa TADPOLE-Share (https://tadpole-share.github.io/), una plataforma para compartir algoritmos de predicción de Alzheimer con TADPOLE Challenge como base. Este proyecto se encuentra todavía en una fase inicial, y tan sólo ofrece una versión funcional del algoritmo EMC-EB. Para este TFG no se ha implementado una versión propia de este algoritmo, sino que se ha accedido a la versión disponible en GitHub.

3 Métodos Desarrollados

En el presente capítulo se detallan los algoritmos de aprendizaje automático implementados en este TFG, junto con la metodología seguida para el entrenamiento de los mismos: preprocesado de datos, selección de hiper-parámetros, división en datos de entrenamiento y validación, etc.

3.1 Aprendizaje Supervisado

Los modelos implementados pertenecen a la clase de algoritmos de aprendizaje supervisado, donde usamos datos de entrenamiento con múltiples atributos x_i para predecir una variable objetivo y_i.

3.1.1 Modelo y Parámetros

El modelo se refiere a la función que permite calcular la predicción de y_i a partir de la entrada x_i. Un ejemplo común sería un modelo lineal, en el que la predicción se da como una combinación lineal de los datos de entrada con pesos

$$\hat{y}_i = \sum_j w_j x_{ij}. \tag{1}$$

Los parámetros son la parte indeterminada que debemos aprender de los datos. En una regresión lineal son los pesos, w.

3.1.2 Función Objetivo: Pérdida y Regularización

Para entrenar un modelo es necesario definir una función objetivo que mida cómo de bien se acomoda la predicción a los datos de entrenamiento. Usualmente, una función objetivo consiste de dos partes: la función de pérdida y el término de regularización

$$obj(\theta) = L(\theta) + \Omega(\theta), \tag{2}$$

donde L es la función de pérdida de entrenamiento y Ω es el término de regularización. Se considera un buen modelo aquel que genera un buen equilibrio entre su sesgo y su varianza, de tal modo que el modelo es sencillo y generaliza correctamente. Por ello, la regularización es esencial para prevenir el subajuste o sobreajuste.

3.2 Gradient Boosting

El método de Gradient Boosting (Potenciación del Gradiente, GB) es una técnica de aprendizaje supervisado basada en la agrupación de árboles de decisión. El modelo de agrupación de árboles se basa en un conjunto de Árboles de Clasificación y Regresión (Classification and Regression Trees, CART), donde el objetivo es que cada árbol haga una clasificación en función de un atributo de los datos (en nuestro caso, visitas de pacientes) en distintas hojas, a las cuales se les asigna una puntuación según una función de optimización. Como un árbol no suele ser suficiente, se suma la predicción de varios árboles distintos al mismo tiempo, es decir, se agrupan los árboles. El valor de la predicción $\hat{y}_i$ se define como

$$\hat{y}_i = \sum_{i=1}^{K} f_k(x_i), f_k \in F, \tag{3}$$

donde K es el número de árboles, y cada f_k es una función en el espacio de funciones F formado por el conjunto de posibles CARTs. La función objetivo a optimizar en GB se define como

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

donde n es el número de muestras de los datos de entrenamiento, $l(y_i, \hat{y}_i)$ es la función de pérdida de entrenamiento y Ω es el término de regularización que controla la complejidad de los árboles. Cabe destacar que este modelo es válido tanto para Gradient Boosting como para Random Forests. La diferencia entre ambos métodos surge del modo en el que estos son entrenados.

En el caso de GB, al ser inviable calcular el gradiente de todos los árboles al mismo tiempo, se utiliza una estrategia aditiva consistente en arreglar lo aprendido y añadir un árbol nuevo iterativamente. Así, la función a optimizar viene dada por

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \Omega(f_k), \tag{5}$$

donde el valor de predicción en cada paso t es $\hat{y_i}^{(t)}$. Las funciones f_i se obtienen de manera recursiva de la siguiente manera

$$\hat{y_i}^{(0)} = 0 \tag{6}$$

$$\hat{y_i}^{(1)} = f_1(x_i) = \hat{y_i}^{(0)} + f_1(x_i) \tag{7}$$

$$\hat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y_i}^{(1)} + f_2(x_i) \tag{8}$$

...

$$\hat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y_i}^{(t-1)} + f_t(x_i). \tag{9}$$

Así, en cada paso, añadimos aquel árbol que optimiza nuestro objetivo

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + const \tag{10}$$

donde la constante agrupa el valor del regularizador en las f_i anteriores. La función de pérdida utilizada es el promedio del logaritmo (Mean Log Loss).

Para su cálculo se utiliza una Expansión de Taylor de la función de pérdida de segundo orden

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + const, \tag{11}$$

donde g_i y h_i se definen como

$$g_i = \partial_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)}) \tag{12}$$

$$h_i = \partial^2_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)}). \tag{13}$$

Simplificando las constantes, el objetivo en el paso t resulta

$$obj^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t). \tag{14}$$

Por último, para calcular el término de regularización necesitamos definir la complejidad del árbol $\Omega(f)$. Para ello, redefinimos $f(x)$ como

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \to \{1, 2, ..., T\} \tag{15}$$

donde w es el vector de puntuaciones en las hojas, q es una función que asigna a cada dato su hoja correspondiente y T es el número de hojas. Con ello, se define la complejidad como

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2. \tag{16}$$

Unificando el desarrollo realizado para la función de pérdida y el regularizador obtenemos

$$obj^{(t)} = \sum_{i=1}^{n} [g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2. \tag{17}$$

Así, la función objetivo puede explicarse de manera muy simplificada como que la ganancia resultante de intentar dividir una hoja en dos hojas nuevas es

$$\frac{1}{2}[(LeftLeafScore) + (RightLeafScore) + (OriginalLeafScore)] - \gamma. \tag{18}$$

De este modo, solo generamos dos hojas nuevas si el resultado es negativo, es decir, si la nueva

puntuación es menor que γ. Este procedimiento se conoce como poda.

Para el desarrollo de este TFG se ha utilizado el mismo paquete que utilizó el equipo Frog, XGBoost [7], una máquina de GB que obtiene resultados de estado del arte en multitud de problemas, tanto de clasificación como de regresión. En concreto, se ha utilizado una implementación disponible en Python de la misma (https://xgboost.readthedocs.io/).

3.3 Random Forest

Al igual que Gradient Boosting, Random Forest (o Bosque Aleatorio, RF) es una técnica de aprendizaje supervisado basada en la agrupación de árboles de decisión. La diferencia entre GB y RF se encuentra en cómo se construye esta agrupación. Como hemos explicado, ambos modelos tienen una base común en la descripción del modelo con la misma función objetivo

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k). \tag{19}$$

Al contrario que GB, RF no define una función de pérdida a optimizar, simplemente define una función de evaluación del modelo. Los árboles irán siendo agregados proceduralmente: no se elegirán con base en una selección razonada, sino al buen rendimiento casual de un clasificador generado aleatoriamente. El procedimiento de agregación de los árboles es similar al método de Bagging (embolsado), que se describe a continuación.

Dado un número B de iteraciones, se selecciona un framgento aleatorio con reemplazo (es decir, un mismo dato puede aparecer varias veces) de los datos de entrenamiento y se entrena un árbol para ese fragmento. Al final, se hace una media de las predicciones para regresión, o una votación por mayoría para clasificación, tal que

$$\hat{y_i} = \frac{1}{B} \sum_{b=1}^{B} f_b(x). \tag{20}$$

En el caso de RF, la diferencia con Bagging reside en que, para cada división en el proceso de entrenamiento, un subconjunto aleatorio de atributos, normalmente de tamaño $\sqrt{p}$ siendo p el número de atributos para un problema de clasificación, es elegido del total, reduciendo la dimensionalidad conforme aumenta la profundidad. Esto se debe a que, en caso de que un atributo sea un buen predictor, será elegido en muchos de los B árboles, causando correlación entre ellos. El desarrollo completo de la materia puede encontrarse en su artículo original [8].

Como el equipo Threedays no presentó una descripción detallada del método utilizado, se decidió utilizar el método de RF disponible en la librería de Python Scikit-Learn (https://scikit-learn.org/stable/modules/generated/

sklearn.ensemble.RandomForestClassifier.html), una de las librerías más estandarizadas sobre aprendizaje automático que permite mucha compatibilidad con paquetes externos, siendo esto interesante de cara a la interpretabilidad del mismo.

3.4 Support Vector Machines

Support Vector Machines (máquinas de soporte de vectores, SVM) son métodos de aprendizaje supervisado que utilizan un algoritmo de aumento de la dimensionalidad de cada dato (o punto p-dimensional) para poder encontrar un hiperplano óptimo que separe de forma óptima los puntos de una clase de la otra, es decir, que permita un maximizar el margen entre ambas clases. Por tanto, funcionan hasta cierto punto como clasificadores multi-lineales muy parecidos a las redes neuronales simples. Se conoce como support vectors al conjunto de puntos más cercanos al hiperplano de separación.

Para este TFG no ha sido necesario implementar un sistema basado en SVM, pues ha sido posible acceder a la versión original presentada por EMC-EB gracias a la iniciativa TADPOLE-Share, por lo que se estudiarán directamente sus resultados.

3.5 Generación de los Modelos

En el proceso de entrenamiento de los modelos realizado en este TFG, se ha buscado un balance de tres aspectos esenciales:

1. Hacer que el proceso de entrenamiento resulte sencillo y que no requiera de muchas transformaciones sobre el dataset original.
2. Utilizar las conclusiones más importantes de nuestro análisis de los resultados de TADPOLE, centrándonos en la escasa información proporcionada por los equipos ganadores, para tratar de utilizar los mismos datos de entrenamiento en ambos modelos, y poder obtener resultados comparables.
3. Obtener resultados con la máxima precisión posible, intentando igualar o superar a los métodos de referencia.

Para ello se han creado dos datasets reducidos, $D1\_RedD3$ (los pacientes de $D1$ con los atributos de $D3$) y $D1D2\_RedD4$ (los pacientes de $D1$ y $D2$ con los atributos de $D4$); y dos datasets aumentados, $D1D2\_Aug$ y $D4\_Aug$. Estos conjuntos de datos utilizan los atributos disponibles para $D4$, aumentados mediante el método utilizado en Frog (medida más reciente, tiempo desde la medida más reciente, máximo y mínimo histórico, tiempo desde el máximo y

mínimo histórico, cambio más reciente en la medida tiempo desde el cambio más reciente). Cabe destacar que la selección de atributos es realizada automáticamente por los modelos.

El procedimiento seguido para la generación de los modelos es pues el siguiente:

1. Obtener los datasets originales ($D1$, $D2$, $D3$ y $D4$).
2. Corregir los datos manualmente para que sean legibles por los modelos. Esta corrección ha consistido en convertir todas las clasificaciones, fechas y comentarios posibles a formato numérico. Por ejemplo, respecto del atributo 'Sexo' = [Hombre, Mujer] se obtiene 'Sexo' = [0, 1]; del atributo 'Fecha' = [January 8th 2015, ...] se obtiene 'Fecha' = [08012015, ...]; etc.
3. Rellenar los datos faltantes con -1, dado que ni Random Forest ni los algoritmos de interpretabilidad los aceptan.
4. Generar 4 modelos por cada clasificador resultantes de: (1) entrenar los modelos con $D1$ para predecir $D2$; (2) entrenar los modelos con $D1\_RedD3$ para predecir $D3$; (3) entrenar los modelos con $D1D2\_RedD4$ para predecir $D4$; (4) entrenar los modelos con $D1D2\_Aug$ para predecir $D4\_Aug$.
5. Realizar una selección óptima de los hiperparámetros de los modelos mediante un método de búsqueda automática. En este caso, se ha utilizado la función GridSearchCV de la librería Scikit-learn. Esta función hace una búsqueda exhaustiva de la mejor combinación de los parámetros explicitados para un estimador mediante un método de validación cruzada (k-fold cross validation, k=10). La lista de hiperparámetros proporcionada para cada modelo se detalla en el Anexo III: Análisis detallado de reproducibilidad.
6. Calcular las métricas utilizadas en el reto y realizar una comparativa con los resultados de los métodos de referencia.

El procedimiento descrito tiene como objetivo replicar de la forma más completa y precisa posible el entorno experimental con el que trabajaron los participantes del TADPOLE Challenge, tratando de respetar las normas del mismo.

4 Intepretabilidad de Modelos de Caja Negra

En su libro *Intepretable Machine Learning* [9], Christoph Molnar introduce la noción de interpretabilidad del siguiente modo:

> *"No hay una definición matemática para la interpretabildad. [...] La intepretabilidad es el grado en el que un humano puede entender la causa de una predicción. [...] La intepretabilidad es el grado en el que un humano puede predecir los resultados de un modelo de manera consistente. Cuanto mayor sea la interpetabilidad de un modelo, más fácil será para una persona entender por qué se han realizado ciertas decisiones."*

Como norma general, los métodos de Inteligencia Artificial Interpretable (Explainable Artificial Intelligence, XAI) se suelen dividir en dos tipos: dependientes del modelo y agnósticos al modelo. Mediante los primeros podemos conseguir ciertas intuiciones muy específicas que pueden ser útiles en el estudio profundo de un modelo concreto. Mediante las segundas, podemos analizar una gran colección de modelos en los mismos términos, siendo una forma útil de comparar su fiabilidad. En el presente capítulo, presentaremos dos métodos de estado del arte en intepretabilidad agnóstica al modelo y, en el capítulo siguiente, los utilizaremos para analizar nuestros modelos. Las explicaciones siguientes son adaptaciones de las presentadas en [9].

4.1 SHAP

SHAP (SHapley Additive exPlanations, [10]) es una aproximación mediante teoría de juegos a la explicación de la salida de cualquier modelo de aprendizaje automático. Conecta la asignación óptima de crédito con explicaciones locales utilizando los valores clásicos de Shapley de teoría de juegos y sus extensiones relacionadas. A partir de ella se ha desarrollado un método de explicación de modelos de árbol, TreeExplainer [11], que permite la interpretabilidad de nuestros modelos de GB y RF de una manera eficiente y precisa.

Los valores de Shapley son un método de distribución de riqueza en teoría de juegos tradicional. La idea es la siguiente: una serie de jugadores cooperan y obtienen una cierta ganancia total por su cooperación. Dado que cada jugador puede haber tenido una participación distinta en la obtención de la ganancia, ¿cómo se reparte la misma de una forma justa entre los jugadores? ¿qué beneficio debe esperar cada jugador? Una posible respuesta es el valor de Shapley definido a partir de la expresión

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S)) \tag{21}$$

donde S es un subconjunto coalicional del total de jugadores, x es el vector de jugadores que participan en la jugada, y p el número de jugadores total. La función val define la ganancia generada por una coalición. La fórmula nos da el valor que un jugador j contribuye durante el juego, con pesos y sumada sobre todas las posibles combinaciones de coaliciones.

