UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Desarrollo de biomarcadores de imagen médica para ELA (Esclerosis Lateral Amiotrófica) con técnicas estadísticas multivariantes y de machine learning

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Gil Chong, Pablo Olayo

Tutor/a: Carot Sierra, José Miguel

Director/a Experimental: CERDA ALBERICH, LEONOR

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Desarrollo de biomarcadores de imagen médica para ELA (Esclerosis Lateral Amiotrófica) con técnicas estadísticas multivariantes y de machine learning

TRABAJO FIN DE GRADO

Grado en Ciencia de Datos

Autor: Pablo Gil Chong

Tutor: José Miguel Carot Sierra

Curso 2021-2022

Resumen

La Esclerosis Lateral Amiotrófica es una enfermedad degenerativa de motoneurona caracterizada por su dificultad de diagnóstico: más del 90 % de los casos son esporádicos y no existe ninguna prueba paraclínica fiable capaz de detectarla. Es por ello que urge el desarrollo de biomarcadores que permitan el diagnóstico y la monitorización.

En este trabajo, se ha utilizado un conjunto de datos de 211 pacientes (114 ELA, 45 mimic, 30 portadores de mutación y 22 control) con atributos de radiómica (morfometría, depósitos de hierro) integrados con variables clínicas y con 6 variables de valoración semicuantitativa de depósitos de hierro.

Se ha enfocado el problema como una tarea de clasificación binaria entre pacientes con y sin ELA. Se ha seguido una metodología de modelado secuenciada abordada desde una perspectiva de mejora iterativa con técnicas de filtrado de variables, reducción de dimensionalidad (PCA, kernel PCA), sobremuestreo (SMOTE, ADASYN) y clasificación (regresión logística, LASSO, Ridge, ElasticNet, Support Vector Classifier, K-vecinos, random forest). Para cada arquitectura propuesta, se ha utilizado varios subconjuntos de los datos disponibles, planteando modelos con un solo conjunto de datos y modelos multimodales.

Los mejores resultados han sido proporcionados por un clasificador de votación compuesto por cinco clasificadores: accuracy=0.896, AUC=0.929, sensitividad=0.886, especificidad=0.929. Los mejores resultados sin uso de las variables semicuantitativas han sido proporcionados por Support Vector Classifier: accuracy=0.815, AUC=0.879, sensitividad=0.833, especificidad=0.794. En ambos clasificadores se ha utilizado un filtrado de variables por feature importance en LASSO.

Finalmente, se ha propuesto un prototipo para integrar la metodología existente necesaria para la obtención de los datos junto a la metodología desarrollada en el trabajo, orientada para diagnóstico y pronóstico en entornos clínicos.

Palabras clave: ELA, biomarcador, radiómica, modelado iterativo, modelo multimodal

Abstract

Amyotrophic Lateral Sclerosis is a degenerative motor neuron disease characterized by its diagnostic difficulty: more than 90% of cases are sporadic and there is no reliable paraclinical test capable of detecting it. The development of ALS biomarkers for diagnosis and monitoring is urgently needed.

This work has used a dataset of 211 patients (114 ALS, 45 mimic, 30 genetic carriers and 22 control) with radiomics attributes (morphometry, iron deposition) integrated with clinical variables and 6 semiquantitative visually-assessed indicators of iron deposition.

A binary classification task approach has been taken to classify patients with and without ALS. A sequential modeling methodology, understood from an iterative improvement perspective, has been followed. It has included variable filtering techniques, dimensionality reduction techniques (PCA, kernel PCA), oversampling techniques (SMOTE, ADASYN) and classification techniques (logistic regression, LASSO, Ridge, ElasticNet, Support Vector Classifier, K-neighbors, random forest). For each proposed architecture, several subsets of the available data have been used, proposing models with single data types and multimodal models.

The best results have been provided by a voting classifier composed of five classifiers: accuracy=0.896, AUC=0.929, sensitivity=0.886, specificity=0.929. The best results without the use of semiquantitative variables have been provided by Support Vector Classifier:

accuracy=0.815, AUC=0.879, sensitivity=0.833, specificity=0.794. In both classifiers a filtering of variables by feature importance in LASSO has been used.

Finally, a prototype has been proposed to integrate the existing methodology necessary to obtain the data together with the methodology developed in this work, oriented to diagnosis and prognosis in clinical settings.

Key words: ALS, biomarker, radiomics, iterative modelling, multimodal model

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

La Esclerosis Lateral Amiotrófica, abreviada ELA, es un grupo de enfermedades degenerativas del sistema nervioso central. La ELA afecta variablemente a las neuronas del sistema nervioso encargadas de provocar contracciones musculares (motoneuronas) de la corteza cerebral (motoneuronas superiores, abreviadas MNS) y del encéfalo y médula espinal (motoneuronas inferiores, abreviadas MNI). Al dejar de recibir señales, el sistema muscular se ve progresivamente atrofiado y la autonomía de la persona se ve vulnerada. Una vez comienzan los síntomas, la supervivencia es de 2 a 3 años en promedio, causando muerte por insuficiencia respiratoria en la mayoría de los casos.

Generalmente, la ELA es una enfermedad cuya detección es complicada. Esto se debe a diversos motivos. El primero y más señalado es la aparición esporádica (ELA esporádica) de la enfermedad en la mayoría de casos. En la literatura, se estima que las pacientes de ELA con antecedentes familiares (ELA familiar) representan tan solo entre el 5 y el 10 % de los diagnósticos totales. Consecuentemente, en más del 90 % de los casos no existe indicio de enfermedad hasta, como pronto, la aparición de los primeros síntomas, ralentizando así el inicio del tratamiento.

A la dificultad anterior se le deben sumar muchas otras. Entre estas, se incluye: la compleja interacción de factores fisiopatológicos para dictaminar la aparición de la enfermedad, los diversos cambios histológicos causados por la enfermedad en proporciones variables, o la heterogeneidad de la expresión fenotípica. Asimismo, las personas portadoras de las mutaciones genéticas causantes de ELA familiar tampoco terminan desarrollando la enfermedad en todos los casos.

Actualmente, el diagnóstico de ELA es principalmente clínico, y se adecúa a los criterios revisados de El Escorial [1]. En ellos se establece que el diagnóstico requiere de, al menos, una de las siguientes: pruebas de degeneración de MNI obtenidas por examinación clínica, electrofisiológica o neuropatológica, pruebas MNS obtenidas por examinación clínica o una expansión progresiva de síntomas dentro de una región o de una región a otras. El diagnóstico de ELA, además, se trata de un diagnóstico por exclusión: lo anterior debe ser acompañado de la ausencia de evidencia de procesos asociados a otras enfermedades. Esto responde al desconocimiento de las causas exactas de la ELA.

La figura 1.1 muestra un diagrama con los criterios para el diagnóstico de El Escorial [1].

No obstante, algunos estudios señalan que la utilidad de este sistema de diagnóstico es limitada. Esto es debido a la sensibilidad limitada con que cuentan como consecuencia de la falta de signos de MNS o MNI en la presentación clínica [2, 3].

Si bien es cierto que existen también técnicas de electrodiagnóstico para valorar cuantitativamente la afectación de la MNI, esto no es así para la MNS, cuya valoración de-

Figura 1.1: Criterios de El Escorial para el diagnóstico de ELA. Fuente: Elaboración propia.

pende enteramente de métodos clínicos. El tratamiento suele retrasarse ante la falta de un diagnóstico temprano y una vez se detectan los síntomas la enfermedad suele llevar entre 10 y 16 meses desarrollándose [4, 5].

Una prueba paraclínica basada en biomarcadores permitiría solucionar, o al menos aliviar, esta problemática para la valoración de la afectación de la MNS.

Por biomarcador se entiende üna característica definida que se mide como indicador de procesos biológicos normales, procesos patógenos o respuestas a una exposición o intervención, incluidas las intervenciones terapéuticas"[6]. Es decir, que se trata de cualquier característica mesurable a la que se puede dar uso como indicador subrogado de cualquier proceso biológico. En la figura 1.2 se muestra un diagrama que ilustra su uso.

Los biomarcadores de imagen médica son principalmente extraídos a través de métodos computacionales aplicados a las imágenes. Las características se obtienen en los conjuntos de vóxeles y se pueden representar en el espacio, como mapas paramétricos, y en el tiempo, analizándose en estudios longitudinales.

Los biomarcadores pueden ser de varios tipos, según su función. Se distingue entre biomarcadores diagnósticos (para detectar una enfermedad concreta o la presencia de una determinada sustancia), predictivos (para predecir la respuesta a un tratamiento o el resultado de una enfermedad) y pronósticos (para determinar la probabilidad de un evento clínico o el tiempo hasta que suceda) [8].

En el caso de la ELA, el desarrollo de biomarcadores es prioritario debido a las dificultades que plantea la metodología clínica actual junto a la naturaleza de la propia enfermedad. De obtener biomarcadores no invasivos y fiables, se podría conseguir no solamente mejorar el proceso diagnóstico, sino también mejorar la comprensión actual de los mecanismos fisiopatológicos de la enfermedad, caracterizar fenotípicamente a las personas pacientes y monitorizar el tratamiento y la progresión de la enfermedad.

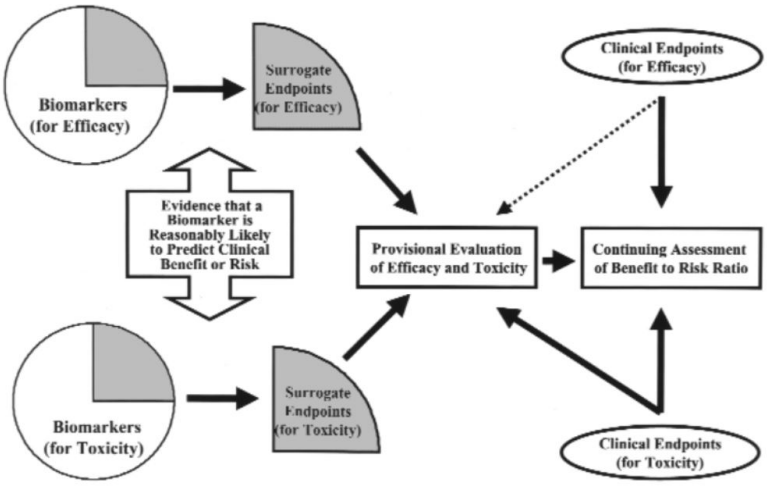

Figura 1.2: Modelo conceptual de la relación entre biomarcadores, procesos surrogados e intervención terapéutica. Fuente: [7].

En este trabajo, se va a tratar de desarrollar biomarcadores de imagen médica para tratar de mejorar la comprensión de la ELA. Esto se hará a partir de datos clínicos y datos de imagen médica de cuatro tipos distintos obtenidos a partir de imágenes de resonancia magnética (RM): datos de volumetría, datos de espesor cortical, datos de acumulación de hierro (ferritina) cerebral y datos de análisis visual. El conjunto de datos utilizado ha sido proporcionado por el Hospital La Fe, y se ha desarrollado en el marco de una beca de colaboración con el departamento de Estadística e Investigación Operativa Aplicadas y Calidad.

1.1 Motivación

La ELA es una enfermedad con una incidencia relativamente elevada, asociada parcialmente al envejecimiento y con una tasa de mortalidad de prácticamente 100 %. Cada vez son más las personas que la padecen, y a medida que la sociedad continúe envejeciendo, este número no hará sino aumentar. A pesar de ello, el progreso en avances son limitados: apenas se ha conseguido frenar el proceso de evolución de la enfermedad y, por descontado, no se ha logrado detener o revertir. Las personas afectadas sufren cambios irreversibles en sus vidas y su esperanza de vida se acorta considerablemente desde el diagnóstico de la enfermedad. Puesto que con el conocimiento disponible no se puede prevenir la enfermedad –visto está que incluso se diagnostica por exclusión–, es necesario continuar realizando avances médicos en pos de frenarla.

El uso diagnóstico que se realiza hoy de imágenes de RM para MNS es eminentemente clínico, utilizando la valoración de la persona especialista para determinar si existe afectación o no y, en caso de que sí, en qué grado. Con este trabajo, en el que se va a trabajar en el desarrollo de biomarcadores, se está buscando comprobar en qué medida puede replicarse el diagnóstico humano con parámetros objetivos. Además, se va a utilizar metodología prácticamente inexplorada, cuya utilidad práctica potencial no se verá limitada al diagnóstico, sino que también podría ser un avance hacia una mejor comprensión de la enfermedad.