Para un modelo de aprendizaje, podemos utilizar las predicciones del modelo como ganancia y los atributos como jugadores. De este modo, obtenemos de qué manera ha contribuido cada jugador-atributo positiva o negativamente a cada predicción.

El problema de los valores de Shapley es que su tiempo computacional es exponencial (sobre 2^k posibles coaliciones) en el número de atributos, pues hay que probar todas las posibles combinaciones, por lo que normalmente sólo se calcula un valor aproximado. Gracias a Shap y su algoritmo TreeExplainer, es posible computar una explicación mediante los valores de Shapley que sólo tiene un coste temporal de $O(AHP^2)$, siendo A el número de árboles, H el número máximo de hojas en un árbol y P la profundidad máxima en un árbol.

SHAP ofrece la herramienta KernelSHAP para clasificadores que no utilizan estructuras de árbol, pero esta presenta una incompatibilidad con el modelo SVM de EMC-EB.

4.2 LIME

Local Interpretable Model-Agnostic Explanations (Explicaciones agnósticas al modelo interpretables localmente, LIME [12]) es un modelo de sustitución local que trata de explicar predicciones individuales de un modelo de caja negra mediante aproximaciones. El funcionamiento de LIME se resume en probar distintas permutaciones y variaciones de los datos de entrada sobre un modelo para observar el efecto que tienen en la salida, de tal modo que se puede hacer un esquema general de qué importancia tiene cada atributo y cuál es su interacción con el resto. El analista puede utilizar estas aproximaciones para entender cuál es el motivo para un cierto resultado. En términos matemáticos, se puede definir la explicación para una instancia como

$$exp(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{22}$$

donde la explicación para la instancia x es el modelo g que minimiza la función de pérdida L, que mide cómo de cercana es la explicación a la predicción del modelo original f, mientras que la complejidad $\Omega(g)$, definida como el número de atributos del modelo g, se mantiene baja. G representa el conjunto de posibles explicaciones y π_x define el tamaño del entorno que se considera alrededor de la instancia x.

En la práctica, LIME sólo optimiza el término relativo a la pérdida, por lo que el usuario determina la complejidad mediante la selección del número de atributos que desea en su explicación.

Ambos algoritmos presentan la posibilidad de obtener gráficas que permiten comprender diferentes aspectos de las elecciones realizadas por los modelos, que serán las herramientas utilizadas para la interpretabilidad de los mismos en este trabajo. Para facilitar su comprensión, se ha añadido en el Anexo IV: Análisis detallado de la interpretabilidad un tutorial de las mismas.

5 Experimentos

5.1 Resultados de reproducibilidad

En este capítulo se detallan los resultados obtenidos por los modelos desarrollados en este TFG para cada uno de los experimentos de entrenamiento y se comparan con los métodos de referencia. Cabe destacar que no existe información disponible sobre el rendimiento de los métodos presentados en el reto para los problemas intermedios, es decir, los resultados de predicción sobre $D2$ y $D3$ utilizando $D1$ como datos de entrenamiento. Sólo se conocen sus puntuaciones finales sobre $D4$. No obstante, incluimos los resultados obtenidos con nuestros métodos por completitud. Las métricas utilizadas se explican detalladamente en el Anexo I: TADPOLE Challenge.

A continuación, se detallan los resultados del problema de diagnóstico clínico sobre cada dataset, presentando comparativamente sus mAUC y BCA:

Método	$D1$ $D2$		$D1\_RedD3$ $D3$		$D1D2\_RedD4$ $D4$		$D1D2\_Aug$ $D4\_Aug$	
	mAUC	BCA	mAUC	BCA	mAUC	BCA	mAUC	BCA
Frog (GB)	-	-	-	-	-	-	93.1	84.9
Threedays (RF)	-	-	-	-	-	-	92.1	82.3
EMC-EB (SVMs)	-	-	-	-	-	-	90.7	80.5
Gradient Booster	98.7	98.3	92.5	90.5	87.3	85.2	97.6	97.1
Random Forest	93.7	92.0	71.6	63.2	69.8	61.1	94.0	92.6
EMC-EB GitHub	-	-	-	-	-	-	87.4	79.8

En la siguiente tabla, se detallan los resultados del problema de volumen de ventrículos sobre cada dataset, presentando comparativamente sus MAE y WES:

Método	$D1$ - $D2$		$D1D2\_Aug$ - $D4\_Aug$	
	MAE	WES	MAE	WES
Frog (GB)	-	-	0.45	0.33
Threedays (RF)	-	-	-	-
EMC-EB (SVMs)	-	-	0.45	0.40
Gradient Booster	0.15	0.54	0.27	0.37
Random Forest	0.03	0.10	0.32	0.44
EMC-EB GitHub	-	-	1.878*	Nan

**Nota: EMC-EB GitHub ofrece los resultados en un formato distinto a los oficiales del reto. Este dato se refiere a VV en % de ICV. Este dato no es comparable ni convertible a la métrica general externamente. Además, no presenta un resultado de WES (salida Not a Number). El paquete está todavía en desarrollo y no ha sido posible corregir estos problemas modificando su código.*

Con los modelos de GB y RF implementados en este TFG se ha conseguido superar los resultados de los modelos correspondientes presentados en el reto (Frog y Threedays, respectivamente). Con GB hemos conseguido un mAUC de 97.6 y con RF hemos conseguido un valor de 94.0 mientras que Frog, el método ganador del reto, obtuvo un mAUC de 93.1. Adicionalmente, también se han obtenido mejores resultados en el pronóstico del volumen de los ventrículos, con un MAE de 0.27 para el GB y 0.32 para el RF, superando el 0.45 de Frog. Cabe destacar que con el método basado en SVM, si bien es el proporcionado directamente por los participantes, hemos obtenido una bajada de rendimiento respecto a la ofrecida en el reto, bajando su mAUC de 90.7 a 87.4. No hay información disponible de por qué puede haber ocurrido esto, pero es posible que al ser una versión abierta haya algún parámetro que los autores hayan modificado para requerir una menor potencia computacional, sacrificando algo de precisión. En el Anexo II: Un comentario sobre EMC-EB se incluye una explicación más detallada de este problema.

5.2 Resultados de Interpretabilidad

Para el análisis de los resultados de interpretabilidad de este TFG es necesario establecer en primer lugar el método de diagnóstico utilizado en ADNI. De la información proporcionada por ADNI en [13] se puede extraer la siguiente cita del apartado de métodos:

> *The subjects for the study were classified as normal controls, subjects with MCI, or subjects with mild AD. The criteria for classification of the subjects were as follows. With respect to memory complaints, the normal subjects had none, while the subjects with MCI and subjects with AD both had to have complaints. On the Mini-Mental State Examination (MMSE), the range for the normal subjects and subjects with MCI was 24–30, and for AD 20–26; all are inclusive. The CDR score for normal subjects was 0 and for subjects with MCI was 0.5 with a mandatory requirement of the memory box score being 0.5 or greater, and the rating for subjects with AD was 0.5 or 1. For the memory criterion, delayed recall of 1 paragraph from the Logical Memory II subscale of the Wechsler Memory Scale–Revised (maximum score of 25) was used with cutoff scores as follows based on education: normal subjects ≥9 for 16 years of education, ≥5 for 8–15 years of education, and ≥3 for 0–7 years of education. For subjects with MCI and subjects with AD, these scores were ≤8 for 16 years of education, ≤4 for 8–15 years of education, and ≤2 for 0–7 years of education.*

Así podemos concluir que para el diagnóstico de la enfermedad se utilizaron las siguientes medidas:

1. Quejas sobre pérdida de memoria, identificadas mediante exámenes cognitivos
2. Puntuación MMSE
3. Puntuación CDR
4. Puntuación Wechsler Memory Scale–Revised

De las cuatro medidas, la única que no existe como atributo en los datos de TADPOLE es Wechsler Memory Scale–Revised. Por lo tanto, estos atributos deberían ser de relevancia para aquellos métodos de aprendizaje más fiables en el diagnóstico de la enfermedad.

A continuación se analiza, para los modelos basados en GB y RF, el resumen de los valores de SHAP y un error por cada clase (CN, MCI, AD) según la interpretación de LIME en su predicción de $D1$ sobre $D2$ y en su predicción de $D1D2\_Aug$ sobre $D4\_Aug$.

Nota: Dada la dificultad para comprender la representación de los resultados generados por SHAP y LIME, se puede encontrar en el Anexo IV: Análisis detallado de la interpretabilidad un pequeño tutorial que puede ayudar con la interpretación de las gráficas que presentamos a continuación. SHAP muestra un resumen, para cada clase, del efecto (positivo o negativo) sobre todo el conjunto de datos. LIME permite analizar para cada clase el efecto de cada atributo en un dato concreto.

5.3 Entrenamiento sobre $D1$, Predicción sobre $D2$

A continuación se presenta la interpretación sobre el problema de datos longitudinales a corto plazo, explicado en la Sección 2.2 utilizando SHAP y LIME.

5.3.1 Gradient Booster

La Figura 3 muestra el resumen de la importancia de los atributos principales para cada clase de diagnóstico (CN, MCI y AD). Para entender estos gráficos, se debe considerar que los valores de SHAP analizan la contribución de cada atributo a una clase en un formato one-vs-all, es decir, dan mayor valor a un atributo si contribuye positivamente a clasificar con el diagnóstico concreto en cuestión, o negativamente si contribuye a clasificar como cualquiera de los demás.

Como podemos observar, en las predicciones de CN y AD hay una evidente predominancia de los atributos que ADNI utiliza para el diagnóstico. Podemos apreciar que los atributos más

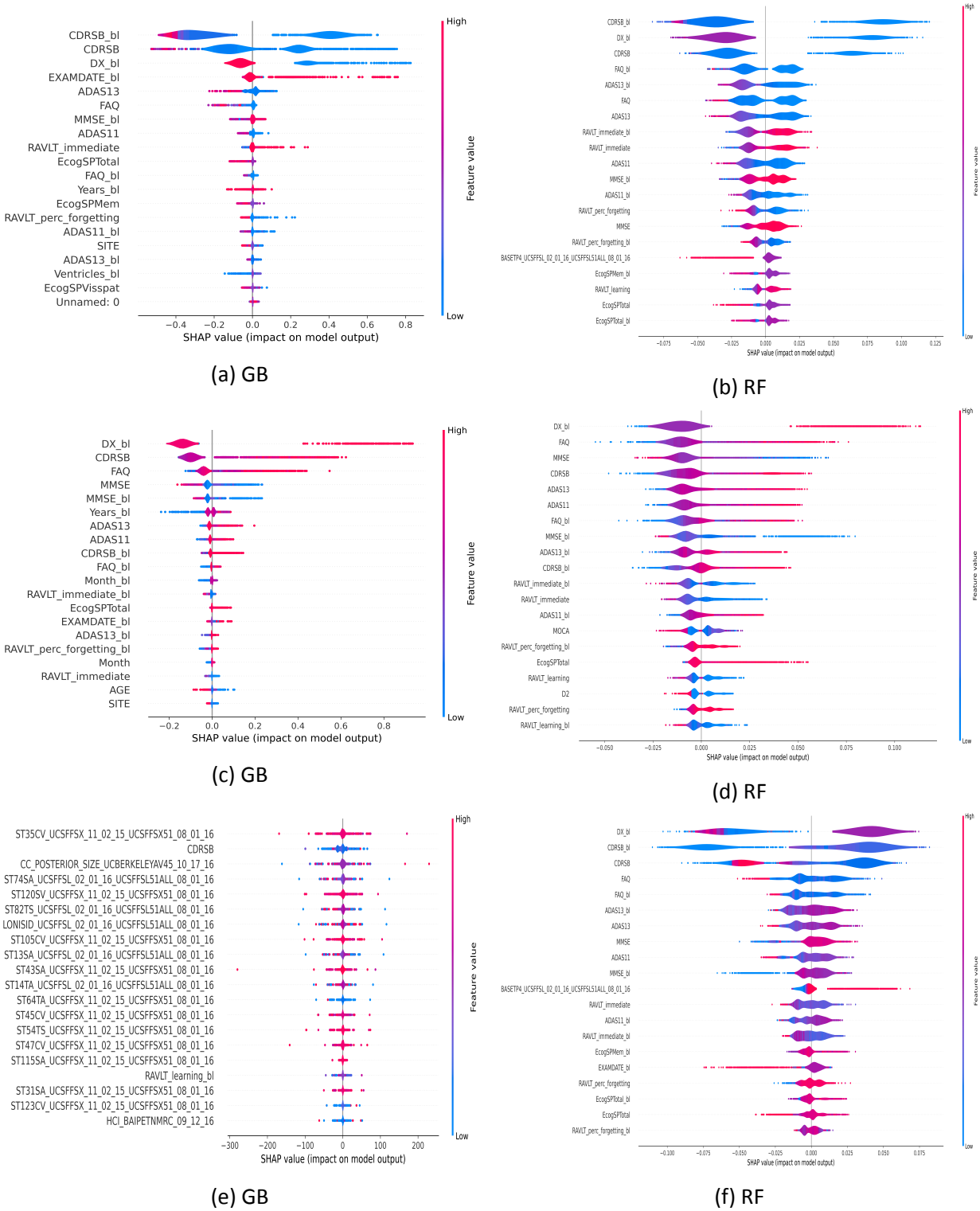


Figure 3: Resumen de valores de SHAP en el problema $D1$ vs $D2$. Arriba, resultados relativos a la clase CN. Centro, clase AD. Abajo, clase MCI.

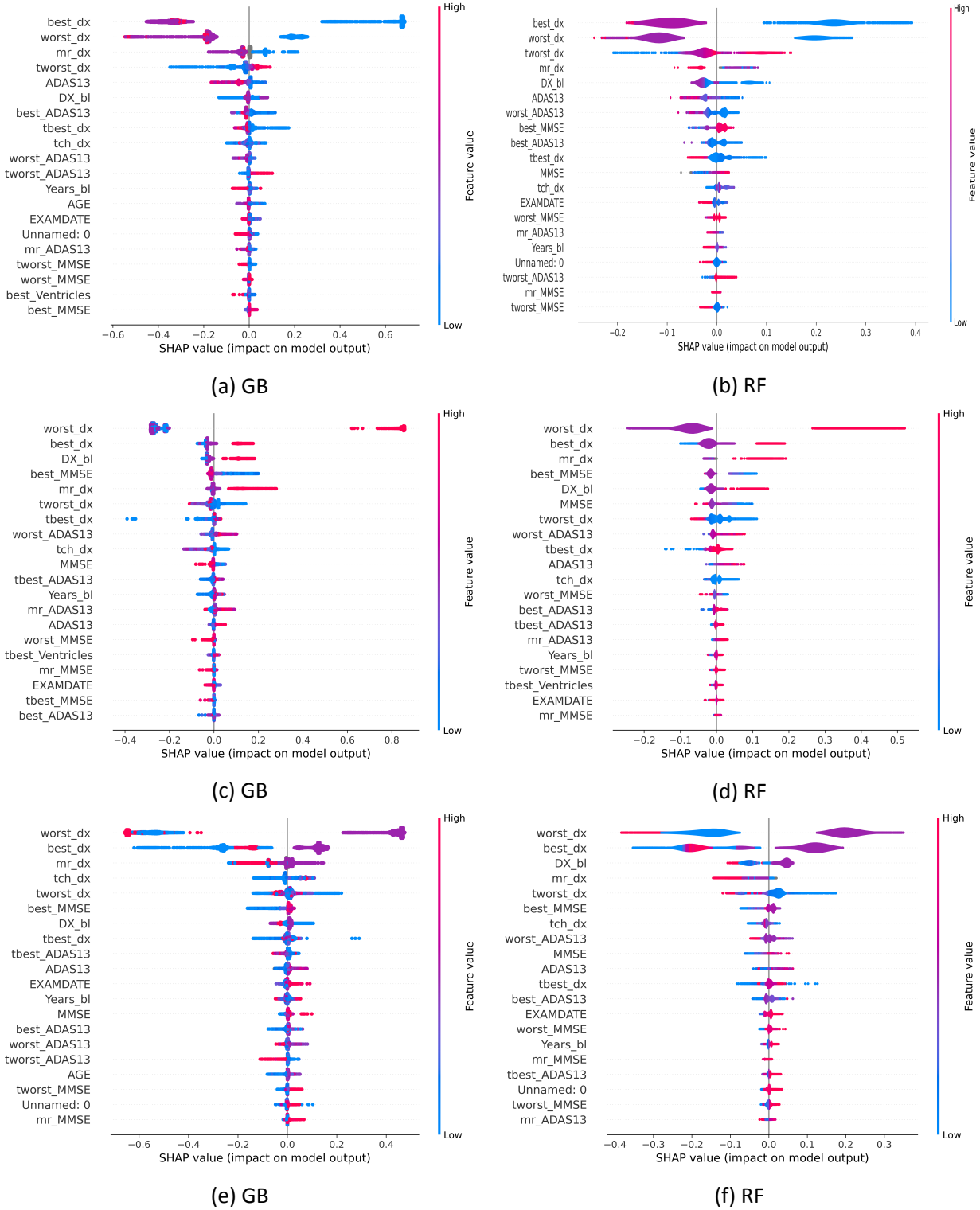


Figure 4: Resumen de valores de SHAP en el problema $D1D2\_Aug$ vs $D4\_Aug$. Arriba, resultados relativos a la clase CN. Centro, clase AD. Abajo, clase MCI.

relevantes, y con una diferencia muy importante respecto de los siguientes, son: CDRSB y CDRSB_bl (puntuación CDR), DX_bl (diagnóstico baseline), ADAS11, ADAS13, FAQ y FAQ_bl y RAVLT (exámenes cognitivos).

De esta información podemos realizar el siguiente análisis. El modelo de GB obtenido tiene una gran capacidad de reconocer la información utilizada para la clasificación en la práctica clínica, ya que SHAP interpreta que GB da una gran importancia a las mediciones utilizadas por los especialistas para el diagnóstico. Del mismo modo, SHAP indica que GB atribuye relevancia a otra serie de mediciones que, si bien no se utilizaron para el diagnóstico en ADNI, se han mostrado relevantes para el diagnóstico de la enfermedad en la literatura.

Además, el modelo da importancia al diagnóstico en la visita de baseline (DX_bl), a la fecha del primer examen (EXAMDATE_bl), al propio orden de aparición de los ensayos (Unnamed: 0), que como ya sabemos están ordenados por paciente y fecha, y a la edad del paciente (Years_bl). De este modo, parece que el sistema entiende que la condición AD es no reversible, y que por tanto el paso del tiempo sólo puede significar un deterioro cognitivo o, en el mejor caso, un sostenimiento de la situación, pero nunca (o casi nunca) una mejora.

Es muy interesante el hecho de que para el diagnóstico de MCI el modelo utilice como referencia los datos derivados de las segmentaciones con FreeSurfer de MRI y PET. En concreto, aparece como atributo más destacado el volumen del sulco lateral occipital izquierdo y, en tercer puesto, el volumen de la circunvolución o giro cingulado posterior. Son medidas que no tuvieron especial interés en [5], pero que aparecen de gran importancia para los pacientes con MCI en este análisis. No queda claro cómo afecta cada una de estas medidas a la clasificación positiva o negativa, ya que la distribución es simétrica e invariante al valor para casi todas ellas. También es destacable que de los atributos relevantes para la clasificación en CN o AD sólo permanecen CDRSB y RAVLT.

Para el análisis mediante LIME hemos seleccionado el que consideramos el error más interesante de cada clase producido por GB. Los resultados pueden visualizarse en la Figura 5.

En la Figura 5a, se muestra un caso muy excepcional en el que un paciente que había sido diagnosticado como MCI ha revertido su estado a CN con el tiempo. En el conjunto de entrenamiento apenas existen casos como este, lo que ha llevado a GB a aprender que un paciente no puede revertir su estado de este modo. En la práctica clínica ningún método de diagnóstico es infalible por lo que sí se permite corregir un diagnóstico con el tiempo. El clasificador obtiene un valor bajo para la probabilidad de la clase CN debido a que varios de los atributos faltantes (marcados con -1) son considerados con una contribución negativa importante. De los atributos presentes, contribuyen negativamente un tamaño reducido del ventrículo lateral derecho, y del giro cingulado anterior, y positivamente un valor de SUVR-PET alto.

En la Figura 5b podemos apreciar cómo los atributos presentan evidencias contrarias. La

medición de AV1451 (update_stamp_UCBERKLEYAV1451_10_17_16 , imagen tau-PET) apuesta fuertemente por la correcta clasificación como MCI, pero los exámenes cognitivos CDR, ADAS y FAQ apuntan hacia la incorrecta clasificación como AD. En este caso podemos apreciar que GB da una excesiva importancia a estos atributos que, aunque en general son buenos indicadores para AD, en este caso no debieron prevalecer sobre los atributos de imagen computacional.

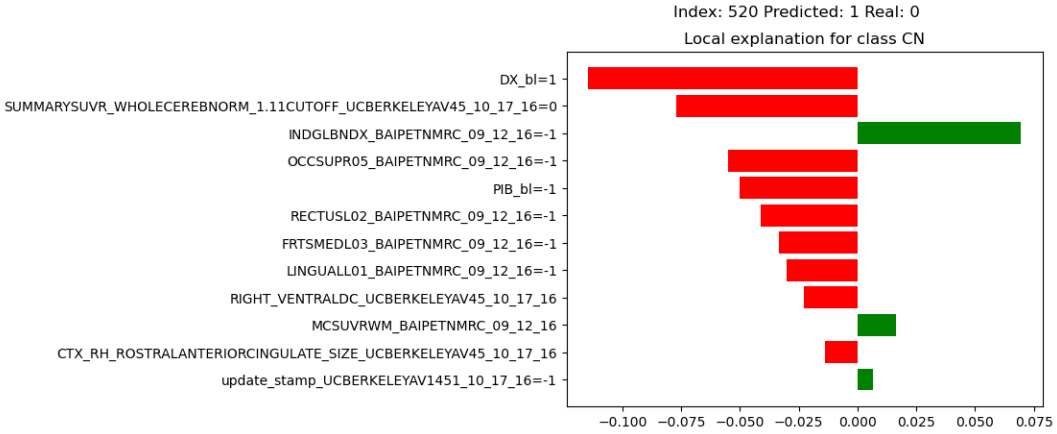
En 5c, tenemos una circunstancia problemática. Los indicadores que el clasificador encuentra como más importantes para determinar al paciente como de AD tienen un valor de -1, es decir, son datos faltantes. Parece ser que el clasificador ha aprendido que existe una cierta correlación entre cómo fueron obtenidos los datos y los diagnósticos. Esto se podría solucionar utilizando métodos que permitiesen interpolar los datos faltantes y reentrenando el clasificador con la esperanza de mejorar la capacidad de generalización del clasificador en estos casos sin perjudicar el resto.

5.3.2 Random Forest

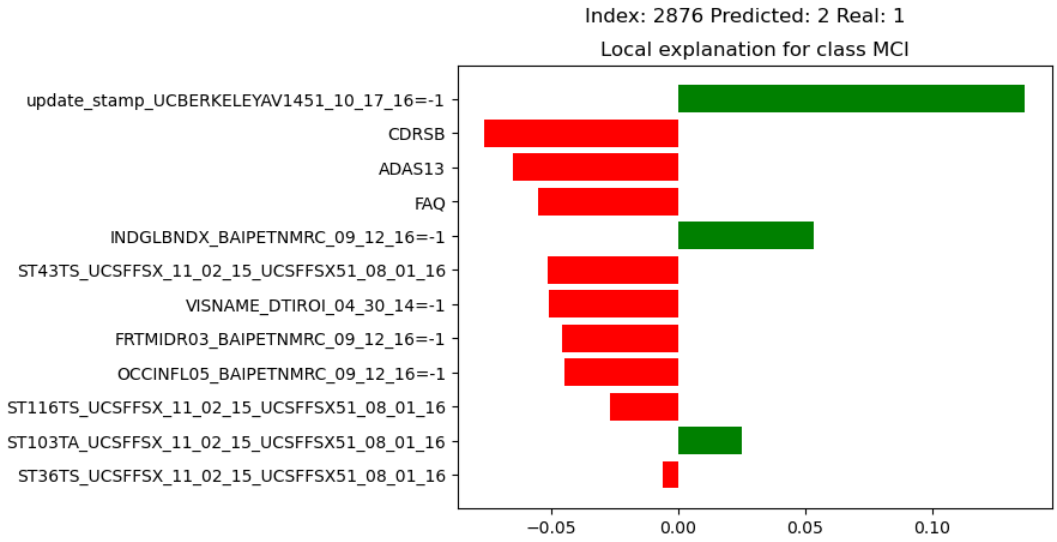
Si algo se hace destacable observando la Figura 3 es que este modelo genera una distribución de la importancia de los atributos muy superior a la de GB. Esto tiene sentido por las diferencias prácticas entre ambos modelos anteriormente explicadas, que hacen que RF distribuya la relevancia de un atributo mediante su repetida aparición en muchos árboles y no mediante la optimización de la función de pérdida. Aun así, los atributos que ambos modelos reconocen como relevantes son esencialmente los mismos: CDR como puntuación más fiable, toma en consideración del diagnóstico base y la información de tiempo para comprobar la evolución y revisión de los exámenes cognitivos ADAS, MMSE y RAVLT. Debe apuntarse además que en este caso, para la clase MCI se utilizan atributos muy similares a los utilizados para CN o AD. Es probable que esto se deba a que RF es un método con una mayor capacidad de generalización que GB, dado su método de generación de árboles. A pesar de ofrecer menos precisión, creemos que en este caso concreto RF ofrece más fiabilidad que GB, pues su método de diagnóstico para casos de MCI son más parecidos a los de la práctica clínica.

En la Figura 6 podemos ver una serie de instancias clasificadas erróneamente que son mucho más complicadas de explicar que las vistas anteriormente.

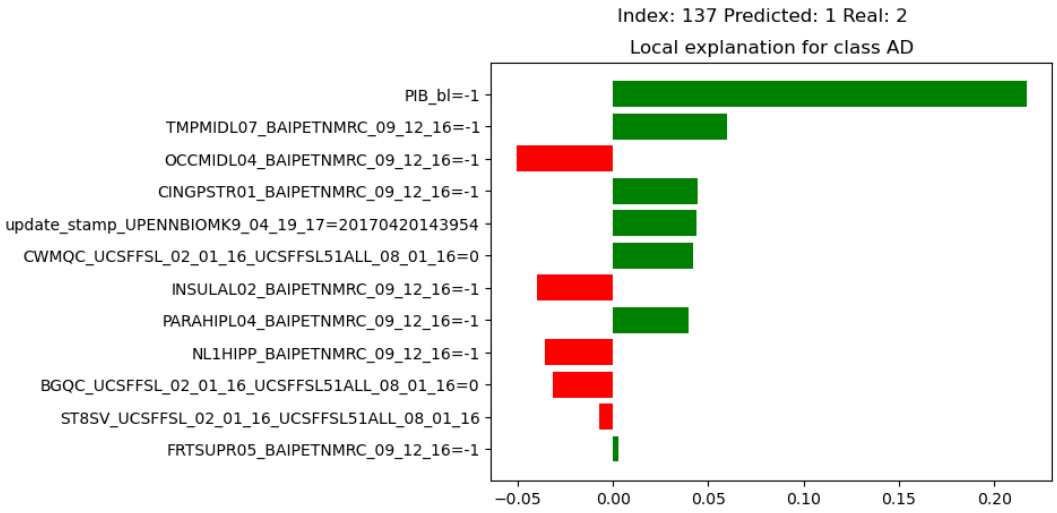
En la Figura 6a podemos apreciar un caso en el que un paciente presenta indicios fuertes de deterioro en los test cognitivos y algunos síntomas leves en distintas medidas de PET. Sin embargo, el paciente fue diagnosticado como CN. No queda claro si esto es debido a un error en el procedimiento o si verdaderamente los médicos consideraron que este paciente con resultados en los test típicos de AD era CN. Quizá se esté omitiendo información relevante, como que el paciente sufriera de alguna enfermedad mental adicional como la depresión que le hiciera fallar en los test cognitivos y el médico tomó la decisión de diagnosticar a este paciente como normal. En concreto, se puede ver como los test cognitivos de FAQ, CDRSB, ECog y RAVLT señalaban que este paciente



(a) Paciente CN clasificado como MCI



(b) Paciente MCI clasificado como AD



(c) Paciente AD clasificado como MCI

Figure 5: Interpretación de errores con LIME para GB en el problema $D1$ vs $D2$

presentaba deterioro cognitivo. Del mismo modo, los datos de imagen de hipointensidades de materia blanca, los ganglios basales, el tamaño del plexo coroideo izquierdo y el de los ventrículos laterales señalaban un deterioro de tamaño que, para el clasificador, indica baja probabilidad de estar sano.