Finalmente, refiriéndome a intereses estrictamente personales, mi preferencia por la ciencia de datos como campo nació por sus potenciales aplicaciones para la búsqueda del bien común y la mejora de las condiciones de vida humanas. Más específicamente, terminar trabajando en aplicaciones médicas de la ciencia de datos fue mi principal

motivación para inscribirme en el Grado en Ciencia de Datos en primera instancia. Este trabajo podría servir como piedra angular para seguir acercándome a dicho objetivo.

1.2 Objetivos

El objetivo de este trabajo es analizar los indicadores de volumetría cerebral, concentraciones de hierro y espesor cortical obtenidos a partir de imágenes de resonancia magnética (RM) junto a las variables clínicas disponibles para contribuir al desarrollo de biomarcadores que permitan pronosticar la ELA o ayudar a diseñar estrategias terapéuticas, mejorando la comprensión de la actuación de la enfermedad sobre el cuerpo. El proceso de obtención de estos biomarcadores será a través de técnicas estadísticas multivariantes y de machine learning y de forma que los resultados puedan ser empleados adecuadamente por los profesionales médicos, particularmente para uso diagnóstico. Es posible desagregar este objetivo en los siguientes:

- Caracterización de tipologías de paciente. Discernir patrones en los indicadores de imagen característicos de las tipologías de paciente a través de técnicas exploratorias y de modelado.
- Clasificación de pacientes. Proponer distintos métodos de modelado con técnicas estadísticas y de machine learning explicables utilizando diferentes subconjuntos de datos para tratar de predecir la tipología del paciente adecuadamente. Sobre las técnicas de modelado:
  - Identificar potenciales biomarcadores pronóstico, interpretando los modelos.
  - Abordar el problema como una tarea de clasificación con una perspectiva de mejora iterativa, observando los errores cometidos por una iteración planteada y tratando de corregirlos en la siguiente.
  - Atender a distintas métricas, encontrando un balance entre errores falsos negativos y falsos positivos.
- Accesibilidad. Plantear enfoques metodológicos accesibles para un público médico no necesariamente familiarizado con las técnicas de análisis y modelado utilizadas.
- Aplicabilidad. Diseñar un prototipo de inferfaz que permita a los profesionales médicos replicar las partes más relevantes del análisis para ayudar al diagnóstico de casos nuevos.

1.3 Impacto esperado

Se espera extraer conocimiento para mejorar la comprensión de la enfermedad: como se verá en los dos próximos capítulos, actualmente apenas hay estudios que utilicen conjuntos de datos similares al disponible. No obstante, sería sorprendente que las posibles pruebas paraclínicas desarrolladas a partir de este trabajo permitiesen una mejoría notable en la predicción de la enfermedad. Esta visión quizá pesimista en apariencia responde, principalmente, al bajo número de casos con los que se cuenta (este problema es compartido con la amplia mayoría de estudios similares, y se hará hincapié sobre ello más adelante). Un número de casos reducido lleva a modelos con necesidad de generalización elevada para no caer en sobreajustes, y que terminan inevitablemente siendo de potencia limitada, aportando información más orientativa sobre la naturaleza y síntomas de la enfermedad, pero no reglas de decisión sólidas para utilizar en un diagnóstico real.

Por tanto, con este trabajo se espera corroborar la información teórica ya conocida sobre la enfermedad (algo apenas desarrollado), así como expandir el conocimiento actual sobre la enfermedad. Podría ser algo optimista, pero quizá razonable, pensar que el conocimiento extraído podría servir para abrir nuevas líneas de investigación médicas futuras.

1.4 Estructura de la memoria

La memoria cuenta con ocho capítulos:

1. Introducción. Presenta un contexto general acerca de la ELA y del potencial del uso de biomarcadores y sintetiza qué se espera del trabajo: motivación, objetivos e impacto esperado.
2. Antecedentes y estado del arte. Resume los antecedentes y aborda y critica el estado del arte actual, focalizando en trabajos con técnicas propias de ciencia de datos.
3. Análisis del problema. Analiza el problema a abordar: presenta el conjunto de datos y describe su obtención, plantea y justifica la metodología, plantea la implementación, lista los recursos utilizados y explica el adecuamiento al marco legal.
4. Preparación y comprensión de los datos. Detalla el preprocesado de los datos y resume los hallazgos principales del análisis exploratorio.
5. Modelado: evaluación y conocimiento extraído. Cada sección de este capítulo describe y justifica los cambios adoptados al diseño para adaptarlo a las nuevas propuestas de modelado y resume e interpreta los resultados de los mejores modelos con dicha propuesta de modelado. Las secciones siguen un orden cronológico con un sentido narrativo acorde a los descubrimientos sobre la naturaleza de los datos.
6. Prototipo de aplicación. Plantea un prototipo de aplicación de utilidad clínica para aplicar el conocimiento adquirido.
7. Conclusiones. Sintetiza los hallazgos más importantes del trabajo.
8. Trabajos futuros. Propone ampliaciones y propuestas de investigación futuras.

CAPÍTULO 2
Antecedentes y estado del arte

2.1 Antecedentes

2.1.1. ELA

Causas

Como se adelantaba en la introducción, el conocimiento de la enfermedad es limitado. No obstante, los avances son cada vez más y llegan más rápido. Desde el descubrimiento del primer gen ligado a la ELA en 1993 [9], se ha descubierto más de 120 variaciones genéticas [10, 11]. El ritmo de estos descubrimientos parece seguir una tendencia exponencial, como se aprecia en la figura 2.1. Algunos estudios longitudinales también muestran que en tiempos recientes se ha mejorado el trato de la atención sintomática de la enfermedad [12], aunque ello podría ser resultado de los avances generales en las ciencias de la salud.

Con todo, a día de hoy sigue sin poder pronosticarse con exactitud el inicio de la enfermedad. Esto es cierto tanto para la ELA familiar, que cuenta con una ventana de edad posible muy amplia [13] a pesar de conocerse variantes genéticas portadoras [11], como para la ELA esporádica, para la cual, pese a conocerse varios factores de influencia [15], aún no se saben causas exactas. Dichos factores incluyen susceptibilidad genética [16], causas ambientales, infección viral persistente o autoinmunidad [15]. Actualmente la literatura ha adoptado un enfoque genómico y se centra principalmente en estudiar la susceptibilidad genética [17] al ser el factor más controlable.

Es importante comentar la dirección generalmente adoptada por la literatura en tiempos recientes. Sin embargo, recordamos que el estudio de las causas de la ELA no es el foco de este trabajo, cuyos objetivos pasan por analizar la sintomatología presentada a través de biomarcadores de imagen, y no genéticos.

Imagen médica

En la literatura, las imágenes RM y las PET (tomografía por emisión de positrones) han sido capaces de revelar en numerosos estudios patrones funcionales y estructurales en la ELA, que se cree que representan el sello patológico de la enfermedad. El avance neurodegenerativo y los síntomas de la ELA se están estudiando a un nivel más fino y detallado gracias al continuo desarrollo de herramientas de neuroimagen [18]. Uno de los puntos de mayor interés en su uso es que se trata de técnicas no invasivas que potencialmente podrían permitir la investigación patofisiológica de la enfermedad y la monitorización del paciente con fines clínicos [19].

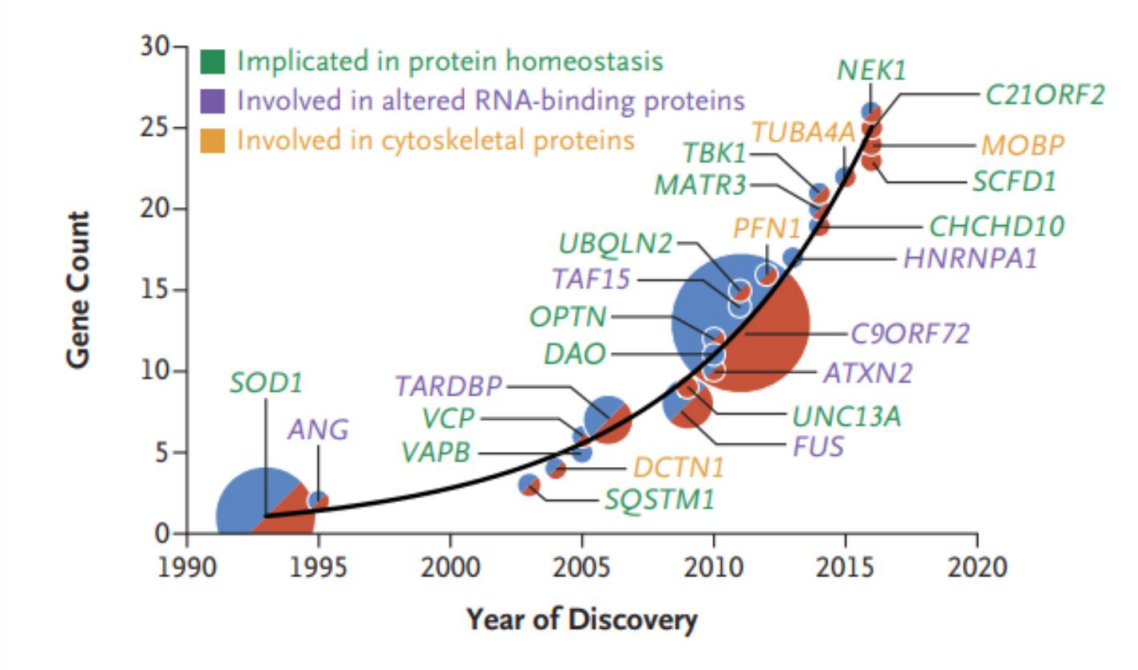

Figura 2.1: Descubrimientos de genes de ELA. El tamaño de cada círculo representa el porcentaje de todos los casos de ELA familiar vinculados a ese gen (p.ej., 20 % para SOD1 y 45 % para C9ORF72). Los círculos azules denotan genes vinculados sólo a la ELA familiar, mientras que los círculos rojos denotan genes vinculados sólo a la ELA esporádica. Los círculos mitad azules y mitad rojos denotan genes vinculados a ambas. Fuente: [10]

En [20] se revisa 151 estudios que trabajan con conjuntos de datos obtenidos a partir de RM y concluye que los resultados obtenidos en estos estudios son acordes a los patrones neuropatológicos de ELA generalmente aceptados. Esto es, los datos de la mayoría de trabajos revisados corroboran la degeneración de las estructuras corticales y subcorticales en pacientes con ELA frente a pacientes control. Algunos estudios no incluidos en [20] también señalan como rasgo de la ELA detectable vía RM la presencia de acumulaciones de hierro patológico en la médula espinal y el córtex motor primario [21, 22].

2.1.2. Biomarcadores

El desarrollo de biomarcadores es un proceso complejo y largo que requiere de la intervención de muchos actores [23]. Pueden distinguirse cuatro pasos en el desarrollo de biomarcadores: descubrimiento, desarrollo y validación de ensayos, validación de utilidad clínica e implementación clínica [24]. En cada punto del proceso intervienen distintos agentes. El descubrimiento comienza con la definición de la normalidad del proceso patógeno o biológico que el biomarcador debería indicar, por lo que se necesita identificar y validar pacientes control y con alteraciones. Para que esta validación sea efectiva se requiere del desarrollo de ensayos sensitivos y selectivos [25].

En el caso de la ELA, los biomarcadores con relevancia para el desarrollo de estrategias terapéuticas y con potencial para utilidad clínica son clínicos, genéticos, basados en fluido, electrofísicos y de neuroimagen [26]. Los biomarcadores de neuroimagen en particular destacan por su capacidad de replicar los hallazgos conocidos desde una perspectiva teórica [27].

2.2 Estado del arte

Con el advenimiento de la ciencia de datos como disciplina ha venido un aumento sin precedentes en cuanto a volúmenes de información generada y capacidad de procesamiento de los mismos, abriendo nuevas líneas de investigació en las ciencias de la salud [28]. En el contexto de este trabajo, es interesante revisar algunos trabajos que apliquen técnicas estadísticas y de machine learning sobre biomarcadores de ELA. Así se podrá analizar cómo son las metodologías más prevalentes y cuáles son actualmente el poder predictivo de los biomarcadores diagnóstico y la capacidad explicativa esperada para la discriminación de pacientes en el desarrollo de biomarcadores.

En [29] se identifica biomarcadores de plasma para diferenciar pacientes con ELA y pacientes mimic (pacientes con alteraciones que muestran sintomatología similar a ELA en primeros diagnósticos) a partir del desarrollo de clasificadores LASSO, Support Vector Machine y random forest entrenados con datos de metabolómica y con N=255. Los mejores resultados obtenidos en validación LOO (Leave One Out) fueron para LASSO, con AUC de 0.76, especificidad de 0.81 y sensibilidad de 0.65 usando todas las variables disponibles (X=353), y AUC de 0.81, especificidad de 0.90 y sensibilidad de 0.58 filtrando a un subconjunto de variables candidatas a biomarcadores (X=32).