En la Figura 6b podemos ver un caso de información conflictiva, en el que ningún atributo de gran relevancia está involucrado. Parece que la evidencia varía de un atributo a otro y esto llevó a confundir al clasificador. Llama nuestra atención que se establece como negativo el hecho de ser mujer (PTGENDER=1) para padecer MCI. Esto se debe a que las mujeres tienen una mayor tendencia a padecer Alzheimer, siendo casi el doble de la de los hombres en la población mundial.

En la Figura 6c tenemos un caso excepcional. Como se puede apreciar, casi toda la evidencia apunta fuertemente a AD: FAQ, CDR, imagen PET, Ecog... Sin embargo, el clasificador lo sitúa como MCI. Utilizando LIME para explicar la clase seleccionada, podemos apreciar en la Figura 7 que lo que lleva esencialmente al clasificador a esta decisión es la relación entre datos faltantes y la clasificación MCI, mezclada con un diagnóstico en la baseline de MCI. Como hemos comentado, una solución a este problema es seleccionar un método de imputación de datos distinto.

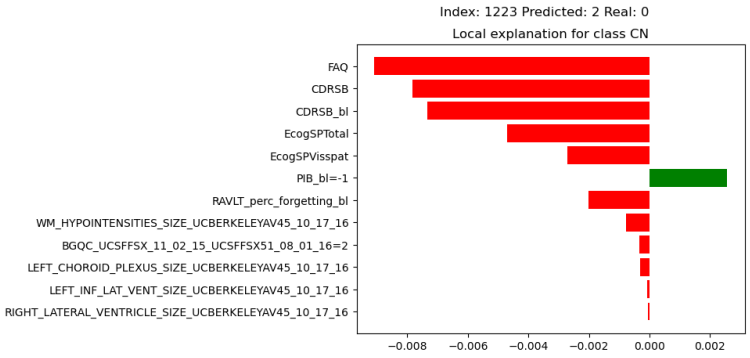
5.4 Entrenamiento sobre $D1D2\_Aug$, Predicción sobre $D4\_Aug$

Se presenta la interpretación sobre el problema de datos reducidos a largo plazo, explicado en la Sección 2.2.

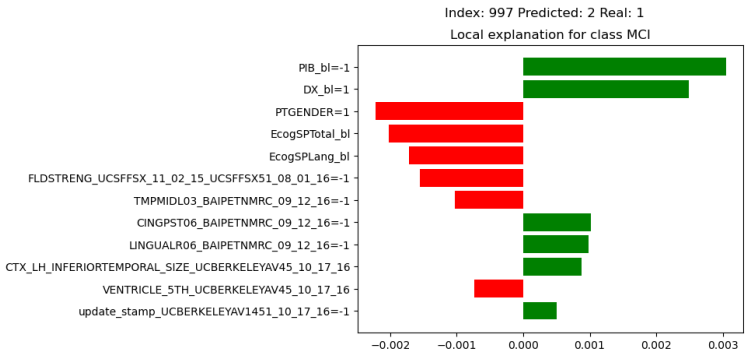
5.4.1 Gradient Booster

Como podemos apreciar en la Figura 4, en este caso la mayor relevancia para el diagnóstico la presentan los atributos aumentados. Del mismo modo que sucedía con la predicción sobre $D2$, el modelo encuentra muy útil la información temporal y de las distintas etapas de diagnóstico para poder establecer si un paciente padecerá AD en un momento concreto, dado que selecciona AGE, Years_bl, EXAMDATE y Unnamed:0 como relevantes.

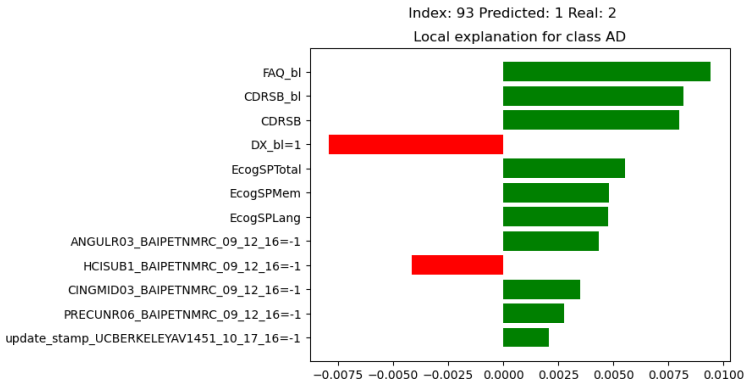
GB establece worst_dx como el atributo de mayor relevancia con enorme diferencia en el diagnóstico de la clase AD. Parece que el modelo es capaz de comprender que un paciente que haya sido diagnosticado previamente como AD tiene grandes probabilidades de seguir siéndolo. Cabe destacar cómo los distintos exámenes cognitivos siguen teniendo una gran relevancia para GB y su aportación se corresponde con la de los diagnósticos clínicos. Por ejemplo, valores bajos de MMSE como indicadores de AD. Como apunte, parece relevante que frente a la falta de datos de imagen, GB selecciona ahora atributos similares a los de las clases CN y AD para predecir MCI.



(a) Paciente CN clasificado como AD



(b) Paciente MCI clasificado como AD



(c) Paciente AD clasificado como MCI

Figure 6: Interpretación de errores con LIME para RF en el problema $D1$ vs $D2$

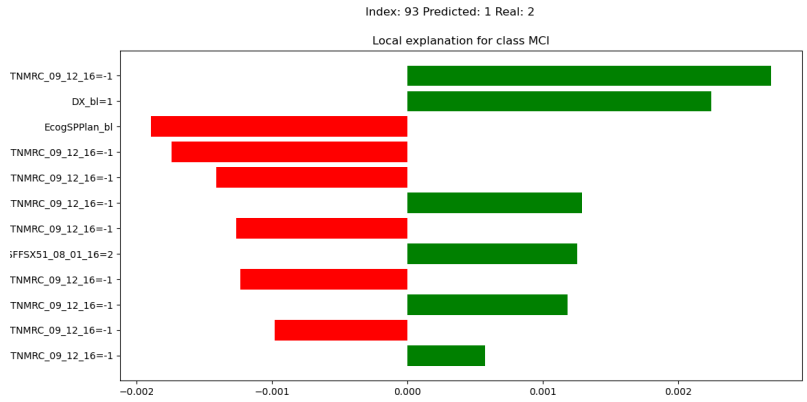


Figure 7: Interpretación con LIME del error de clasificación del RF en el problema $D1$ vs $D2$ (AD clasificado como MCI), visto en 6c, explicado desde el punto de vista de la clase MCI

En este caso GB no ha presentado errores en la clasificación de grupo de AD, por lo que la Figura 8 sólo presenta un error de clase CN y uno de clase MCI. Tanto en 8a como en 8c se aprecian dos casos excepcionales en los que un paciente ha sido diagnosticado con un estado de deterioro superior al real.

Como se ha comentado anteriormente, una solución a este problema es que se añadan al conjunto de entrenamiento suficientes datos para que el clasificador pueda llegar a tener en cuenta estas excepciones. Si esto no es posible, se podría hacer una selección de características previa o se podría transformar el espacio de datos para acomodarlo a estos casos. Un detalle interesante es que, en ambos casos, la coincidencia del diagnóstico en la visita de baseline (DX_bl) con la clase real no es considerada como un punto a favor del diagnóstico. Esto quiere decir que el clasificador no toma el diagnóstico base como muy fiable, aunque no podemos asegurar el motivo.

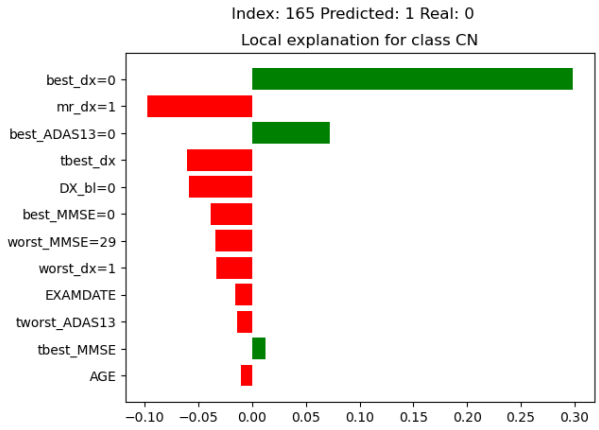
5.4.2 Random Forest

En este caso y de igual modo que para GB, encontramos una dominancia muy importante de los atributos aumentados. Incluso con la tendencia del RF a distribuir la carga de importancia de los atributos, en este caso resulta demasiado relevante el diagnóstico previo de los pacientes para determinar su estado actual. Además, se percibe la relevancia del tiempo en el modelo con la inclusión de la edad, la edad base o las fechas de examen de cara al diagnóstico. Cabe destacar que para RF, ADAS13 parece ser la puntuación cognitiva más precisa, ya que aparecen tanto la misma como sus distintos aumentos en muy buena posición en todas las clases de diagnóstico, especialmente MCI.

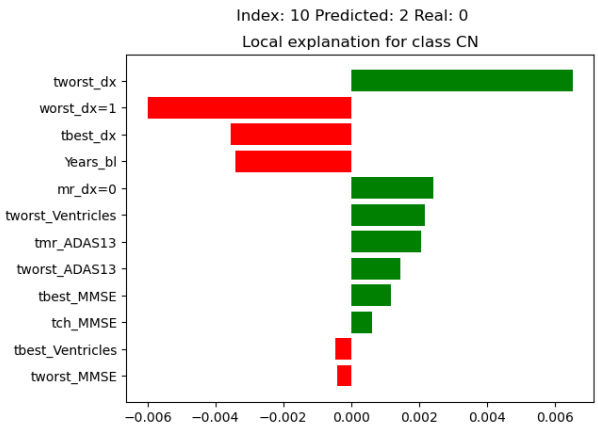
En la Figura 8 apreciamos los errores más interesantes, sin incidencias al clasificar AD.

En 8b es interesante que, a pesar de ser uno de los típicos fallos de retorno a un diagnóstico de menor deterioro, se da mucha importancia a que dicho diagnóstico fue hace mucho tiempo (tworst_dx, tiempo desde el peor diagnóstico). Los exámenes cognitivos apuntan a CN, pero el clasificador sigue dando mucha importancia al diagnóstico previo.

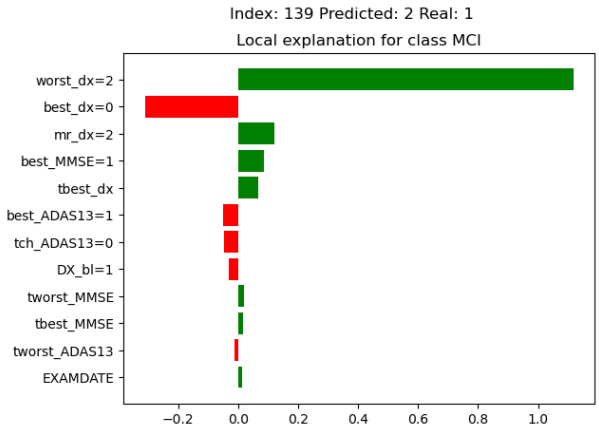
En 8d vemos un problema relevante. Que el peor diagnóstico hasta la fecha de un paciente sea MCI no lo hace muy propenso a ser diagnosticado como MCI, sino a ser diagnosticado como AD. Este es un sesgo de los datos de entrenamiento que resulta problemático, porque generaliza la rápida evolución de MCI a AD de algunos pacientes a todos los pacientes. En el caso presentado, para un paciente que no hace mucho tiempo era CN y que ha presentado buenas puntuaciones en MMSE, la generalización no es correcta.



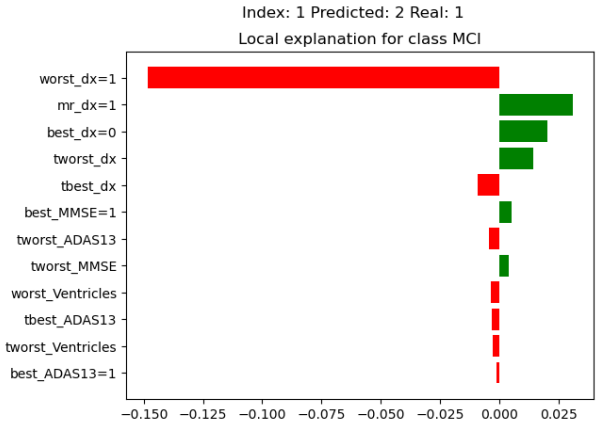
(a) GB, Paciente CN clasificado como MCI



(b) RF, Paciente CN clasificado como AD



(c) GB, Paciente MCI clasificado como AD



(d) RF, Paciente MCI clasificado como AD

Figure 8: Interpretación de errores con LIME en el problema $D1D2\_Aug$ vs $D4\_Aug$. Izquierda, GB. Derecha, RF.

6 Conclusiones y Trabajo Futuro

Gracias al TADPOLE Challenge ha sido posible obtener una colección de métodos de aprendizaje automático muy potentes para el diagnóstico y pronóstico de la enfermedad de Alzheimer. Si bien los resultados de dichos métodos son importantes, no es posible avanzar el estado del arte, profesionalizar y comercializar soluciones como las presentadas si no existe un acceso directo a los mismos, por lo que comprobar su reproducibilidad y su interpretabilidad es una tarea de vital importancia.

Mediante este TFG, se ha demostrado que los métodos ganadores de TADPOLE no sólo son reproducibles y precisos, sino también son fiables y se puede extender su uso a espacios de tiempo mayores. Las librerías más avanzadas de interpretabilidad propuestas hasta la fecha han posibilitado que comprobemos que estos modelos son capaces de establecer una correcta relación entre las distintas mediciones clínicas de un paciente y su diagnóstico clínico, pues los mejores algoritmos del reto dan importancia a las mismas mediciones que se utilizan en la práctica clínica. Además, los resultados de interpretabilidad han destacado la utilidad de algunos atributos adicionales que se han destacado como relevantes en otros trabajos de estado del arte (CDR como puntuación más fiable, consideración del diagnóstico base y la información de tiempo y revisión ADAS, MMSE y RAVLT). Por tanto, creemos que este hallazgo justifica la creación de herramientas comerciales basadas en métodos de Gradient Boosting o Random Forest para el apoyo de las instituciones sanitarias en materia de diagnóstico y pronóstico de Alzheimer.

Como trabajo futuro, sería muy interesante comprobar el rendimiento de estos modelos en las mediciones publicadas en las futuras fases de ADNI. Se ha encontrado como bastante limitante el diagnóstico de los pacientes de $D4$ con tan sólo 11 atributos. Creemos que del estudio realizado de la pérdida de potencia de los modelos pone de manifiesto la necesidad de poder acceder a todos los datos posibles, o al menos los más relevantes para los mejores algoritmos. Adicionalmente, sería de interés estudiar en qué atributos se basa un mal algoritmo predictivo, para plantear el descartarte de atributos.

El diseño de los datos de TADPOLE ha dificultado que los algoritmos de deep-learning hayan ocupado las mejores posiciones del ranking. Sería un trabajo futuro interesante aumentar el dataset para que estos algoritmos pudiesen mostrar todo su potencial. La interpretabilidad en deep-learning es más complicada que la utilizada en los algoritmos de aprendizaje automático convencional, pero tanto SHAP como otros algoritmos comienzan a ofrecer soluciones al respecto.

Finalmente, los resultados de este TFG constituyen un primer paso para la creación de una herramienta de ayuda al diagnóstico estandarizada y fiable que, de forma automática, sea capaz de seguir todo el proceso que se ha seguido en este trabajo: corrección y estandarización de datos, entrenamiento del modelo, interpretación general del modelo, pronóstico de resultados futuros e interpretación de los resultados para AD u otras enfermedades, ofreciendo un gran valor social.

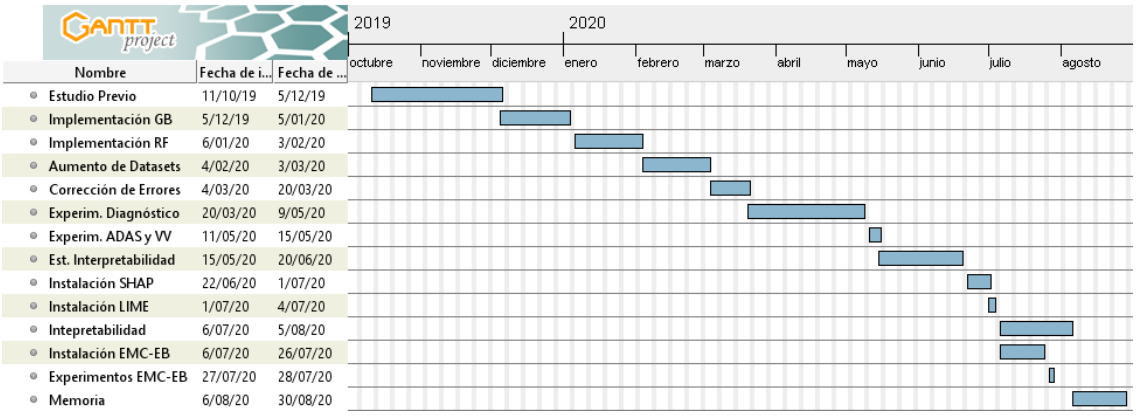


Figure 9: Diagrama de Gantt del proyecto

Este TFG se ha desarrollado entre octubre de 2019 y agosto de 2020. Parte del trabajo se desarrolló como prácticas en la Universidad de Zaragoza. En la Figura 9 se muestra un diagrama de Gantt con las tareas desarrolladas.

Anexo I: TADPOLE Challenge

En este anexo se presenta información adicional relevante acerca del análisis del TADPOLE Challenge realizado en este TFG.

Atributos más relevantes

En esta subsección se recopilan y explican los atributos del dataset de TADPOLE que han destacado como más relevantes en los resultados de este TFG:

CDRSB: Clinical Dementia Rating - Sum of Boxes. Puntuación que mide las capacidades en memoria, orientación, juicio, asuntos comunitarios, hobbies y cuidado personal en una escala de 1 a 5.
DX: Diagnóstico clínico (CN, MCI o AD).
EXAMDATE: Fecha del exámen clínico.
ADAS11: Medición del estado cognitivo en 11 pruebas, respectivamente, recordar palabras, recordar el nombre de objetos, seguir órdenes, praxis en construcción, praxis en ideación, orientación, reconocimiento de palabras, recordar direcciones, leguaje hablado, comprensión, dificultad para encontrar palabras.
ADAS13: Puntuación AdasCog-13. Versión extendida de AdasCog-11, consistente de 13 pruebas que miden el decline cognitivo.
FAQ: Functional Activities Questionnaire. Mide la capacidad del paciente de realizar actividades para la autogestión, como prepararse una comida sin ayuda o llevar las finanzas personales.
MMSE: Mini Mental State Examination. Cuestionario de 24 puntos que reconoce una serie de habilidades. Orientación, enfoque, atención, memoria, nominación, repetición, lectura, habilidad y escritura.
RAVLT: Rey's Auditory Verbal Learning Test. Es un examen basado en recordar listas de palabras a lo largo de periodos de tiempo cortos. Se dividen sus puntuaciones por períodos temporales (inmediato, corto, medio...).
Ecog: Everyday Cognition. Exámen que mide de manera general y específica las capacidades cognitivas del desarrollo diario de los pacientes (memoria, lenguaje, capacidad visual-espacial, planificación, organización y atención dividida).
Ventricles: Volumen de los ventrículos estimado mediante la herramienta de simulación de imagen neurológica computacional FreeSurfer.
PIB: Pittsburgh compound B. Se refiere a los análisis realizados con este radioactivo usado para escáneres PET.
Unnamed:0: Posición de la visita en el dataset, referencia de orden en el documento.
ST35CV_UCSFFSX_11_02_15: Volumen del lóbulo occipital lateral izquierdo.
CC_POSTERIOR_SIZE: Tamaño de la corteza cingulada.

ST74SA_UCSFFSL: Sección medio-frontal del núcleo caudado derecho.
SUMMARYSUVR_WHOLECEREBNORM: Standardized Uptake Value Ratio. Consumo de florbetapir amiloide en el cerebro.
APOE4: Gen de la Apolipoproteína E, con cuatro alelos (APOE1, 2, 3 y 4), donde el APOE4 está relacionado con el AD.
FDG: Tomografía por Emisión de Positrones con 18F-fluorodeoxiglucosa (PET-FDG) es una técnica de diagnóstico por imagen, este atributo resume sus datos en una puntuación.
AV45: Puntuación del diagnostico por imagen SUVR mediante el compuesto florbetapir-fluorine-18 (18F-AV45).
HippoCampus: Volumen del hipocampo mediante FreeSurfer.
WholeBrain: Volumen del cerebro completo mediante FreeSurfer.
Ethorinal: Volumen de la corteza entorrinal.
Fusiform: Volumen del giro fusiforme.
Temp_Lobe: Volumen del lóbulo temporal.

Nota: Cualquier atributo que presenta la etiqueta '_bl' se refiere al valor de dicho atributo en la visita de baseline

No todos los atributos que aparecen en las gráficas han sido explicados por carecer de relevancia suficiente. En caso de requerir todas las descripciones detalladas se pueden encontrar los datos en la propia web de TADPOLE (https://tadpole.grand-challenge.org/Data/)

Evaluación

El TADPOLE Challenge establece la selección de los ganadores del reto según la evaluación de sus predicciones en base a las siguientes métricas:

1. Área Multiclase Bajo la Curva Operativa del Receptor (Multi-class Area Under the Receiver Operating Curve, mAUC):

   Esta métrica generaliza la idea de la ROC para clasificación binaria a un problema de múltiples clases. Dadas L clases, el número de pares de clases es $L(L-1)/2$ de tal modo que para una clasificación de clase c_i contra la clase real c_j sea:

$$mAUC = \frac{2}{L(L-1)} \sum_{i=2}^{L} \sum_{j=1}^{i} \hat{A}(c_i, c_j) \tag{23}$$

   donde:

$$\hat{A}(c_i, c_j) = \frac{\hat{A}(c_i|c_j) + \hat{A}(c_j|c_i)}{2} \tag{24}$$

Se utiliza la forma estándar de la AUC, $\hat{A}(c_i|c_j)$ donde n_i y n_j son el número de puntos pertenecientes a las clases i y j respectivamente, mientras que S_i es la suma de rangos de los puntos de test de clase i tras ordenar todos los puntos de las clases i y j en orden creciente de probabilidad de pertenecer a la clase i:

$$\hat{A}(c_i|c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}. \tag{25}$$

2. Precisión de Clasificación Equilibrada (Balanced Classification Accuracy, BCA):

   La BCA general se calcula como la media de las BCA para cada clase:

$$BCA = \frac{1}{L}\sum_{i=1}^{L} BCA_i, \tag{26}$$

   donde BCA_i se define del siguiente modo. Asignando a cada dato una clasificación dura, es decir, clasificándolo en aquella clase de máxima probabilidad tal que $max[P_{CN}, P_{MCI}, P_{AD}]$, la BCA para cada clase i será:

$$BCA_i = \frac{1}{2}[\frac{TP}{TP+FN} + \frac{TN}{TN+FP}] \tag{27}$$

3. Error Medio Absoluto (Mean Absolute Error, MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{M}_i - M_i|, \tag{28}$$

   donde N es el número de datos adquiridos, M_i es el valor real del individuo i en los datos a predecir y $\hat{M}_i$ es la predicción del participante sobre M_i.

Además, se utilizaron medidas auxiliares como la Puntuación con Pesos del Error (Weighted Error Score) y la Precisión de Cobertura de Probabilidad (Coverage Probability Accuracy), pero estas no son relevantes de cara a la evaluación final, por lo que no las tendremos en cuenta en este TFG.

Análisis de Resultados

Los resultados obtenidos en TADPOLE se recogen en 2 artículos, uno hecho público en ArXiv en mayo de 2018 [14], y una continuación de los resultados anteriores obtenidos un año después, hecho público en ArXiv en febrero de 2020 [5]. A fecha del inicio de este TFG, solo estaba disponible el primer artículo, por lo que el estudio y análisis sobre el segundo artículo se han realizado con el TFG bastante avanzado.

Un total de 33 equipos de todo el mundo se presentaron al reto, donde sus modelos

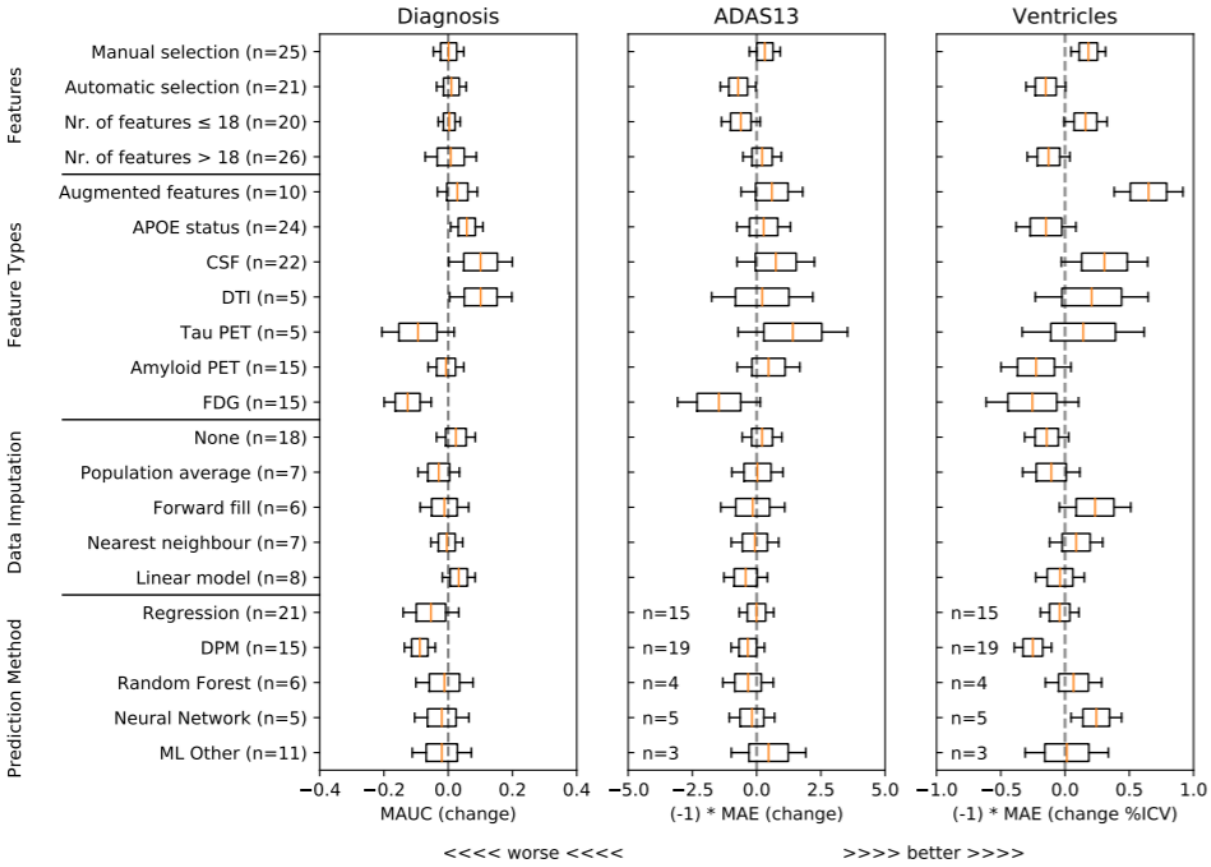


Figure 10: Resumen del efecto que la selección de distintos atributos y métodos ha tenido en cada una de las pruebas.

fueron evaluados para la predicción de las tres métricas anteriormente mencionadas (Diagnóstico, ADAS13 y VV) sobre un total de 219 participantes de ADNI-3 con los datos presentes en el dataset $D4$. Ningún método fue el mejor en todos los ámbitos. Tanto en diagnóstico como en estimación de VV, los modelos presentados mejoran sustancialmente los resultados de referencia, mientras que para la predicción de ADAS13 ningún método es substancialmente mejor que una estimación aleatoria. Bajo los datos limitados de $D3$, el rendimiento general de los modelos sólo se redujo en un 3% de media. Se estima una mejora del rendimiento en diagnóstico clínico a aquellos métodos que incluyeron atributos relacionados con el CSF y de Proyección de Imagen del Tensor de la Difusión (Diffusion Tensor Imaging, DTI). Por otro lado, una mejor estimación del VV se relaciona con la inclusión de estadísticas de resumen de atributos. La Figura 10 resume cómo afectaron dichos tipos de atributos. En oposición, podemos ver cómo la inclusión de atributos sobre amyloid-PET y FDG reducen como norma la calidad de las predicciones tanto para diagnóstico clínico como para VV, y la inclusión de tau-PET sólo sobre diagnóstico. Estos resultados llaman la atención, porque todos estos atributos están bastante establecidos como biomarcadores para la ayuda al diagnóstico (amyloid-PET, tau-PET) y al diagnóstico temprano (FDG-PET).