En [30] se busca potenciales biomarcadores para diferenciar pacientes control de pacientes con ELA con modelos de regresión logística entrenados sobre datos de genómica, proteómica y de concentración de hierro con N=65. Con un ratio entrenamiento:validación de 2:1, se llega a un *accuracy* de 0.82 y se identifica varios biomarcadores, entre los cuales también se registra interacciones.

En [31] se utiliza atributos de morfometría (volumetría y espesor cortical) obtenidos por radiómica para entrenar modelos de regresión *Ridge* integrados en un proceso automatizado de diagnóstico para ELA, con pacientes control y con ELA (N=141). Los resultados que obtiene en validación son *accuracy* de 0.784 y sensibilidad de 0.857.

En [32] se utiliza suero de microRNA obtenidas de pacientes control y con ELA (esporádico y familiar) y pacientes portadores sanos para encontrar biomarcadores que permitan discriminar entre grupos de pacientes (N=53). A través del análisis visual de resultados de PCA, se termina seleccionando hasta 51 potenciales biomarcadores distintos.

Por último, [33] es uno de los estudios basados en biomarcadores de imagen con mejores resultados hasta la fecha, con métricas de clasificación muy notables para la arquitectura multicapa de red de aprendizaje profundo planteada: *accuracy* de 0.90 ± 0.01 y AUC de 0.94 ± 0.04. No obstante, el trabajo se basa en el uso de biomarcadores de imagen CT obtenidos de células iPS (células madre pluripotentes inducidas) generadas artificialmente, así que no es razonable pensar que sus resultados puedan aplicarse en cualquier centro con la tecnología disponible.

2.2.1. Crítica y propuesta

En primer lugar, es importante señalar la falta de atención prestada generalmente en la literatura al potencial de la radiómica. Encontramos incluso revisiones como [34] donde se desarrollan los hallazgos y nuevas posibilidades brindadas por los distintos enfoques ómicos y, sin embargo, no hay mención alguna a la radiómica. Asimismo, los trabajos tampoco suelen ser multiómicos o multimodales, y solamente incorporan variables obtenidas por un mismo procedimiento.

Otro problema fácilmente identificable, pero también señalado por los propios estudios revisados, es la ausencia de conjuntos de datos con grandes números de pacientes,

públicos o no. Este problema aplica incluso a algunos de los estudios multicéntricos más destacados. En [35], por ejemplo, no se llega a los 300 casos.

El número bajo de casos en estudios limita el tipo de modelos que se puede desarrollar. Es por ello que la mayoría de técnicas encontradas en la bibliografía son relativamente sencillas, en un intento de preservar el poder de generalización del modelado y evitar sobreajustes que afecten a la calidad de los resultados. Los trabajos revisados que utilizan aprendizaje profundo son únicamente aquellos que cuentan con valores considerablemente grandes de atributos, siendo en casi todos los casos datos minables obtenidos por genómica. Con todo, las estructuras de modelado estadístico y machine learning que priman en la literatura son muy sencillas y apenas combinan técnicas entre sí, siendo las combinaciones de filtrado con clasificadores si acaso.

Un problema adicional que se puede relacionar con el bajo número de casos es que en la mayoría de estudios se trabaja sobre un problema en el que se distingue únicamente entre dos clases. Estas clases suelen ser paciente con ELA y otro tipo de paciente, que puede ser paciente control, portador o *mimic*, aunque puede también ser paciente con ELA esporádico y familiar.

Finalmente, en los estudios no existe integración de la valoración del profesional junto a los atributos ómicos. Es cierto que desde una perspectiva de automatización total del modelado dentro de una secuencia única de diagnóstico o pronóstico este punto no tendría cabida. Sin embargo, la intervención del profesional es siempre necesaria para supervisar los modelos, especialmente durante el desarrollo de los mismos. Esta secuencia debe ser planteada como última meta en la investigación, pues primero es necesario completar el desarrollo de los biomarcadores pertinentes. Para este proceso, la integración del conocimiento de la figura del profesional podría ser de utilidad.

En vista de la crítica realizada, se ha propuesto un trabajo que desarrolle una *pipeline*, o secuencia estructurada de trabajo, para el desarrollo de biomarcadores a partir de un conjunto de datos minables derivados de imagen médica utilizando combinaciones de técnicas de modelado y machine learning para la clasificación de pacientes. En el marco de la literatura actual, este trabajo es novedoso en varios sentidos: por el planteamiento multimodal resultado de la integración de variables radiómicas de distintos tipos junto a variables semicuantitativas medidas por el profesional médico, por la incorporación de hasta cuatro tipos de paciente distintos, por la metodología empleada que incorpora el uso de técnicas estadísticas apenas trabajadas, por la propuesta de una arquitectura secuenciada para la automatización del análisis y por la combinación de todo lo anterior.

CAPÍTULO 3
Análisis del problema

3.1 Descripción del conjunto de datos

El conjunto de datos del que se dispone se resume en la tabla 3.1.

Contamos con 232 casos y 393 atributos en total. De los casos, 23 son control, 30 son sanos, 55 son mimic y 124 son pacientes con ELA. La diferencia entre caso control y caso sano radica en que el caso sano cuenta con antecedentes familiares de ELA; caso mimic se recuerda que es aquel que siendo inicialmente sospechoso de ELA es finalmente diagnosticado con otra enfermedad. De los atributos, 2 son clínicos (edad y sexo), 188 son de volumetría, 60 son de hierro, 72 son de espesor cortical, 6 resultan del análisis visual y 65 son atributos descartados. Se detallará el motivo de descarte en el capítulo siguiente.

3.1.1. Preprocesado médico y físico

En el conjunto de datos se puede distinguir hasta entre 6 grupos de variables distintos. Como se expuso en la sección 2.4, en la literatura se suele trabajar con conjuntos de datos de un solo tipo de variable y, sin embargo, este trabajo cuenta con variables de distintos tipos. El conjunto de datos sobre el que se ha trabajado proviene de la colaboración de múltiples equipos. A continuación, se describe brevemente y en orden lógico qué pasos se ha seguido hasta la obtención del conjunto de datos que se ha utilizado en este proyecto.

El primer paso fue la obtención de imágenes fuente. Con el protocolo seguido, las RM se realizaron en el momento del diagnóstico con un escáner de RM de alto campo $3T$ (Signa HDxt, GE Healthcare, Milwaukee, EE.UU.) y una antena de cabeza transmisora-receptora de 8 elementos. Las seis secuencias incluidas en el protocolo y sus utilidades resumidas son:

- Secuencia FLAIR-FSE potenciada en T2. Útil para valorar otras enfermedades que simulen ELA (mimic).

232 casos x 393 atr.		Atributos					
Casos		Clínicos	Volum.	Hierro	Espesor	Visuales	Descartes
Control	N=23	X=2	X=188	X=60	X=72	X=6	X=65
Sano	N=30						
Mimic	N=55						
ELA	N=124						
Total	N=232	X=393					

Tabla 3.1: Resumen del conjunto de datos proporcionado.

- Secuencia SPGR potenciada en T1 de alta resolución. Útil para obtener información de volumetría (vía morfometría basada en vóxel) y de espesor cortical (vía morfometría basada en superficie).
- Secuencia de Susceptibilidad SWI potenciada en T2 estrella. Útil para la detección y cuantificación indirecta del hierro depositado en parte del córtex.
- Secuencia potenciada en T2 estrella multi-eco con mapa de relajatividad T2 estrella. Útil para detectar y cuantificar indirectamente el hierro calculando la relajatividad.
- Secuencia BOLD EPI T2 estrella de estudio funcional en reposo. Útil para valorar la conectividad funcional en reposo.
- Secuencia DWI con Tensor difusión. Útil para valorar la composición y arquitectura microestructural encefálica.

El segundo paso fue la preparación de imágenes. Este paso es necesario para reducir la variabilidad técnica y adaptarlas a un formato estándar universal. Las técnicas aplicadas fueron:

- Eliminación de ruido de la imagen y corrección de homogeneidades. En las imágenes de RM existen variaciones de la señal que responden a factores intrínsecos al procedimiento pero que no guardan relación con el objeto estudiado. Esto es lo que se conoce como ruido, y debe ser eliminado para mejorar la calidad de la imagen.

  Para esta tarea, se aplicó la herramienta de eliminación de ruido de *Advanced Normalization Tools* [36].

- Registro de imágenes. Se entiende por registro de las imágenes su alineación y traslación a un mismo espacio para que en las secuencias de vóxeles homólogas se represente regiones anatómicas equivalentes, garantizando una coherencia anatómica.

  El registro de imágenes se realizó sirviendo el módulo de registro de *Statistical Parametric Mapping* [37].

- Eliminación de artefactos. Los artefactos añaden características que realmente no son parte de lo que se pretende capturar en la imagen. Se relacionan con los procesos fisiológicos de los sujetos y con las liimtaciones técnicas de reconstrucción de imágenes, entre los cuales además existen interacciones complejas.

  Se utilizó *Artifact Detection Tools* [38] para eliminar los artefactos.

- Normalización espacial. La normalización consiste en transformar todas las imágenes a un marco de referencia anatómico estándar común para permitir la comparativa vóxel a vóxel entre imágenes. El objetivo es corregir diferencias anatómicas globales para identificar diferencias locales.

  Se utilizó el atlas MNI-152 del *Montreal Neurologial Institute* como plantilla en SPM, complementándolo con herramientas disponibles en ANTs.

- Segmentación. En la segmentación se etiqueta las distintas regiones o vóxeles de la imagen para distinguir a vóxeles que comparten características dadas establecidas antes del tratamiento. Permite diferenciar los tejidos, áreas y regiones de interés.

  Se hizo servir SPM para este segmentar las imágenes.

- Suavizado y parcelación. El suavizado reduce la influencia de frecuencias altas y aumenta la ratio señal-ruido, consiguiendo una distribución normal de la información de los vóxeles y facilitando el encuentro de diferencias locales significativas. El parcelado, por su parte, se realizó con una modificación del atlas de Harvard-Oxford [39], incorporando las regiones cerebelosas para un total de 132 regiones parceladas.

El tercer paso consistió en el procesado de las imágenes. Procesar las imágenes significa obtener la información no identificable para el ojo humano contenida en ellas. Se trata de información cuantitativa y objetiva sobre los tejidos, estructuras y procesos metabólicos, representada en variables cuantitativas continuas. Este proceso de conversión de imagen médica a datos minables también se conoce como radiómica, y se realiza de forma automatizada dada su naturaleza. Se listan los métodos de procesado para obtener las variables de morfometría (volumetría y espesor cortical) y cuantificación de hierro seguidos en este punto:

- Análisis volumétrico cerebral: Morfometría Basada en Vóxel (VBM). Método de análisis morfométrico basado en técnicas estructurales para localizar las alteraciones morfológicas, segmentando y cuantificando el volumen de sustancia gris y blanca por regiones. La VBM es una aproximación objetiva que detecta las variaciones en volúmenes locales de componentes del cerebro comparando la intensidad de los vóxeles con una valoración regional que emplea un modelo lineal general. El proceso se realizó utilizando SPM.

  Del análisis volumétrico cerebral se terminó obteniendo la volumetría total, la volumetría por materias (gris, blanca y líquido cefalorraquídeo) y la volumetría de las 132 regiones parceladas.

- Espesor cortical: Morfometría Basada en Superficie (SBM). Método de análisis morfométrico, como el anterior, pero que descompone el volumen cortical en espesor cortical y área superficial. Esto se consigue reconstruyendo la superficie tras definir el límite entre las materias. FreeSurfer [40] fue el software empleado para obtener la SBM.

  El resultado es el espesor cortical de las 132 regiones, de las que se descarta el espesor de las regiones subcorticales por no ser relevante para el análisis.

- Cuantificación de hierro: Mapas T2 estrella y R2 estrella e imagen de susceptibilidad (SWI). Los depósitos de hierro asociados a ELA cuentan con una gran susceptibilidad ferromagnética que contribuye a detección y cuantificación.

  La primera técnica utilizada fue mapeo T2 estrella o R2 estrella (mapeo T2 invertido). T2 estrella obtiene la distribución espacial y valores por región de acumulación de hierro a través de la aplicación de métodos de relajometría sobre secuencias de gradiente multieco. Las herramientas utilizadas se encuentran disponibles en SPM.