En el estudio se reporta cómo los métodos de TADPOLE han conseguido avanzar el estado del arte en cuanto a predicción de diagnóstico clínico, siendo el mejor método propuesto en TADPOLE

el mejor de una colección de 15 estudios en la que se reportó una mAUC máxima de 0.902 frente a los 0.931 del mejor método de TADPOLE. Los métodos presentados para TADPOLE resultaron bastante precisos a la hora de predecir valores de VV (MAE 0.4 en el caso del mejor método). Sorprendentemente, no se consiguió ningún resultado relevante para ADAS13 (MAE 3.44 para el mejor método).

Del análisis de los resultados se puede observar que no existe ningún método bueno para todo, dado que ningún método ha ganado en más de una categoría. Se estima un buen rendimiento general para los métodos de agrupación (ensemble), dado que equilibran la sobreestimación de ciertos métodos con la subestimación de otros respecto del deterioro futuro del paciente.

Teniendo en cuenta todo lo anterior, se puede concluir que TADPOLE ha excedido en su capacidad para hacer avanzar las fronteras de la ayuda al diagnóstico clínico utilizando métodos de aprendizaje automático, al menos sobre el resto de sus propósitos. En pos de obtener los resultados de mayor valor posible con este TFG, es necesario enfocar el esfuerzo en un problema acotado. Por ello se decidió, que los métodos seleccionados para estudiar su reproducibilidad e intepretabilidad sean los mejores respecto al diagnóstico clínico.

Anexo II: Un comentario sobre EMC-EB

Se realizó un intento de aplicar las técnicas de intepretabilidad al algoritmo EMC-EB como parte del desarrollo de este trabajo. No obstante, no se consiguió este objetivo por los motivos detallados a continuación.

En lugar de reproducir el método manualmente se hizo uso de la iniciativa TADPOLE-Share, que permite acceder al método original mediante un cuaderno de la herramienta Jupyter Notebook. Esta herramienta es muy útil en desarrollo de tutoriales web, pero la idea es que sólo sea accesible de forma sencilla la parte del código que sea de interés para el usuario del código. En este caso, sólo se dispone acceso a la lectura de datos, su procesado automático, el entrenamiento del modelo y la predicción sobre $D4$. Este formato es extremadamente cómodo para probar el modelo con datos distintos, pero no está pensado para incluir nuevos paquetes ni tratar con los algoritmos subyacentes. En concreto, encontramos que existe una incompatibilidad de una de las herramientas de PyTorch que utiliza EMC-EB y las librerías de SHAP y LIME. Aunque el objetivo inicial de este TFG era reproducir e interpretar los tres mejores métodos de TADPOLE, esta eventualidad lo hizo inviable.

Anexo III: Análisis detallado de reproducibilidad

A continuación se muestran las medidas de Precision, Recall y Accurcy para GB y RF en cada uno de los problemas de diagnóstico.

Método	$D1$ $D2$			$D1\_RedD3$ $D3$			$D1D2\_RedD4$ $D4$			$D1D2\_Aug$ $D4\_Aug$		
	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
Gradient Booster	98.7	98.3	98.7	90.1	90.5	89.9	88.7	85.2	81.4	95.3	97.1	96.2
Random Forest	90.8	92.0	92.2	68.2	63.2	64.5	64.8	61.1	61.9	90.1	92.6	91.4

Como se puede apreciar, GB excede ligeramente a RF en los problemas principales de este trabajo ($D1$vs$D2$, $D1D2_Aug$vs$D4_Aug$), pero muestra una diferencia mucho más importante en los secundarios ($D1_RedD3$vs$D3$, $D1D2_RedD4$vs$D4$). De hecho, la precisión de RF para los problemas secundarios es sorprendentemente baja en comparación con los principales. Los resultados originales de TADPOLE no presentan ninguna de estas medidas para los métodos participantes, por lo que no es posible hacer una comparativa. Del mismo modo, la versión GitHub de EMC-EB tampoco permite calcular estas métricas. En la Figura 11 se muestran las matrices de confusión referentes a las anteriores métricas.

En la siguiente tabla, se muestran los resultados obtenidos para los problemas principales de pronóstico de ADAS13. Según lo mostrado en [5], un predictor aleatorio es capaz de conseguir un MAE de 4.6 y WES de 4.5. Los resultados obtenidos no son sustancialmente mejores, ni por los métodos originales ni por los propios.

Método	$D1$ - $D2$		$D1D2\_Aug$ - $D4\_Aug$	
	MAE	WES	MAE	WES
Frog (GB)	-	-	4.85	4.74
Threedays (RF)	-	-	-	-
EMC-EB (SVMs)	-	-	6.75	6.66
Gradient Booster	2.41	2.37	7.86	7.82
Random Forest	1.18	1.41	25.71	25.53

Hiperparámetros de Gradient Booster

Dado un entrenamiento automático con optimización de los siguientes hiperparámetros para el GB, se muestran a continuación los hiperparámetros seleccionados y las matrices de confusión resultantes de cada prueba.

Hiperparámetro	Descripción	Valores	Selección
eta	Tiempo de acceso estimado	[0.01, 0.1, 0.5]	0.01
max_depth	Profundidad máxima de cada árbol	[5, 10, 200]	200
colsample_bytree	Sub-muestra de columnas al construir cada árbol	[0.5, 0.75, 1]	0.75
nrounds	Número de iteraciones de entrenamiento	[500, 1000]	500
subsample	Submuestra de datos a entrenar (recorte aleatorio de datos antes del primer árbol)	[0.5, 1]	0.5
lambda	Regularización L2 en los pesos	[0, 0.01, 0.05, 0.1]	0.01
alpha	Regularización L1 en los pesos	[0, 0.1, 0.5]	0

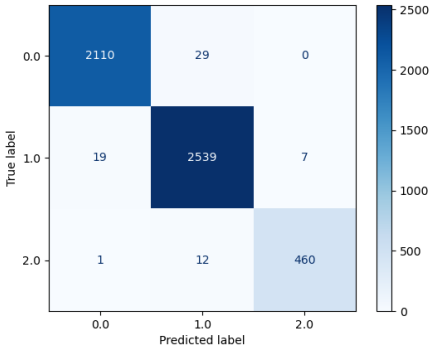
El rango de valores de cada hiperparámetro fue obtenido mediante un proceso experimental de prueba y error, tomando como referencia una serie de artículos divulgativos, métodos de Kaggle y recomendaciones de la documentación de XGBoost. Para su selección, se utilizó la herramienta GridSearchCV de la librería scikit-learn de Python, que realiza una búsqueda de hiperparámetros óptimos exhaustiva.

Hiperparámetros de Random Forest

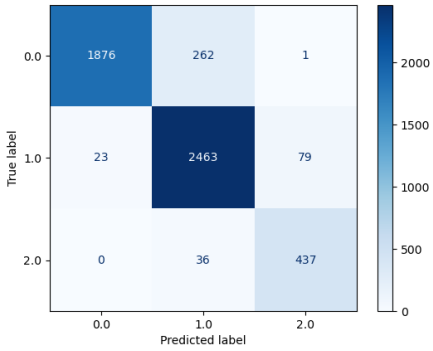
Dado un entrenamiento automático con optimización de los siguientes hiperparámetros para el RF, se muestran a continuación los hiperparámetros seleccionados y las matrices de confusión resultantes de cada prueba.

Hiperparámetro	Descripción	Valores	Selección
n_estimators	Número de árboles	[50, 100, 250, 500]	250
max_features	Número máximo de atributos por árbol	[10, 100, 1000, auto]	auto
max_depth	Profundidad máxima de cada árbol	[5, 10, 100, 200, 300]	200
criterion	Criterio de optimización	[gini, entropy]	entropy
min_samples_split	Mínimo de muestras necesarias para formar dos hojas	[2, 4, 8]	2
min_samples_leaf	Mínimo de muestras necesarias en cada hoja	[1, 2, 4]	1

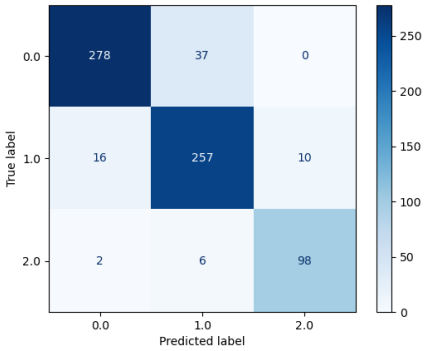
El rango de valores de cada hiperparámetro fue obtenido mediante un proceso experimental de prueba y error, tomando como referencia una serie de artículos divulgativos, métodos de Kaggle y recomendaciones de la documentación de RandomForestClassifier. Para su selección, se utilizó la herramienta GridSearchCV de la librería scikit-learn de Python, que realiza una búsqueda de hiperparámetros óptimos exhaustiva.



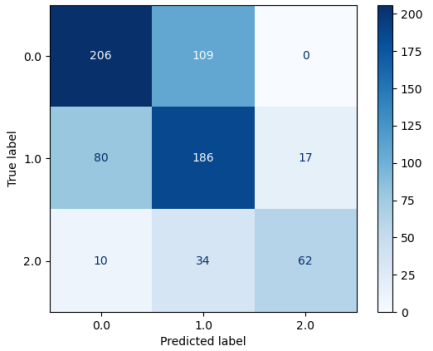
(a) GB, $D1$ vs $D2$



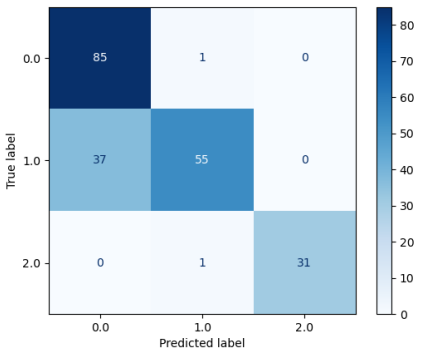
(b) RF, $D1$ vs $D2$



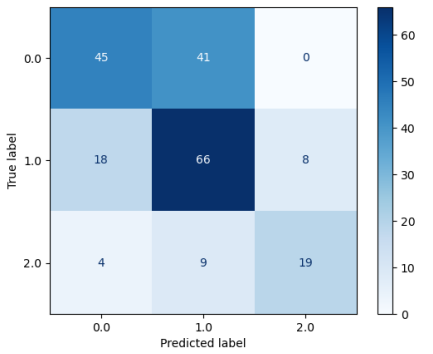
(c) GB, $D1\_RedD3$ vs $D3$



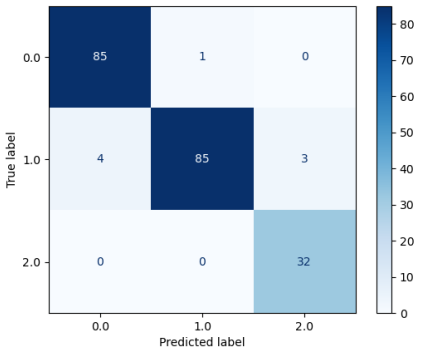
(d) RF, $D1\_RedD3$ vs $D3$



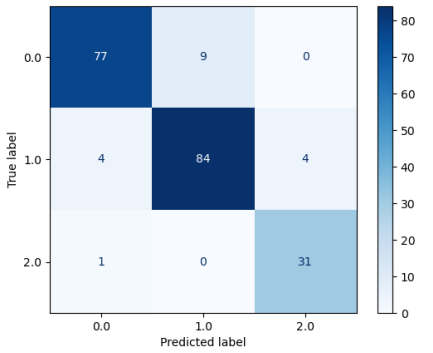
(e) GB, $D1D2\_RedD4$ vs $D4$



(f) RF, $D1D2\_RedD4$ vs $D4$



(g) GB, $D1D2\_Aug$ vs $D4\_Aug$



(h) RF, $D1D2\_Aug$ vs $D4\_Aug$

Figure 11: Matrices de confusión de GB y RF para cada problema. Izquierda, GB. Derecha, RF.

Anexo IV: Análisis detallado de la interpretabilidad

En este anexo se presentan una serie de cuestiones relacionadas con la interpetabilidad de los modelos que complementan los resultados mostrados en la memoria.

¿Cómo se interpretan las gráficas?

Los resultados de interpretabilidad proporcionados por SHAP y LIME se suelen presentan mediante unas gráficas de interpretabilidad cuya comprensión no resulta intuitiva. En este apartado explicamos cómo leerlas.

Resumen de valores de SHAP

Las gráficas de resumen de valores de SHAP presentan los atributos más relevantes para el modelo de cada clase predicha, en nuestro caso CN, MCI o AD. Estos atributos se presentan ordenados por su contribución total a clasificar una clase como probable en un problema one-vs-all. Se muestra la distribución de todos los datos del set (independientemente de si han sido clasificados como la clase que se esté analizando) utilizado para cada atributo 3 dimensiones:

- Dimensión Horizontal: El valor de SHAP de cada dato para ese atributo.
- Dimensión Vertical: La acumulación de puntos en ese valor de SHAP. Cuando dos o más datos toman el mismo valor de SHAP para un atributo sus puntos se sitúan en el mismo lugar horizontalmente y a igual distancia del eje horizontal verticalmente, a la mínima posible.
- Color: El valor del atributo en una escala de su mínimo a su máximo registrado.

Pongamos el caso de la Figura 12. En este caso estamos analizando los valores de SHAP de la clase CN para GB en el problema $D1$ vs $D2$. Primeramente, podemos ver que CDRSB_bl es el atributo más importante, porque aparece primero. El segundo sería CDRSB, el tercero DX_bl, etc. Tomemos CDRSB_bl. Como podemos apreciar, se generan dos 'burbujas': una, alrededor del valor de SHAP -0.35 (aprox.) y otra, alrededor del valor de 0.45 (aprox.). Tomemos la burbuja de la izquierda. Como sabemos, lo que estamos viendo no es más que la acumulación de un número muy grande de puntos. Por tanto, sabemos que hay unos pocos datos con un valor de -0.1, algunos más con -0.2, bastantes con -0.3, muchos con -0.35, bastantes con -0.4, y unos pocos con -0.5. Esto es interesante porque podemos establecer que, dentro de estos datos, estamos viendo que todos ellos tienen un valor de SHAP negativo, es decir, nos indican que es menos probable que uno de estos datos sea clasificado como CN por nuestro GB. Pero, ¿por qué sucede esto? La

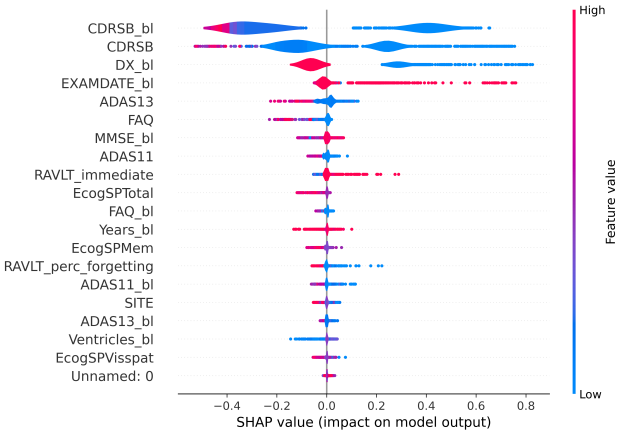


Figure 12: Resumen de valores de SHAP para el GB en el problema $D1$ vs $D2$

explicación la podemos encontrar si nos fijamos en el color de la gráfica. Como podemos apreciar, el extremo izquierdo de la acumulación es de color rojo. Esto significa que los datos con un valor de CDRSB_bl relativamente alto, al obtener valores SHAP muy negativos, predicen de forma intensa que un dato con ese valor no será clasificado como CN. Si nos fijamos, el color se torna de rojo a azul progresivamente conforme los valores de SHAP suben, implicando una fuerte correlación negativa entre el valor de CDRSB_bl y la probabilidad de ser clasificado como CN. Además, esto nos indica que hay muchos datos (y por tanto muchos pacientes) que tienen un valor bastante bajo de CDRSB_bl.

Por contra, si nos fijamos en la burbuja de la derecha vemos que es de color azul claro y que se reparte entorno al valor SHAP de 0.45 (aprox.), implicando que un bajo valor de CDRSB_bl es un buen indicador de que el dato será clasificado como CN. Ocupar el segundo lugar en la gráfica indica que no es tan bueno como un valor alto de CDRSB_bl para indicar que no será clasificado como CN.

Este tipo de gráficas no sólo se usan individualmente, del modo que hemos comentado sino también comparativamente. Por ejemplo, si sabemos que un alto valor de CDRSB_bl implica que un dato no será clasificado como CN, será de nuestro interés comprobar cómo de relevante resulta un alto valor de CDRSB_bl para clasificar un dato como AD. En nuestro caso, mediante la Figura 3c podemos comprobar que CDRSB_bl ocupa el segundo puesto en importancia para el diagnóstico AD. No obstante, un alto valor sigue teniendo cierta correlación con ser clasificado como AD. Por tanto, podemos establecer que CDRSB_bl es un atributo que funciona muy bien como discriminante entre personas sanas (CN) y no sanas (MCI y AD).

Interpretación Local con LIME

Las gráficas de interpretación de LIME analizan qué atributos han sido más relevantes para la clasificación de un dato en cada clase (CN, MCI o AD) en un problema one-vs-all. La diferencia

con SHAP es que permite establecer un estudio individual de cada dato, viendo la contribución de cada atributo con valores concretos. Por ello, se utiliza habitualmente LIME para interpretar aquellos datos que el clasificador ha fallado al clasificar, ya que permite hacer un estudio de cuál es el motivo del error.

En estas gráficas los atributos se presentan ordenados por su contribución a la predicción para una determinada clase en la dimensión vertical, mientras que en la dimensión horizontal se muestra la contribución a la predicción. En cuanto más a la izquierda más negativo, y por tanto más contribuye a establecer que el dato no será clasificado en la clase que se está estudiando; en cuanto más a la derecha más positivo, y por tanto más contribuye a clasificar el dato en la clase estudiada. El color en este caso es puramente estético. Cabe destacar que en ciertos casos en los que el valor de un atributo sea un entero pequeño o un booleano, este será presentado junto al nombre del atributo tal que (atributo=valor).

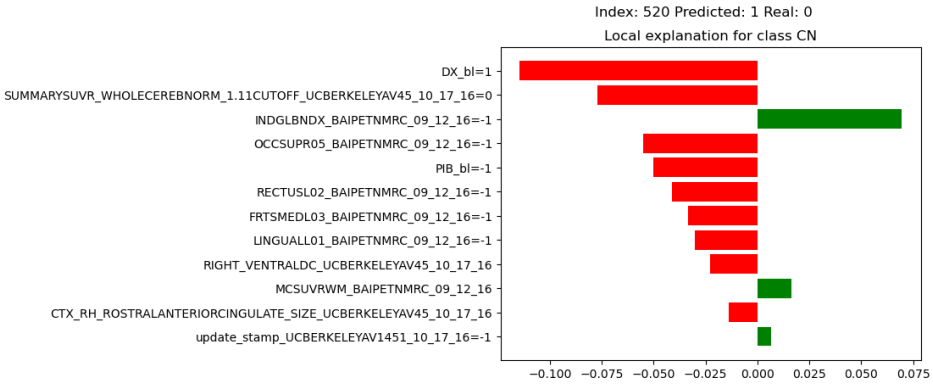


Figure 13: Interpretación con LIME de un dato mal clasificado por el GB en el problema $D1$ vs $D2$

Ejemplifiquemos esto mediante la figura 13. En ella, estamos interpretando un caso en el que el GB ha tomado un dato del problema $D1$ vs $D2$ cuya clase real era MCI y lo ha clasificado como AD. Utilizando LIME, mostramos la interpretación de la clase real (MCI) para este dato, para poder ver qué atributos contribuyen negativamente y, por tanto, que hacen al clasificador fallar. Primero, podemos ver de arriba a abajo el orden de los atributos según su contribución (update_stamp_UCBERKELEY, CDRSB, ADAS13...). Si tomamos el más relevante, update_stamp_UCBERKELEYAV1451_10_17_16, podemos apreciar que su contribución es muy positiva (+0.1 para la clase MCI). Esto debe llamar nuestra atención, pues el valor del atributo es explicitado junto con el atributo. Su valor es -1. Esto implica que es un caso de atributo faltante, y el clasificador ha aprendido que los datos faltantes son un buen indicativo de ser clasificado como MCI. Esto nos señala una flaqueza de los datos de entrenamiento, y es que al ser un set no muy grande es posible aprender qué atributos aparecen mayoritariamente en ciertos pacientes, a pesar de no ser indicadores reales de la presencia de Alzheimer.

Pero la parte interesante es la de los atributos posteriores. En este caso, vemos como CDRSB, ADAS13 y FAQ nos indican una probabilidad negativa de pertenecer a la clase MCI. Esto se debe a que el paciente, para estos test cognitivos, presenta valores excesivamente altos para MCI en

comparación con otros pacientes y, al ser estos tres atributos muy buenos indicadores de AD, han llevado al clasificador a elegir dicha clase. Esto nos hace pensar que un mayor número de ejemplos en los datos de entrenamiento con MCI y AD podrían llevar al clasificador a distinguir mejor entre estas clases o a no dejarse llevar por malas puntuaciones en los test cognitivos, pues no son un indicador infalible de diferencia entre MCI y AD.

Otros recursos

Adjuntamos aquí varios artículos en los que se enseña a comprender los resultados proporcionados por SHAP y LIME.

SHAP

- https://christophm.github.io/interpretable-ml-book/shap.html
- https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values
- https://mlconference.ai/blog/tutorial-explainable-machine-learning-with-python-and-shap/

LIME

- https://homes.cs.washington.edu/~marcotcr/blog/lime/
- http://gael-varoquaux.info/interpreting_ml_tuto/content/02_why/04_black_box_interpretation.html

Atributos de interés

Se ha obtenido mediante SHAP, para cada uno de los clasificadores en el problema $D1$ vs $D2$, un ranking de los atributos más relevantes de manera total, calculando la relevancia como

$$\frac{\sum_{i=0}^{D}\sum_{j=0}^{N}|ShapValue(i,j)|}{N} \tag{29}$$

siendo D el número de datos en $D2$ y N el número de clases (CN, MCI, AD). A partir de dicho ranking se han obtenido las posiciones que ocupan ciertos atributos de interés clínico, en cada uno de los clasificadores.

El valor de cada atributo representa su posición en una clasificación según sus valores de SHAP, es decir, a menor posición mayor importancia. Como se puede apreciar, RF tiene una tendencia

Table 1: Ranking de atributos según sus valores de SHAP

Método	APOE4	FDG	PIB	AV45	CDRSB	ADAS11	ADAS13	MMSE
Gradient Booster	319	191	259	435	2	54	93	91
Random Forest	51	171	1095	920	3	10	7	9

Método	Ventricles	HippoC.	WholeBrain	Ethorinal	Fusiform	Temp_Lobe
Gradient Booster	219	67	286	42	337	307
Random Forest	138	60	174	64	151	64

mucho mayor a dar importancia a los atributos más habituales en la práctica clínica. Hemos podido apreciar anteriormente esto se debe a que el GB utiliza de forma más intensa los atributos de imagen computacional para los casos de MCI dudoso, por lo que los atributos de test cognitivo toman menos importancia general al sólo estar entre los principales en 2 de 3 clases para el GB.

Entrenamiento sobre $D1$, Predicción sobre $D2$

A continuación se presentan algunos datos recavados adicionalmente con SHAP y LIME para el problema $D1$ vs $D2$. En concreto, se resumen los atributos principales elegidos por ambos clasificadores en la Figura 14, y un caso de buena clasificación para cada predictor en la Figura 15.

Gradient Booster

En la Figura 14a podemos apreciar el efecto de los atributos principales de $D1$ vs $D2$. Es muy destacable el impacto general de los atributos de imagen computacional, que supera a prácticamente cualquier otra medición. Esto se debe a que su efecto sobre la clase 1 (MCI) es increíblemente superior a las demás, haciendo que parezca que es la única clase que se está analizando dada la diferencia en los valores de SHAP.

En la Figura 15a podemos apreciar un caso de paciente bien clasificado por el GB en el que se interpreta la clasificación con LIME. En este caso, el paciente se clasifica como MCI. Para ello utiliza algunos datos de imagen que no habían aparecido anteriormente. El clasificador toma como buenos indicadores para padecer MCI un diagnóstico base de MCI, la puntuación de imagen PET del giro cingulado anterior izquierdo, la imagen PET del precúneo izquierdo y derecho, y otras imágenes PET como el frontal o el angular. Por otro lado, toma como malos indicadores de MCI la imagen PET del tálamo derecho, la imagen PET del parahipocampo derecho, el giro cinguldo medio y algunos datos faltantes, con el flag -1.

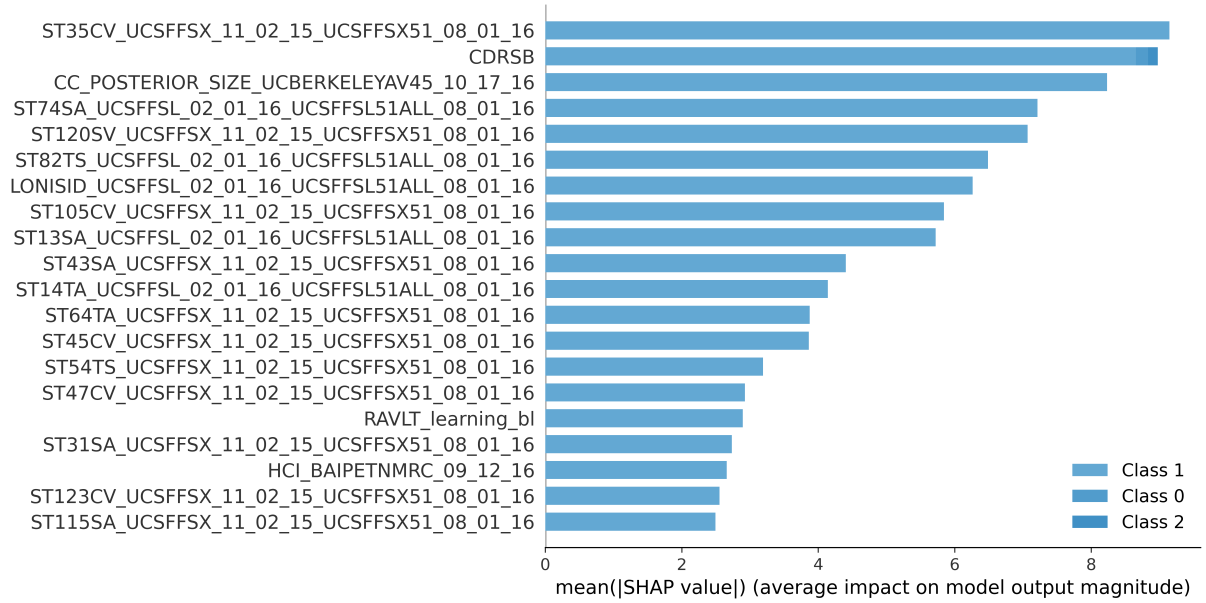
Random Forest

En la Figura 14b podemos apreciar el efecto de los atributos principales de $D1$ vs $D2$. En comparativa con el GB, es claro como el RF utiliza más los elementos de forma proporcionada, dando especial relevancia a los test cognitivos como FAQ o ADAS.

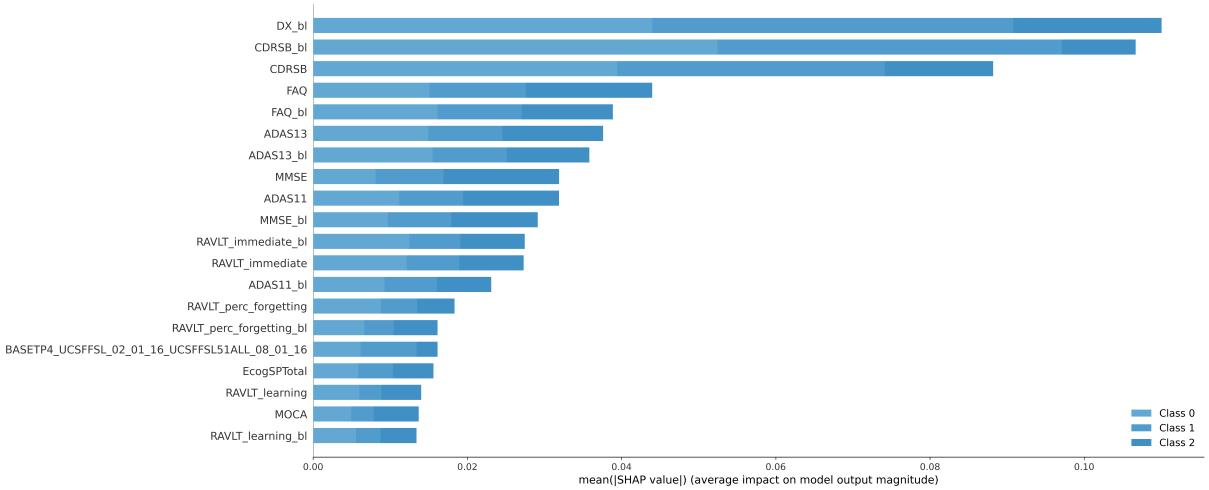
En la Figura 15b podemos apreciar un caso de paciente bien clasificado por el RF en el que se interpreta la clasificación con LIME. En este caso podemos ver cómo RF está dando mucha importancia a una serie de atributos de imagen computacional que no habíamos visto anteriormente. Esto es interesante porque nos confirma que, aunque el RF tiende menos a utilizar estos atributos, también les da importancia en el caso de clasificación MCI. Al mismo tiempo, podemos percatarnos de que varios de estos datos tienen el valor -1, por lo que el RF también ha aprendido a relacionar la falta de un dato con su diagnóstico. Descontando esto, se seleccionan como buenos indicadores de MCI un diagnóstico base de MCI, una potencia de campo magnético de 1 en imagen DTI, el volumen del Occipital Inferior Izquierdo y el del Occipital Medio Derecho, junto con varios datos de imagen PET. Por otro lado, se determinan como malos indicadores la puntuación de Study Partner Ecog, el índice de convergencia hipermetabólica y la imagen PET de la ínsula derecha.