  El resultado es la media, la mediana, la desviación estándar y los cuartiles primero y tercero para cada región de utilidad de acuerdo con la literatura.

  La segunda técnica fue SWI, secuencias eco de gradiente con tiempos elevados de eco y alta resolución espacial con tamaño reducido de vóxel. Se utilizó la secuencia SWAN (*Susceptibility Weighted ANgiography* de GE Healthcare, una secuencia tridimensional eco de gradiente que únicamente utiliza imágenes de magnitud y promedia en diferentes tiempos de eco. Es más sencilla que otras técnicas SWI en cuanto a post-procesado pero pierde información al no tener en cuenta imágenes de fase.

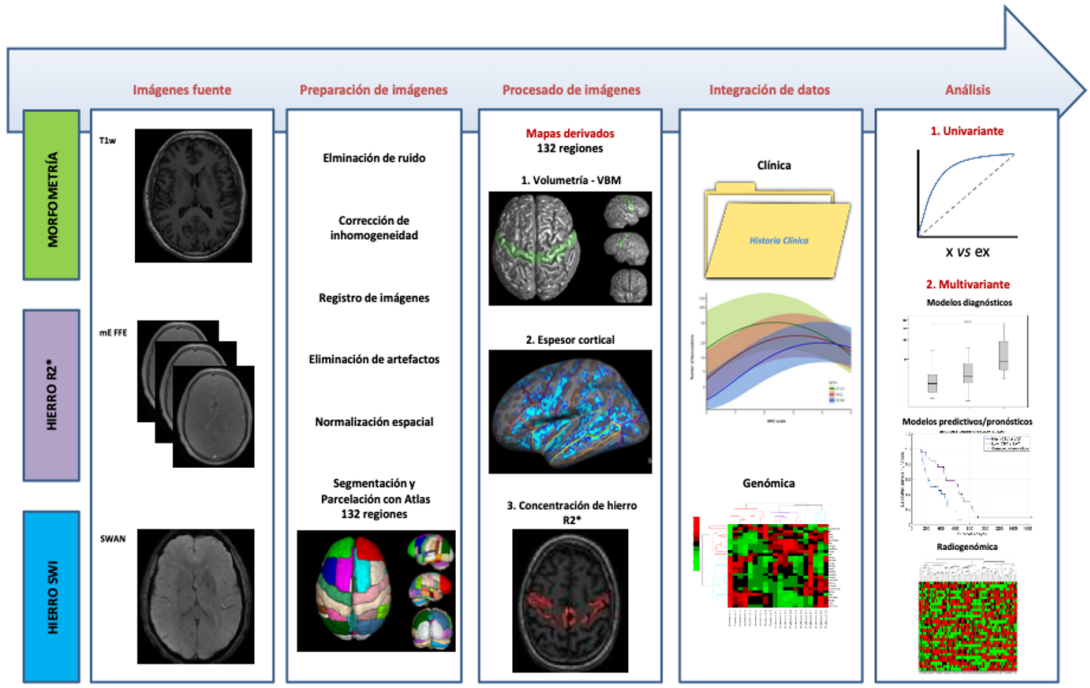

Figura 3.1: Secuencia de trabajo para la obtención del conjunto de datos. Fuente: [41].

Sobre las imágenes de SWI se realizó una valoración de afectación semicuantitativa en el córtex motor posterior izquierdo y derecho. Se dividió en tres regiones, asociadas a los miembros inferiores, superiores y a la musculatura bulbar. Sobre cada una de las seis regiones se aplicó la siguiente escala de valoración visual: 0 para una intensidad de señal normal, 1 para una hipointensidad discreta (similar a la del cuerpo calloso) y 2 para una hipointensidad marcada (similar a las venas subependimarias).

El último paso fue la integración de los datos. Junto a los datos minables resultado del procesado de imágenes, se añadieron las variables clínicas y genéticas de los pacientes para completar el conjunto de datos. El resultado en este punto es el conjunto de datos con el que se comenzó a trabajar en este proyecto como se mostraba, una vez limpiado de variables residuales, en la tabla 3.1. Es importante mencionar que el conjunto de datos fue también anonimizado durante este paso. Para la anonimización se asignó nuevos identificadores a los pacientes.

Si entendemos los varios procesos descritos como las partes de una única tarea -el desarrollo de biomarcadores de imagen médica para ELA-, podríamos pensar en la labor realizada en este trabajo como la parte final de la tarea, siguiendo a todos los pasos previos. Teniendo este trabajo en cuenta, la secuencia aproximada seguida por los equipos para la confección de este conjunto de datos se ilustra en la figura 3.1.

3.2 Metodología

Para la consecución de los objetivos, se ha propuesto el desarrollo de una secuencia estructurada automatizable de modelado, o *pipeline*. La construcción de esta *pipeline* se ha abordado desde una perspectiva de mejora iterativa: para cada arquitectura propuesta, se analizará los resultados y se identificará puntos de mejora. En base a estos, se planteará revisiones a la estructura anterior, repitiendo el proceso una y otra vez.

La *pipeline* pretende afrontar el problema como una tarea de clasificación. En este sentido, su utilidad para el desarrollo de biomarcadores diagnóstico es bastante clara, pues puede sugerirse el uso de los modelos con mejores resultados para ayuda en el diagnóstico, y de las variables más relevantes en estos modelos como potenciales biomarcadores. En lo que a biomarcadores pronóstico respecta, los mismos biomarcadores identificados para el diagnóstico pueden ser monitorizados para su uso pronóstico.

La arquitectura base de la *pipeline* es una secuencia lineal de escalado y centrado, filtrado de variables, reducción de dimensionalidad y clasificador. Decimos que la arquitectura es base porque está sujeta a los cambios que se considere pertinentes para mejorar su rendimiento.

En la *pipeline*, el escalado y centrado es necesario para la aplicación de técnicas de filtrado y reducción de dimensionalidad y reduce el tiempo de las ejecuciones al disminuir los valores en los cálculos. El filtrado de variables y la reducción de dimensionalidad sirven propósitos similares. El filtrado de variables descarta aquellas variables que parezcan de menor relevancia desde un punto de vista univariante, mientras que la reducción de dimensionalidad condensa la información disponible en menos variables. El uso de reducción de dimensionalidad es particularmente relevante porque de forma individual es muy difícil, si no imposible, asegurar que un único biomarcador puede indicar adecuadamente un proceso surrogado concreto, pues el biomarcador puede verse afectado por una multitud de factores distintos. Los procesos surrogados deben monitorizarse utilizando varios biomarcadores.

Todas las técnicas utilizadas son técnicas explicables, de modo que los procesos no son opacos, sino que el funcionamiento puede ser comprendido. Esto es importante porque buscamos una metodología accesible desde un punto de vista de interpretabilidad clínica. Las técnicas utilizadas en la pipeline se explican en la siguiente subsección.

Debe tenerse en cuenta que, para tener en cuenta todas las interacciones de técnicas posibles del modelo, los conjuntos de hiperparámetros de las técnicas deberán ser optimizados para asegurar el mejor rendimiento posible y facilitar las comparativas entre combinaciones de técnicas. Realísticamente, no se espera que la diferencia entre algunas combinaciones de hiperparámetros y otras sea más significativa que las diferencias en la arquitectura de la *pipeline*, pero sí se espera que nos permita comparar técnicas y resultados para una única arquitectura.

La optimización de hiperparámetros se ha realizado a través de un *grid search* (búsqueda en rejilla) exhaustivo con validación cruzada. *Grid search* es un método utilizado para encontrar la mejor solución a un problema dado. Parte el espacio de búsqueda en regiones y aplica un algoritmo de búsqueda a cada región. La mejor solución será aquella región donde el algoritmo de búsqueda obtenga el mejor resultado. En optimización de hiperparámetros, esto implica probar todas las combinaciones de hiperparámetros posibles, siendo cada combinación una región distinta, y guardándose como mejor solución aquella combinación que mejores métricas presente. Se dice que es exhaustiva porque agota todas las combinaciones posibles. Esto no es óptimo, pero para este problema, en el que el tiempo no es una restricción, no es necesario que lo sea. La validación cruzada se utiliza para calcular las métricas sobre conjuntos de validación que el modelo no haya visto antes. En este caso, se ha utilizado validación cruzada en 10 *folds*.

Además de utilizar diferentes técnicas en la arquitectura de modelado, también se ha empleado distintos subconjuntos de los datos disponibles para poder valorar sus capacidades individualmente. Para la estructura base y la mayoría de arquitecturas, se ha utilizado los siguientes subconjuntos: variables volumétricas, variables de hierro, variables de espesor, todas las variables salvo variables visuales, variables visuales y todas las variables. Las variables de análisis visual se distinguen del resto por su naturaleza distin-

ta: son semicuantitativas y su obtención no es automatizable. De hecho, en las iteraciones de la *pipeline*, se introducirá las variables visuales después de haber trabajado sin ellas, para poder comparar rendimientos adecuadamente. Esta partición en subconjuntos de datos permitirá valorar mejor los beneficios del enfoque multimodal.

El trabajo se ha propuesto como una tarea de clasificación binaria entre pacientes ELA y el resto de pacientes. Inicialmente, esta clasificación fue propuesta como multiclase para tratar de evitar la pérdida de información, pero los resultados iniciales fueron muy poco prometedores, sobre todo comparando con la literatura, y se decidió abandonar por completo el enfoque multiclase. El resumen de los resultados iniciales del enfoque multiclase puede consultarse en el apéndice A.

Durante el desarrollo del trabajo, también se planteó afrontarlo como una tarea binaria entre alteración (pacientes ELA y mimic) y no alteración (pacientes control y sanos) secuenciada con una tarea binaria entre las dos clases comprendidas en alteración. No obstante, los resultados fueron significativamente peores que para el enfoque ELA versus resto, por lo que se ha dejado fuera del cuerpo principal. Aun así, gran parte del análisis efectuado para la tarea ELA-resto se ha replicado también para la primera parte de este planteamiento (alteración-no alteración). Se puede encontrar los un resumen en el apéndice B.

El rendimiento de los modelos y las arquitecturas ha sido medido atendiendo a varias métricas. Esto quiere decir que no se ha obedecido estrictamente a una o dos métricas para determinar los mejores modelos, sino que se ha servido un criterio presumiblemente subjetivo en el que se ha valorado múltiples métricas. Las métricas utilizadas han sido, principalmente, *accuracy*, AUC (área bajo la curva), especificidad y sensibilidad. El *accuracy* mide el porcentaje de aciertos del modelo, el AUC mide el balance entre la especificidad y la sensibilidad, la especificidad indica la probabilidad de clasificación correcta de un caso negativo (en este caso de un paciente sin ELA) y la sensibilidad estima la probabilidad de clasificación correcta de un caso positivo (de un paciente con ELA).

La interpretación visual de los resultados también ha jugado un papel importante para determinar los mejores clasificadores. Se ha utilizado principalmente matrices de confusión, con las clases binarias y con las clases reales frente a la predicción binaria.

No todos los resultados de todas las combinaciones se han analizado, pues el número asciende a más de 140 modelos optimizados por cada propuesta de arquitectura. En su lugar, se ha hecho un primer filtrado con los valores de *accuracy* de los resultados almacenados del *grid search* y después se ha comparado los mejores modelos.

Como el número de casos del que se dispone es limitado (N=211 después del preprocesado), se ha rechazado hacer una partición en entrenamiento y validación, pues queremos contar con el máximo número de puntos posibles para entrenar los modelos. Consecuentemente, las métricas se han calculado con validación cruzada 10-fold, como durante el *grid search*. Se destaca el cálculo de métricas tanto para los conjuntos de validación como para los conjuntos de test, a fin de evitar seleccionar modelos sobreajustados (aquellos con mucha diferencia entre ambos conjuntos). En los cálculos de métricas para entrenamiento, cada punto del conjunto de datos está repetido 9 veces, una por cada vez que ha sido usado para entrenar el modelo.

La arquitectura de la *pipeline* base propuesta se muestra en la figura 3.2.