Entrenamiento sobre $D1D2\_Aug$, Predicción sobre $D4\_Aug$

Finalmente, se resumen los atributos principales elegidos para el problema $D1D2\_Aug$ vs $D4\_Aug$ por ambos clasificadores en la Figura 16.

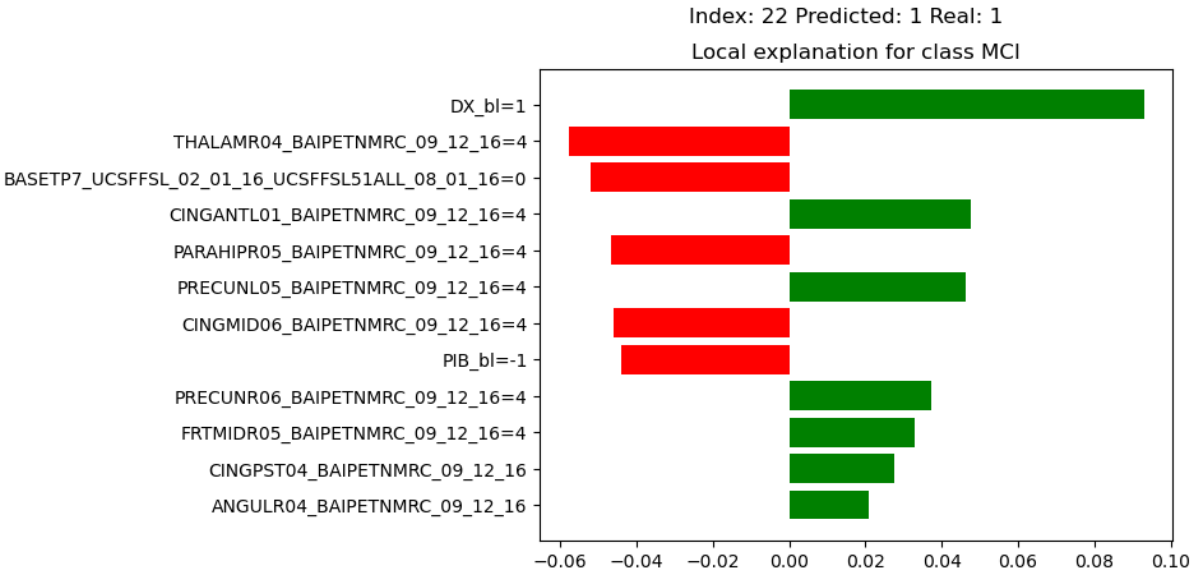


(a) Relevancia de los atributos principales de GB $D1$ vs $D2$

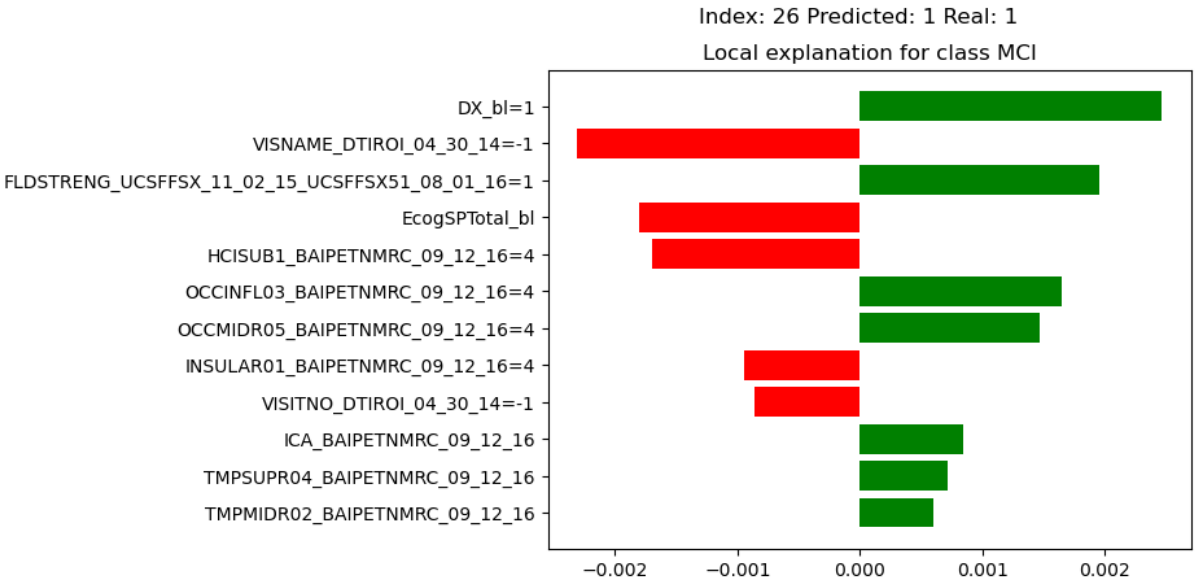


(b) Relevancia de los atributos principales de RF $D1$ vs $D2$

Figure 14: Atributos según relevancia para GB y RF para el problema $D1$ vs $D2$

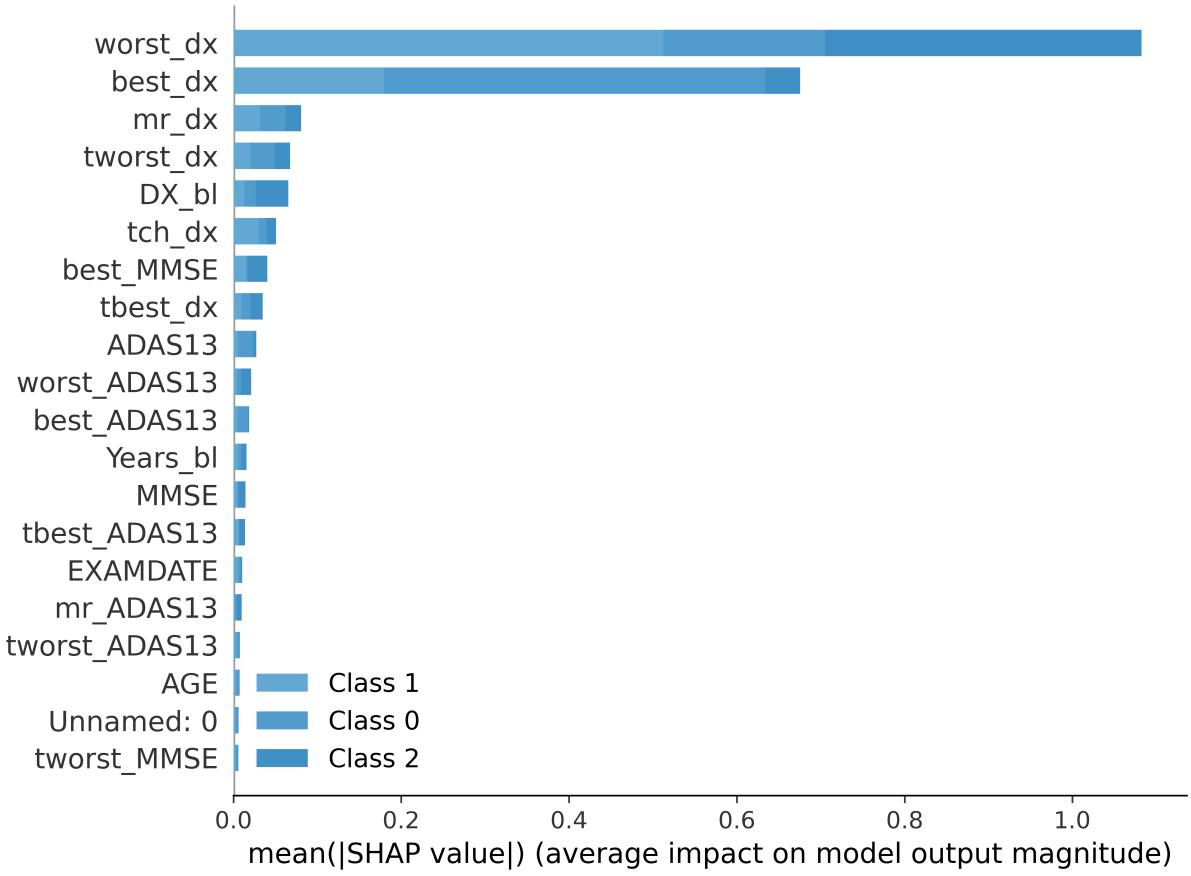


(a) Interpretación con LIME de un paciente bien clasificado por GB en el problema $D1$ vs $D2$

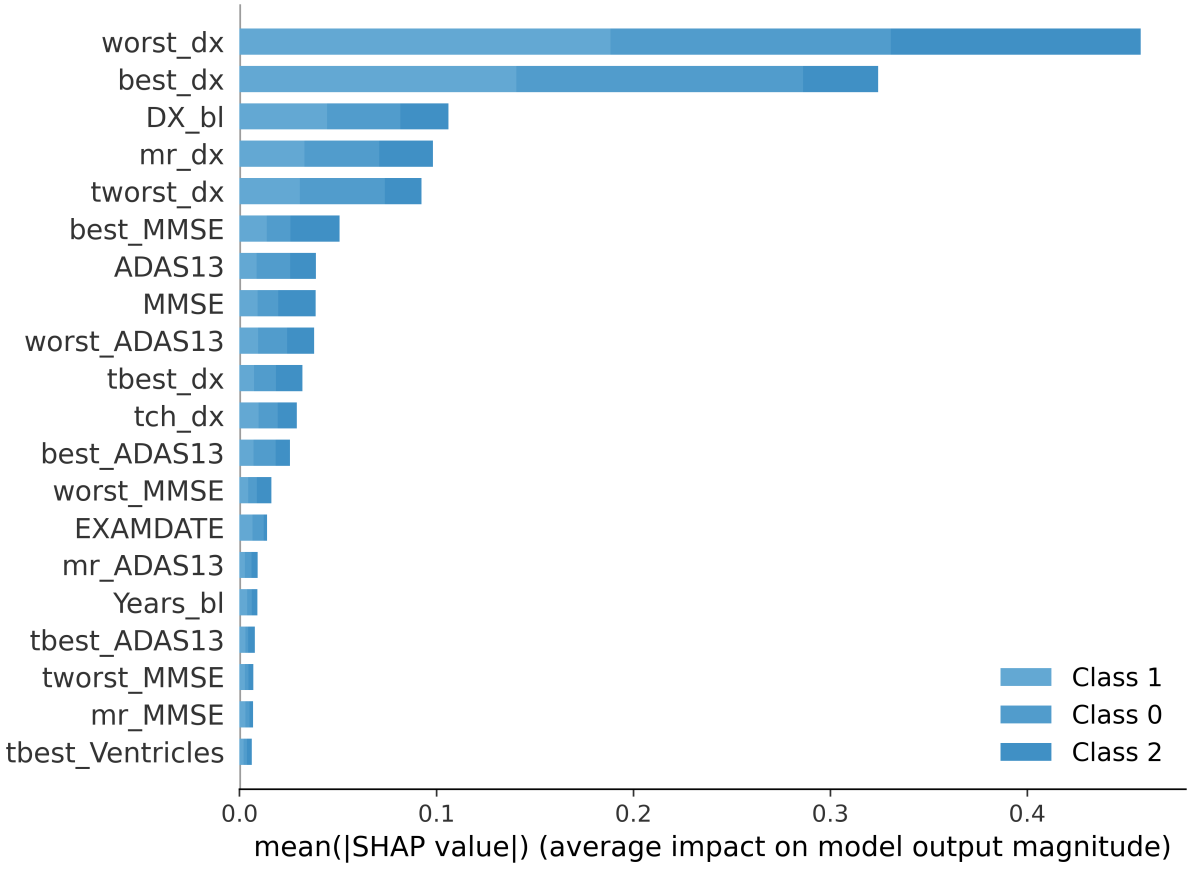


(b) Interpretación con LIME de un paciente bien clasificado por RF en el problema $D1$ vs $D2$

Figure 15: Interpretaciones con LIME de pacientes bien clasificados como MCI



(a) Relevancia de los atributos principales de GB $D1D2\_Aug$ vs $D4\_Aug$



(b) Relevancia de los atributos principales de RF $D1D2\_Aug$ vs $D4\_Aug$

Figure 16: Atributos según relevancia para GB y RF para el problema $D1D2\_Aug$ vs $D4\_Aug$

Referencias

[1] A. Nordberg and A. Svensson, "Cholinesterase Inhibitors in the Treatment of Alzheimer's Disease," *Drug-Safety*, vol. 19, no. 1, pp. 465–480, 1998.

[2] R. Khoury and E. Ghossoub, "Diagnostic biomarkers of Alzheimer's disease: A state-of-the-art review," *Biomarkers in Neuropsychiatry*, vol. 1, p. 100005, 2019.

[3] J. K. e. a. Kueper, "The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review." *Journal of Alzheimer's disease : JAD*, vol. 63, no. 2, pp. 423–444, 2018.

[4] I. e. a. Arevalo-Rodriguez, "Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)." *The Cochrane database of systematic reviews*, vol. 2015, no. 3, 2015.

[5] R. V. M. et al., "The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up," 2020.

[6] G. J. Moore PJ, Lyons TJ, "Random forest prediction of Alzheimer's disease using pairwise selection from time series data," *PLoS ONE*, vol. 14, no. 2, 2019.

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 08 2016.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, p. 5–32, 2001.

[9] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[11] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[13] R. C. e. a. Petersen, "Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization." *Neurology*, vol. 74, no. 3, pp. 201–9, 2010.

[14] R. V. M. et al. and the EuroPOND Consortium, "TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease," 2018.