3.2.1. Técnicas utilizadas

Las técnicas de filtrado de variables, reducción de dimensionalidad y clasificadores que se ha utilizado durante todo el trabajo como parte del núcleo de la *pipeline* se presen-

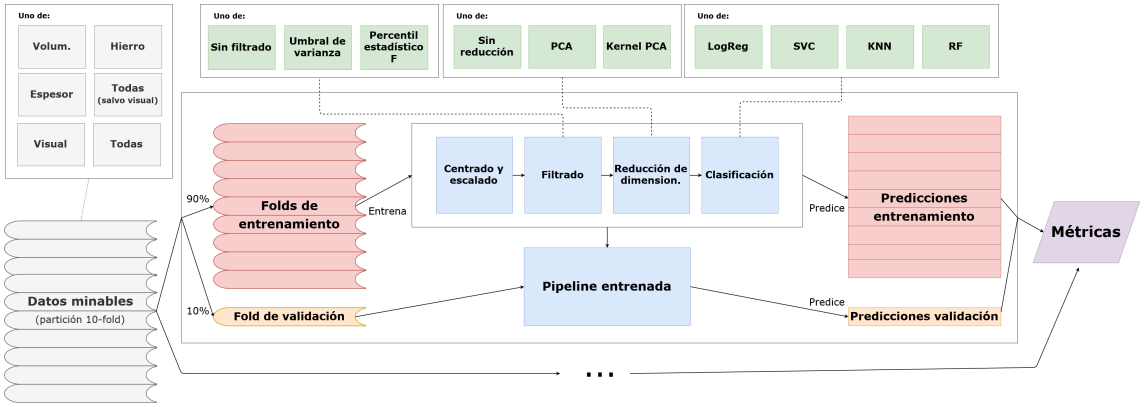

Figura 3.2: Diagrama de la *pipeline* de entrenamiento y validación. Incluye los hiperparámetros optimizados. Fuente: elaboración propia.

tan en esta subsección. También se especifica qué hiperparámetros han sido optimizados en cada una de las técnicas.

Los métodos de filtrado de variables que se ha utilizado han sido umbral de varianza y percentil de estadístico F. Se resume brevemente cada uno de los métodos:

- **Umbral de varianza**. Se filtra aquellas variables con valores de varianza menores al umbral seleccionado. Debe tenerse en cuenta que las variables han sido escaladas previamente para tener magnitudes comparables, por lo que los valores de varianza aplican por igual a todas las variables.

  Opciones: 0.2, 0.5, 0.8.

- **Percentil de estadístico F**. El estadístico F es el ratio de la varianza intergrupal a la varianza intragrupal. Se utiliza para determinar si la diferencia entre las medias de dos grupos o clases es estadísticamente significativa. Un valor elevado indica que la diferencia entre las medias de dos grupos también lo es, mientras que un valor bajo indica lo opuesto. Desde una perspectiva univariante y asumiendo que los datos han sido estandarizados de antemano, las variables con estadísticos F bajos deberían tener menor capacidad de discriminación entre grupos.

  Opciones: 20, 50, 80.

Los métodos de reducción de dimensionalidad utilizados han sido PCA (Análisis de Componentes Principales) y kernel PCA. Sigue una síntesis de cada uno de los métodos junto al listado de los parámetros optimizados para el segundo:

- **PCA**. PCA transforma el conjunto de datos a un nuevo conjunto de coordenadas a las que se llama componentes principales. Estas nuevas coordenadas maximizan la varianza del conjunto de datos y son ortogonales entre sí. Se encuentran calculando los vectores propios sobre la matriz de covarianza de los datos estandarizados y ordenándolos según sus valores propios, de manera que cada componente principal retenga más varianza que los siguientes y puedan eliminarse los componentes que expliquen poca varianza.

  Opciones (varianza retenida): 80 %, 85 %, 90 %, 95 %.

- **Kernel PCA** [42]. Kernel PCA es similar a PCA con la principal e importante diferencia de que el cálculo de los vectores propios se realiza sobre otra matriz. Esta matriz se obtiene mapeando los datos estandarizados en un nuevo espacio con una

función kernel, y centrándolos. Un kernel es una función que mapea los puntos dados en un espacio con dimensión superior, permitiendo resolver problemas no-lineales con métodos lineales.

- Kernel. Opciones: radial, polinómico orden 2, polinómico orden 3, sigmoide, coseno.
- Número de componentes retenidas. Se utiliza en lugar de varianza retenida porque con la implementación utilizada no incluía un parámetro de varianza retenida para kernel PCA, debido a que su cálculo es más costoso que para PCA.
  Opciones: 5, 10, 30, todas.

Las técnicas de clasificación que se ha seleccionado han sido regresión logística (en adelante LogReg), SVC (*Support-Vector Classification*), KNN (*K-Nearest Neighbours*) y *random forest* (en adelante RF). A continuación se resume brevemente cada uno de los modelos y se explicita qué parámetros se han optimizado:

- **LogReg**. LogReg es un modelo lineal que calcula la probabilidad de que se produzca un resultado binario. Utiliza una función logística que mapea las variables predictoras a una probabilidad entre 0 y 1. Esta probabilidad se asocia a una de las clases binarias siguiendo un umbral de decisión.
  Hiperparámetros optimizados:
  - Penalización. Tipo de penalización a añadir a la función de pérdida. Penalización $L1$ es la suma absoluta de los coeficientes, penalización $L2$ es la suma cuadrática de los coeficientes y penalización *ElasticNet* calcula un ratio $L1$:$L2$.
    Opciones: $L1$, $L2$, *ElasticNet* 0,2:0,8, *ElasticNet* 0,5:0,5, *ElasticNet* 0,8:0,2, ninguna.
  - C. Inverso de la fuerza de regularización; a menor valor, mayor regularización.
    Opciones: 0,001, 0,01, 0,1, 1, 10.
- **SVC** [43]. SVC encuentra el hiperplano de margen máximo entre las clases a separar en entrenamiento resolviendo por optimización cuadrática. Se determina los puntos más cercanos a la frontera de separación entre clases (vectores soporte) y se crea el hiperplano para que sea equidistante de los vectores soporte.
  Hiperparámetros optimizados:
  - Kernel. Opciones: lineal (equivalente a no aplicar kernel), radial, polinómico orden 2, polinómico orden 3, sigmoide.
  - C. Opciones: 0,001, 0,01, 0,1, 1, 10.
  - Gamma. Coeficiente del kernel.
    Opciones: $1/n\_vars$, $1/(n\_vars * varianza)$.
- **KNN** [44]. KNN encuentra los K-vecinos más cercanos a un punto dado y clasifica al punto en la clase mayoritaria de los K-vecinos.
  Hiperparámetros optimizados:
  - K. Número de K-vecinos.
    Opciones: 3, 5, 7, 10, 15.
  - Pesos. Función de ponderación de los puntos para predicción.
    Opciones: uniforme, inversa a distancia (los puntos más cercanos tienen más peso).

- P. Coeficiente del kernel. Potencia para el cálculo de la distancia Minkowski; con $P = 1$ equivale a distancia Manhattan y con $P = 2$ equivale a distancia euclídea.
  Opciones: 1, 1,2, 1,5, 1,8, 2.

- **RF**. RF [45] es un método de ensamble. Los ensambles combinan varios modelos para obtener un rendimiento predictivo mejor que el que podría obtenerse con cualquiera de los modelos constitutivos por separado – RF en particular consta de muchos árboles de decisión. Cada árbol se entrena con un subconjunto aleatorio de datos y las predicciones finales se realizan por votación.

  Hiperparámetros optimizados:

  - Número de árboles. Número total de árboles que forman el modelo.
    Opciones: 50, 100, 200.
  - Criterio. Función para medir la calidad de un *split* en un árbol.
    Opciones: impureza de Gini, entropía.
  - Mínimo de muestras por *split*. Mínimo de muestras necesarias para considerar un *split*.
    Opciones: 2, 5, 10, 20.
  - Máximo de predictoras por *split*. Máximo de predictoras a considerar para decidir cada *split*.
    Opciones: $\sqrt{n\_vars}$, $\log_2(n\_vars)$, 5.

3.3 Material y recursos utilizados

Casi la totalidad de la implementación del proyecto se ha realizado en Python 3.7 [46], con Jupyter Notebook como interfaz de programación. Las principales librerías que se ha utilizado han sido pandas [47], numpy [48], scikit-learn [49], matplotlib [50], seaborn [51], plotly [52] e imbalanced-learn [53]. De forma puntual, también se ha hecho servir R [54].

La mayoría del trabajo se ha realizado en un ordenador portátil tipo usuario, de prestaciones limitadas que impedían estar realizando cálculos durante días y permitir su uso normal. Es por ello que se ha hecho servir AWS (Amazon Web Services) para casi la totalidad de la optimización de hiperparámetros.

AWS ofrece un conjunto de servicios de computación en nube a través de una red de centros de datos seguros y fiables distribuidos por todo el mundo que permite construir aplicaciones en poco tiempo. Dentro de los servicios prestados por AWS, se ha empleado AWS EC2 (Elastic Cloud Computing) y AWS S3 (Simple Service Storage). De forma resumida, EC2 permite realizar cálculos en línea y S3 ejerce de almacén seguro de archivos.

En este proyecto, una instancia de EC2 se encargó de ejecutar el script de entrenamiento e ir actualizando los resultados de mejores parámetros, así como los tiempos de entrenamiento, en un *bucket* de S3.

La elección de AWS sobre otras plataformas de naturaleza similar responde a, principalmente, tres motivos. El primero es que AWS es la plataforma de computación en la nube más popular, convirtiéndola en la plataforma más accesibles, para la que más recursos existen y en una de las más seguras. El segundo es que AWS permite mantener aplicaciones en ejecución de forma ininterrumpida y sin ninguna forma de participación del usuario. El tercer y último motivo es que AWS pone a disposición del usuario un *tier* de uso gratuito siempre que no se exceda un determinado nivel de memoria utilizada, lo cual es accesible asumiendo un tiempo absoluto de ejecución mayor.

3.4 Implementación

La *pipeline* se ha implementado íntegramente en Python. Debe destacarse en particular el uso de las clases disponibles en scikit-learn, librería de la cual provienen casi la totalidad de los métodos estadísticos y de machine learning utilizados.

3.4.1. Optimización de hiperparámetros

Como se ha adelantado ya, para la optimización de hiperparámetros se ha utilizado AWS. El orden de pasos hasta la obtención de los parámetros optimizados ha sido el siguiente:

1. Creación de un *bucket* S3. Se creó un *bucket* donde se almacenaron los datos de entrenamiento, el archivo con los hiperparámetros de los modelos optimizados, el script de optimización y un archivo con los tiempos de ejecución. El *bucket* se configuró con acceso público bloqueado para garantizar la seguridad de los datos.

2. Lanzamiento de una instancia EC2. Se lanzó una imagen de máquina virtual con sistema operativo Amazon Linux a la que se concedió acceso de lectura y edición al *bucket* creado.

3. Ejecución del script de optimización. Con el par de claves creado al generar la instancia, se accedió a la instancia utilizando SSH (Secure SHell). Una vez allí, lo primero fue configurar Python y abrir un entorno virtual en el que utilizarlo. Después, se accedió al *bucket* S3, se copiaron los archivos a la memoria local y se comenzó la ejecución del script. Para cada combinación de hiperparámetros óptima nueva encontrada, el script actualizó los archivos del *bucket*.

4. Obtención de los hiperparámetros. Desde la interfaz de usuario de AWS, se descargaron los datos una vez terminada la optimización.

El tiempo total de ejecución de la instancia para la optimización de hiperparámetros fue en total de alrededor de 50 días. Este tiempo podría haber sido menor consumiendo más recursos en menos tiempo, pero ello no hubiera sido posible utilizando la *tier* gratuita de AWS.

3.5 Marco legal

El marco legal aplicable a este trabajo se relaciona con la naturaleza del conjunto de datos utilizado y con su tratamiento.

De acuerdo con la normativa europea vigente, el RGPD europeo de 2016 [55], los datos sobre los que se ha trabajado en este proyecto son sensibles. Los datos sensibles son aquellos que están sujetos a condiciones de tratamiento específicas e incluyen datos relativos a la salud [56]. En el apartado 4.15 del RGPD se define datos relativos a la salud como "datos personales relativos a la salud física o mental de una persona física, incluida la prestación de servicios de atención sanitaria, que revelen información sobre su estado de salud [55]. El conjunto de datos utilizado encaja en esta categoría, y, por tanto, el trabajo debe estar sujeto a las condiciones de tratamiento específicas de los datos sensibles.

El apartado 9.1 del RGPD prohibe el tratamiento de, entre otros, datos relativos a la salud. No obstante, en el apartado 9.2, que indica en qué casos el apartado 9.1 no aplica,

se especifica que se permite el tratamiento cuando .[el] tratamiento es necesario para fines de medicina preventiva o laboral, (...), diagnóstico médico, (...) sobre la base del Derecho de la Unión o de los Estados miembros o en virtud de un contrato con un profesional sanitario y sin perjuicio de las condiciones y garantías contempladas en el apartado 3"[55]. El apartado 9.3 mencionado explica que se autoriza el tratamiento cuando "sea realizado por un profesional sujeto a la obligación de secreto profesional, o bajo su responsabilidad, de acuerdo con el Derecho de la Unión o de los Estados miembros o con las normas establecidas por los organismos nacionales competentes, o por cualquier otra persona sujeta también a la obligación de secreto de acuerdo con el Derecho de la Unión o de los Estados miembros o de las normas establecidas por los organismos nacionales competentes"[55].

A partir de lo anterior, podemos concluir que el tratamiento entra dentro del marco legal vigente. Los datos entran dentro de una categoría de tratamiento específico, pero su tratamiento se permite al ser con una finalidad médica preventiva y de diagnóstico y bajo la obligación de secreto profesional. Se recuerda también que, en cualquiera de los casos, los datos han sido anonimizados, como es práctica estándar.

CAPÍTULO 4
Preparación y comprensión de los datos

4.1 Preprocesado

Debido a la falta de datos ausentes justificadamente imputables y a la naturaleza del conjunto de datos inicial, ya procesado previamente, el preprocesado en este trabajo se ha limitado a la eliminación de atributos residuales, por una parte, y de casos con ausencias, por otra.

Entre los atributos descartados se puede diferenciar varios grupos: identificadores redundantes o atributos residuales del proceso de obtención de los datos descrito en la subsección inmediata (X=35), atributos con información genética solo para los pacientes sanos (X=3) y atributos con información clínica, tipo y grado de avance, de ELA solo para los pacientes con ELA (X=27). El primer grupo se descarta por motivos evidentes; el segundo y el tercero, pese a que podrían utilizarse para proyectos futuros, especialmente para el desarrollo de biomarcadores pronóstico, se alejan del enfoque primariamente diagnóstico adoptado en este trabajo y seguramente para su correcto uso se debieran seguir líneas metodológicas distintas a la tomada, especialmente por la notable presencia de ausentes. Esta justificación se desarrolla más en profundidad al inicio de la siguiente sección.

Los datos presentan valores ausentes, además de en el grupo de variables descartadas, en 21 de los casos. 19 casos cuentan con ausencias para todos los atributos de hierro. Los otros 2 restantes muestran ausentes en más de la mitad de entradas. Estos casos ausentes pertenecen mayoritariamente a las clases mimic (10 casos) y ELA (10 casos), siendo el último caso control. El segundo subconjunto se descarta automáticamente por tener un porcentaje tan elevado de ausencias.

Para el primer subconjunto, decimos que los datos ausentes no son justificadamente imputables por dos motivos distintos. Desde un punto de vista teórico no sería correcto imputar las variables ausentes con las variables de otros grupos, ya que a priori no guardan ningún tipo de relación. Esto lo corrobora un vistazo rápido a la matriz de correlaciones mostrada en la figura 4.1. Si bien las variables de morfometría sí se correlacionan ligeramente las unas con las otras, las variables de hierro solamente tienen valores visibles de correlación consigo mismas.

El único procedimiento adicional en lo que a preprocesado respecta fue la recodificación de la variable clinica sexo a numérica. El conjunto de datos resultante y que se ha utilizado para el modelado tiene la estructura que se plantea en la tabla 4.1.

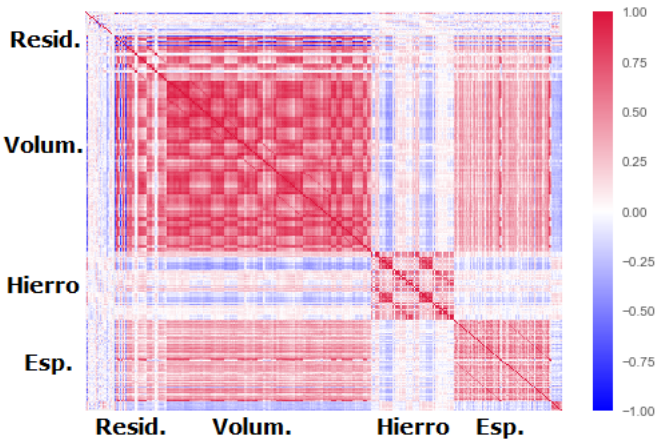

Figura 4.1: Matriz de correlaciones de las variables. Por orden, los grupos de variables que se distingue.

211 casos x 328 atr.		**Atributos**				
Casos		Clínicos	Volum.	Hierro	Espesor	Visual
Control	N=22	X=2	X=188	X=60	X=72	X=6
Sano	N=30					
Mimic	N=45					
ELA	N=114					
Total	N=211	X=328				

Tabla 4.1: Resumen del conjunto de datos limpio utilizado.

4.2 Análisis exploratorio

El resumen que sigue del análisis exploratorio se centra en las variables de radiómica. No se ha encontrado nada particularmente interesante en las dos variables clínicas y las seis variables de análisis visual.

Univariante y bivariante

En primer lugar, se realizaron análisis univariantes y bivariantes, con la clase de paciente como segunda variable, con dos fines: comprobar qué tipos de distribuciones presentaban las variables del estudio y verificar si, de partida, existían diferencias significativas en las distribuciones de las variables entre los grupos de pacientes. El objetivo fue comprobar si las muestras de los grupos trataban la misma población o, por el contrario, las muestras eran pertenecientes a poblaciones distintas.

Para estudiar las distribuciones, se utilizó un test de normalidad basado en los tests de D'Agostino y Pearson sobre las distribuciones conjuntas y desagregadas por clases,

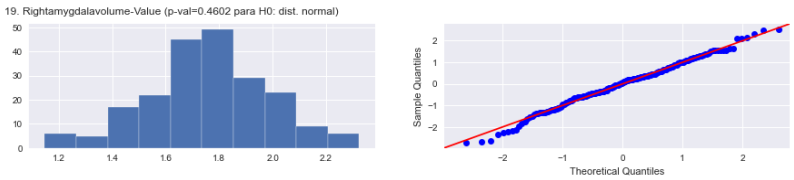

Figura 4.2: Histograma y PNN para la variable Rightamygdalavolume-Value.

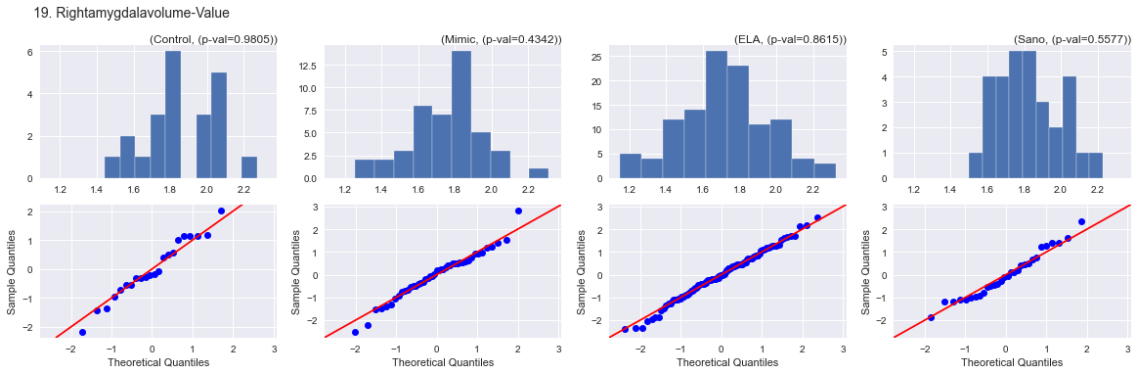

Figura 4.3: Histogramas y PNNs para la variable Rightamygdalavolume-Value, desagregada por grupos.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Control	Mimic	-0.1195	0.1608	-0.2679	0.0288	False
Control	Paciente ELA	-0.1535	0.0162	-0.2863	-0.0207	True
Control	Sano	-0.0486	0.8436	-0.2087	0.1114	False
Mimic	Paciente ELA	-0.0339	0.794	-0.1343	0.0664	False
Mimic	Sano	0.0709	0.5194	-0.0635	0.2053	False
Paciente ELA	Sano	0.1048	0.0965	-0.0122	0.2218	False

Figura 4.4: Test Tukey HSD para la variable Rightamygdalavolume-Value.

se analizó los histogramas de las distribuciones conjuntas y desagregadas por clases, se examinaron los gráficos de probabilidad normal y se realizó tests HSD de Tukey. Las figuras 4.2, 4.3 y 4.4 muestran los procedimientos seguidos para una variable cualquiera, Rightamygdalavolume-Value.

En líneas generales, las distribuciones se encontraron normales para la mayoría de variables de morfometría y exponenciales para la mayoría de variables de hierro. Por otra parte, se encontraron diferencias entre el grupo de ELA y al menos uno de los otros grupos en el 72.4 % de las variables.

Análisis de Componentes Principales

Como ya se explicó en el capítulo anterior, PCA es una técnica de reducción de dimensionalidad que resume la información de la varianza encontrada en las variables. Analizar las proyecciones de PCA permite entender con mayor profundidad las relaciones entre las variables y la naturaleza de las clases.

En la figura 4.5 se muestra las proyecciones de PCA en las dos primeras componentes. Se observa fácilmente la dificultad para diferenciar entre clases, al menos linealmente, lo que lleva a pensar que los métodos de transformación del espacio via kernel funcionarán mejor en la tarea de clasificación. La varianza se concentra en muy pocas componentes principales para todas las componentes principales. Esto puede relacionarse con las correlaciones entre variables ya vistas en la matriz de correlaciones.

Por otra parte, en la figura 4.6 se plasma todas las combinaciones de proyecciones en las dos primeras PCs para todos los grupos de pacientes con todos los subconjuntos

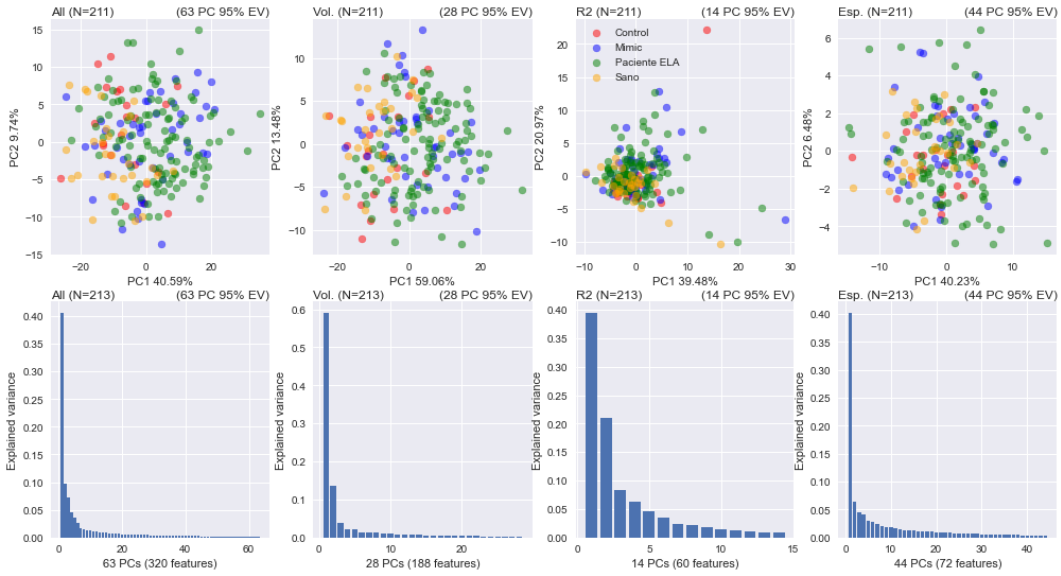

Figura 4.5: Proyecciones de los datos en las dos primeras componentes principales (arriba) y reparto de la varianza explicada (abajo). Desagregado por subconjuntos de datos.

utilizados. Los resultados no hacen sino dar más peso a la sospecha anterior, pues resulta difícil detectar diferencias entre los grupos de pacientes.

Finalmente, también se analizaron las VIP (*Variable Importance on Projection*) del PCA. Las conclusiones obtenidas fueron que las variables que más varianza aportaron en promedio a las primeras componentes fueron las de volumetría, seguidas de lejos por las de espesor en segundo lugar. Se espera, por tanto, que los modelos con variables de un solo tipo que mejor performen sean aquellos que utilicen datos de volumetría. Se muestra un ejemplo del tipo de gráfico analizado en la figura 4.7.

Hotelling T2

Hotelling T2 es una técnica utilizada para la detección de datos anómalos en conjuntos de datos con dimensionalidad alta. En este trabajo se utilizó para analizar posibles atípicos dentro de una misma clase de pacientes y descartarlos del estudio. Sin embargo, los resultados no fueron los esperados, detectándose una cantidad muy elevada de puntos como anómalos en los casos mimic y especialmente en ELA, al tiempo que apenas en sano y control. Se muestra como ejemplo el gráfico para los grupos sano y ELA con todas las variables en la figura 4.8.

Estos resultados indican que existe una dispersión muy elevada del estadístico en los grupos mimic y ELA, seguramente debido a la heterogeneidad que deben presentar sus datos (diferentes tipos de mimic o distintos grados de avance de ELA). Por ello podemos pensar que el espacio ocupado por los grupos sano y mimic será menor y los modelos capturarán mejor este espacio. Por el contrario, para mimic y ELA seguramente encontremos más desafíos.

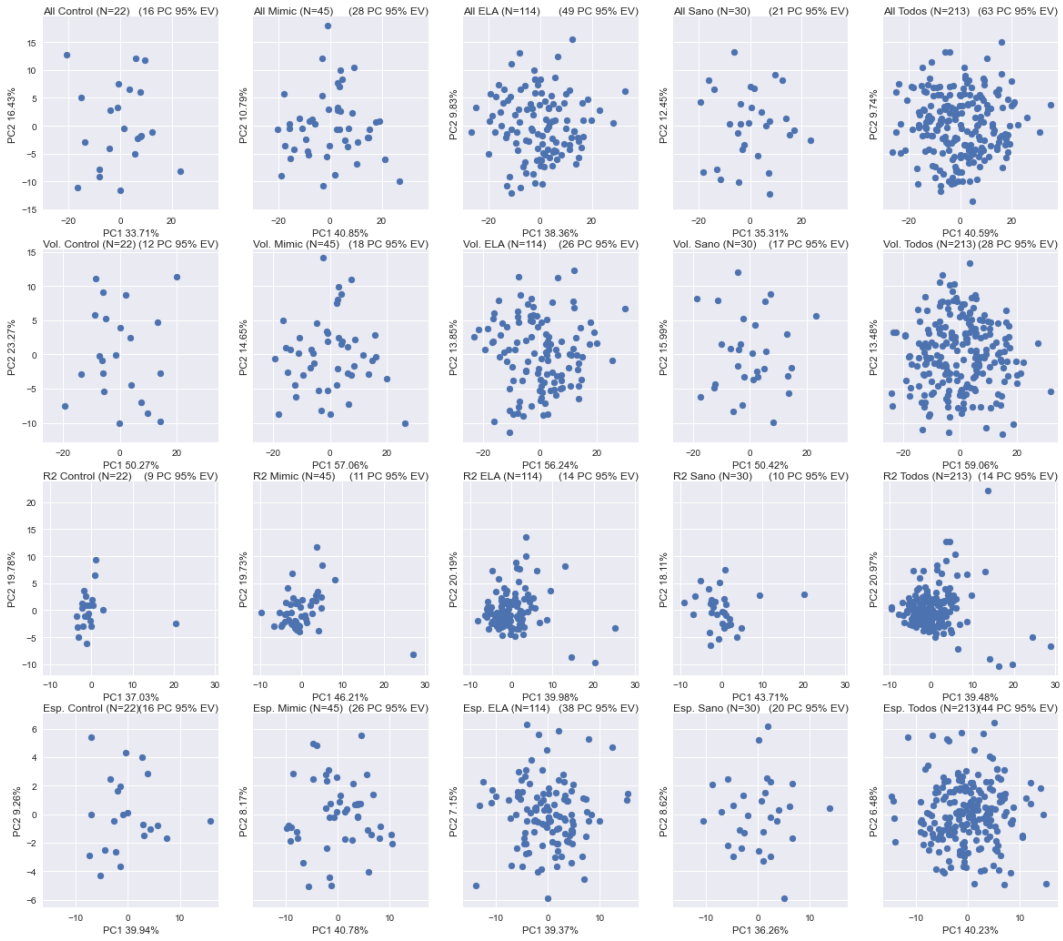

Figura 4.6: Proyecciones PCA, desagregado.

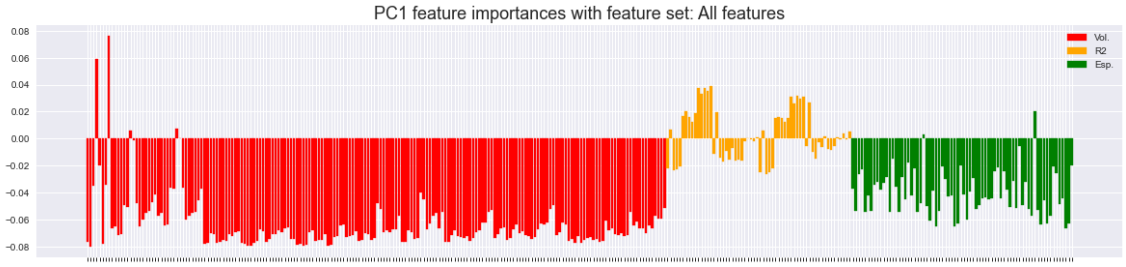

Figura 4.7: Proyecciones PCA, desagregado.

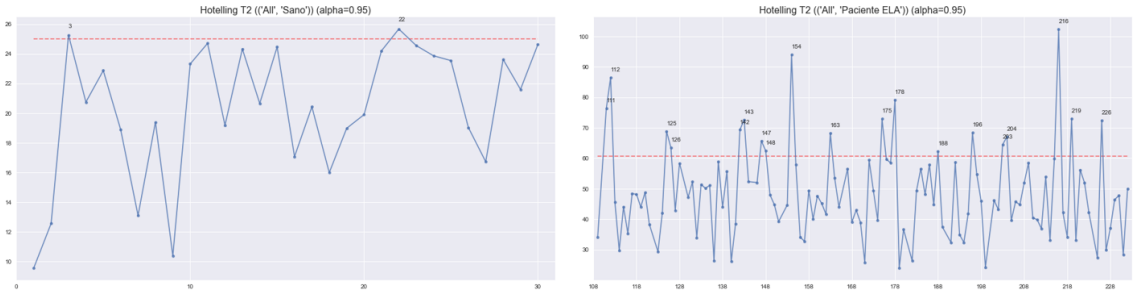

Figura 4.8: Hotelling T2 con todas las variable para el grupo sano (izquierda) y el grupo ELA (derecha).

CAPÍTULO 5
Modelado: evaluación y conocimiento extraído

Dado que con la implementación desarrollada se obtuvieron más de 140 modelos optimizados para cada una de las propuestas, solamente se mostrará los resultados del mejor modelo para cada tipo de clasificador con cada conjunto de tipos de variable. Siguiendo la misma lógica, en cada sección solamente se interpretará los resultados del mejor modelo.

5.1 ELA v. resto

Justificación

El primer enfoque probado fue clasificar ELA frente al resto de pacientes. Si buscamos alejarnos de la clasificación multiclase y tratar de balancear los tamaños de las clases, esta es la tarea de clasificación más intuitiva. Es probablemente también la más útil desde un punto de vista exclusivamente diagnóstico. Su utilidad a nivel pronóstico se encuentra en poder identificar qué diferencias existen en los pacientes con ELA respecto al resto, independientemente de su tipología, y así corroborar los hallazgos existentes y facilitar la monitorización de estos atributos en pacientes sanos.

Resultados

Los resultados de los mejores modelos para esta arquitectura se muestran en la tabla 5.1. Se ha seleccionado como mejor clasificador SVC con todas las variables. La composición completa, obviando el centrado y escalado, es la siguiente: sin método de filtrado, PCA (varianza retenida 0.95) y SVC (C=10, kernel lineal, gamma escalada a varianza).

Pese a ser el mejor modelo, el resultado obtenido es muy susceptible de mejora. Roza las predicciones perfectas en entrenamiento, lo que, junto a su elevado valor de regularización, indica que está enormemente sobreajustado. Además, obtiene métricas de especificidad y sensibilidad cercanas a las de un clasificador aleatorio para la clase mimic (figura 5.1. No obstante, no es el único modelo con este problema: ningún modelo con *accuracy* superior a 0.68 tiene una especificidad mayor de 0.6 para el grupo mimic.

Teniendo en cuenta lo anterior, las principales diferencias entre este modelo son la ligera, pero notable superior capacidad predictiva para el resto de clases, como punto positivo, y el nivel de sobreajuste no presente en otros modelos, como contrapartida. Diferencias de 0.25 en *accuracy* y 0.35 en AUC respecto a los siguientes mejores modelos

Figura 5.1: Matrices de confusión para el mejor clasificador encontrado para ELA v. resto, SVC con todas las variables.

		Validación				**Entrenamiento**			
Muestreo	Clasif.	Métrica				Métrica			
		Acc.	AUC	Sens.	Esp.	Acc.	AUC	Sens.	Sens.
Todas	LogReg	0.706	0.765	0.719	**0.691**	0.815	0.891	0.831	0.795
	SVC	**0.735**	**0.799**	**0.781**	0.68	**0.984**	**0.991**	**0.982**	**0.985**
	KNN	0.711	0.71	0.737	0.68	0.705	0.77	0.721	0.685
	RF	0.673	0.71	0.728	0.608	1.0	1.0	1.0	1.0
Volum.	LogReg	**0.701**	**0.72**	**0.772**	0.619	0.761	0.808	0.803	0.711
	SVC	0.692	0.696	0.702	0.68	0.711	0.76	0.712	0.71
	KNN	0.673	0.674	0.702	**0.639**	**0.775**	**0.847**	**0.811**	**0.733**
	RF	0.664	0.683	0.781	0.526	1.0	1.0	1.0	1.0
Hierro	LogReg	0.583	0.618	0.675	0.474	0.631	0.665	0.729	0.517
	SVC	**0.611**	**0.625**	**0.684**	**0.526**	**0.667**	**0.697**	**0.747**	**0.573**
	KNN	0.573	0.579	**0.684**	0.443	1.0	1.0	1.0	1.0
	RF	0.564	0.548	0.798	0.289	1.0	1.0	1.0	1.0
Espesor	LogReg	0.668	0.714	**0.737**	0.588	0.689	0.741	0.724	0.647
	SVC	**0.697**	**0.793**	0.746	**0.639**	1.0	1.0	1.0	**1.0**
	KNN	0.673	0.723	**0.737**	0.598	0.715	0.78	0.741	0.684
	RF	0.682	0.711	0.737	0.619	0.771	0.86	0.783	0.758

Tabla 5.1: Resultados de los mejores modelos para ELA v. resto. Las filas con fondo rojo corresponden al mejor modelo para el grupo de variables en validación. Los valores en negrita representan las mejores métricas obtenidas para el conjunto de datos acorde. Los valores subrayados pertenecen al mejor modelo. El resto de valores se muestran grises para mejor legibilidad.

no parecen ser triviales considerando la cercanía del resto de modelos entre sí, y es por eso que se ha considerado este como mejor modelo. Con todo, no se considera particularmente relevante analizar este modelo en mucha mayor profundidad debido al margen de mejora existente.

Un punto a resaltar es la notable diferencia entre las métricas de los modelos multimodales, que superan el 0.7 de *accuracy* en muchos casos, frente a los modelos con variables de un solo tipo. Dentro de estos, los modelos de volumetría son los mejores, seguidos de los modelos de espesor cortical. Esto se pudo anticipar atendiendo a las VIPs de las variables para las componentes principales durante el análisis exploratorio.

Con estos resultados, parecería que existen grandes dificultades para discriminar entre pacientes ELA y pacientes mimic con los datos disponibles. Es por eso que se planteó un enfoque de clasificación alternativo: clasificar primero entre alteración (pacientes ELA y mimic) y no alteración (pacientes sano y control) para después clasificar entre pacientes ELA y mimic. Teóricamente, esto debería facilitar a los modelos la discriminación entre los dos grupos. No obstante, esto no funcionó como se esperó en un primer momento y los métodos terminaron reduciendo sustancialmente la capacidad predictiva de las clases sano y control, clasificando la mayoría de puntos en la clase mayoritaria. Se muestra parte de este análisis en el apéndice B.

5.2 Sobremuestreo

Justificación

Analizando los resultados anteriores, resulta evidente la necesidad de corregir de alguna manera las predicciones de los pacientes mimic. Se plantea la posibilidad de que

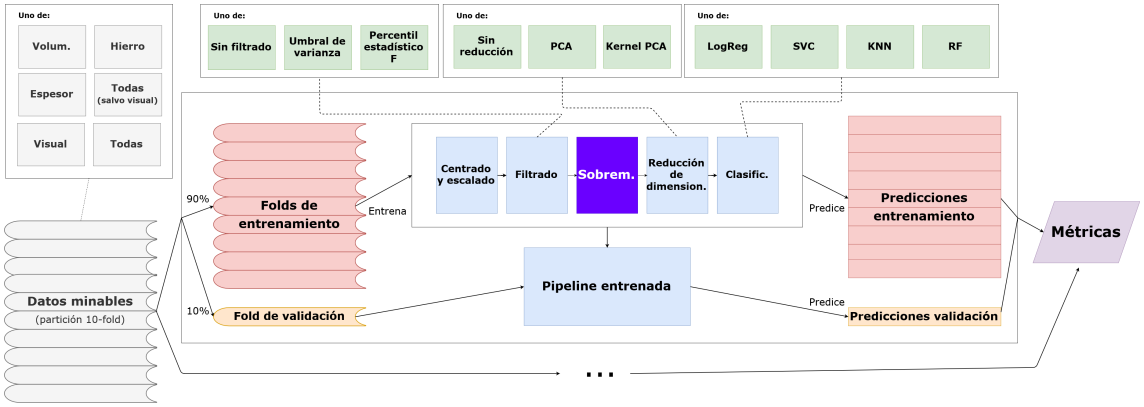

Figura 5.2: *Pipeline* con sobremuestreo incluido.

generando nuevos puntos, los clasificadores mejoren sus capacidades de delimitar el espacio asociado a cada clase. Teniendo en cuenta la distribución multivariante errática presentada por los datos de mimic y ELA en Hotelling T2, parece razonable pensar que podría beneficiar a los modelos generar nuevos puntos adicionales con los que recubrir el espacio. Se introduce dos técnicas de sobremuestreo con este propósito: SMOTE y ADASYN.

SMOTE (*Synthetic Minority Oversampling TEchnique*) es una técnica de sobremuestreo que genera nuevos puntos de datos a partir de los existentes seleccionando puntos que se encuentran entre los conocidos. Por ejemplo, un punto nuevo para un conjunto de datos con puntos (0, 0, 0) y (5, 5, 5) podría ser el punto (1,2,3), pero no el punto (6,3,3). ADASYN (*ADAptive SYNthetic*) es otra técnica de sobremuestreo que se diferencia de SMOTE por atender a las distribuciones de los datos al operar: genera puntos sintéticos utilizando KNN con distancia euclídea para que se ajusten a la distribución estimada de los datos.

El uso que se ha hecho de SMOTE y ADASYN ha sido únicamente para sobremuestrear la clase minoritaria, que es la que incluye el espacio que queremos recubrir, y obtener un ratio 1:1 entre clases. La implementación utilizada para ambas ha sido la provista en la librería imbalanced-learn. Para esta iteración, solamente se mostrará los resultados utilizando el conjunto entero de datos, vista ya la notable diferencia en la iteración anterior.

El esquema de la nueva propuesta queda como se ve en la figura 5.2.

Resultados

Los resultados obtenidos son, en general, peores que para el apartado anterior. Las métricas resultado son peores y los clasificadores siguen sin ser capaces de identificar correctamente los casos mimic. La tabla 5.2 muestra los resultados.

Destaca un modelo con una especificidad excepcionalmente elevada: el seleccionado como mejor para esta iteración, LogReg con ADASYN. Este modelo no incorpora métodos de filtrado o reducción de dimensionalidad. Sus parámetros para el clasificador son C=0.1 y penalización ElasticNet (ratio L1 0.2:0.8 L2). Aunque parece ligeramente sobreajustado, a juzgar por las métricas generales de entrenamiento, la generación de datos sintéticos parece haber generado modelos con tendencia al sobreajuste. Lo particularmente llamativo de este modelo es que ha sido el primero que ha parecido saber identificar en cierto grado los pacientes mimic. Esto puede observarse en sus matrices de confusión, figura 5.3.

Muestreo	Clasif.	Validación				Entrenamiento			
		Métrica				Métrica			
		Acc.	AUC	Sens.	Esp.	Acc.	AUC	Sens.	Esp.
SMOTE	LogReg	0.697	**0.795**	0.702	**0.691**	0.924	**0.972**	0.919	**0.93**
	SVC	**0.706**	0.75	**0.728**	0.68	**0.997**	1.0	**0.995**	1.0
	KNN	0.626	0.625	0.596	0.66	0.87	0.949	0.867	0.874
	RF	0.531	0.529	0.491	0.577	1.0	1.0	1.0	1.0
ADASYN	LogReg	0.692	**0.781**	0.632	**0.918**	0.98	0.887	0.955	0.947
	SVC	**0.725**	0.752	**0.728**	0.722	**0.997**	1.0	**0.994**	1.0
	KNN	0.578	0.61	0.57	0.588	0.913	**0.983**	0.878	**0.953**
	RF	0.502	0.543	0.404	0.619	1.0	1.0	1.0	1.0

Tabla 5.2: Resultados de los mejores modelos para ELA v. resto con sobremuestreo. Los modelos incluyen todas las variables. Las filas con fondo rojo corresponden al mejor modelo para cada técnica de sobremuestreo en validación. Los valores en negrita representan las mejores métricas obtenidas para la técnica de sobremuestreo acorde. Los valores subrayados pertenecen al mejor modelo. El resto de valores se muestran grises para mejor legibilidad.

Figura 5.3: Matrices de confusión el mejor clasificador encontrado con sobremuestreo, LogReg con ADASYN.

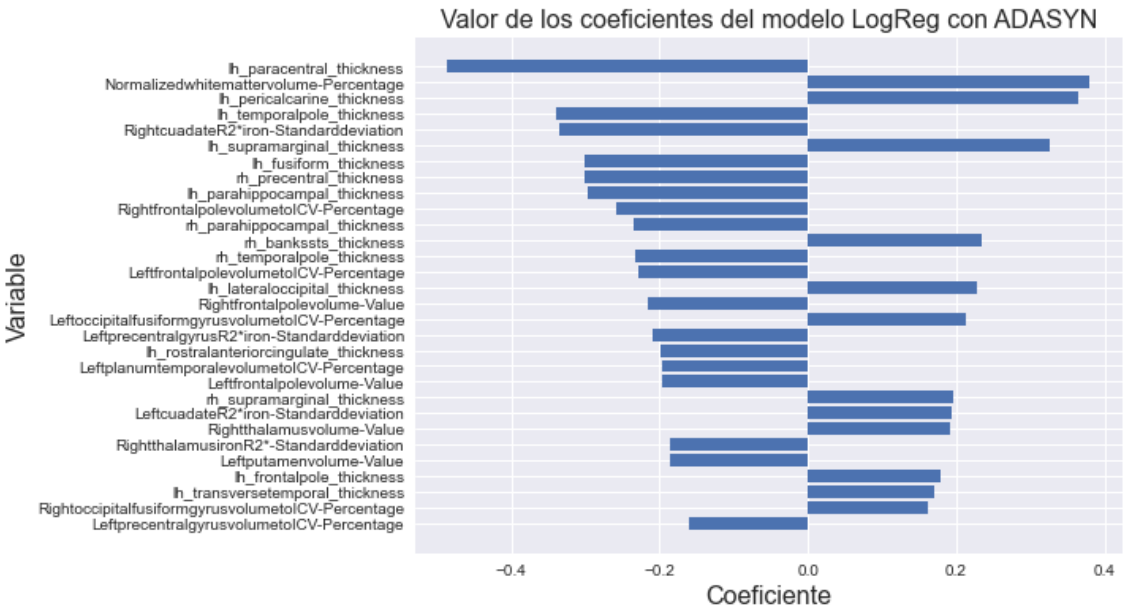

Figura 5.4: Coeficientes de las variables para el modelo LogReg con ADASYN.

Este modelo es fácil de interpretar: bastará con identificar las variables con coeficientes más elevados en el modelo. El gráfico de la figura 5.4 muestra en orden las 30 variables con coeficientes más elevados (absolutos). A pesar de que hay presencia de los tres grupos, las variables de espesor son las más abundantes entre las variables con coeficientes más elevados. Es probable que estas variables sean candidatas a biomarcadores. La primera variable, lh_paracentral_thickness, llega a doblar a variables unos pocos puestos por debajo suya.

5.3 Doble filtrado de variables

Justificación

Si la inclusión de datos sintéticos no funcionó, puede deberse a la presencia de variabilidad no relevante para la tarea de clasificación que puede verse magnificada con la generación de nuevos puntos. De ser así, sería necesario encontrar métodos que reduzcan este ruido. Una manera sencilla, generalmente efectiva y relacionada con la interpretación del mejor modelo anterior es el uso de las variables con mayor *feature importance* medida a través de coeficientes de LogReg.

Se plantea, para cada subconjunto de datos, filtrar utilizando una regresión LASSO con un parámetro de regularización bajo. Se selecciona LASSO por su tendencia a anular el efecto de variables asignándoles coeficientes de cero para reducir lo máximo posible el hipotético ruido. Se propone un threshold arbitrario de 0.1 para este filtrado.

A pesar de que ya se están aplicando técnicas de filtrado, no son incompatibles y pueden secuenciarse. El esquema para esta arquitectura se ilustra en el diagrama de la figura 5.5.

Resultados

Esta arquitectura proporciona resultados notablemente mejores que las anteriores para el conjunto de todas las variables, lo cual confirmaría la presencia de ruido en los datos.

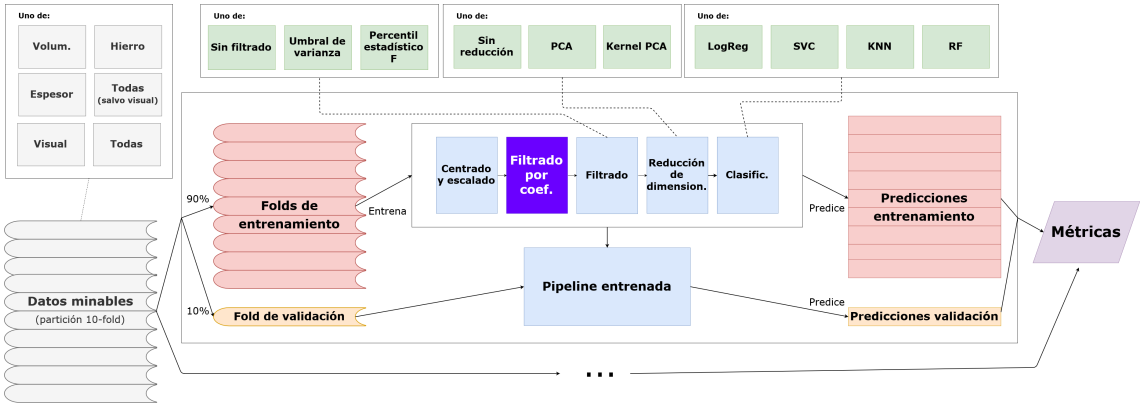

Figura 5.5: *Pipeline* con filtrado por *feature importance* incluido.

Se amplia considerablemente la diferencia respecto al resto de conjuntos, para los cuales también se percibe una mejora en métricas, si bien más ligera. Esta diferencia seguramente pueda explicarse porque el ruido tenía un efecto mayor en el conjunto de todas las variables. Esto puede ser resultado de la interacción entre las variables de distintos grupos.

El mejor clasificador seleccionado es el que mejores métricas presenta en validación, SVC con todas las variables. Tiene la misma estructura que el mejor clasificador de la primera iteración de la *pipeline* –PCA (0.95 varianza retenida) y SVC (C=10, kernel lineal, gamma escalada a varianza)–y, sin embargo, con el filtrado de variables consigue aumentar la especificidad de la clase mimic considerablemente, a 0.71, y aumentar la *accuracy* y AUC en alrededor de un 10 % cada una. Es también el único clasificador hasta el momento en superar la barrera del 0.8 de *accuracy*, con 0.815. Las variables utilizadas son muy similares a las destacadas por el clasificador anterior.

5.4 Incorporación de análisis visual

Justificación

En las iteraciones anteriores se ha comprobado la capacidad diagnóstica y se ha identificado potenciales biomarcadores utilizando únicamente variables de obtención automatizable. Uno de los puntos fuertes de este proyecto es la incorporación de variables semicuantitativas resultado del análisis visual de las imágenes para generar modelos multimodales, con datos radiómicos y valoraciones médicas.

En esta iteración, se propone utilizar la arquitectura base y la arquitectura de doble filtrado, que es la que mejor ha funcionado, para evaluar cuánto se consigue mejorar la capacidad diagnóstica de los modelos incorporando el juicio médico. También se propone utilizar las seis variables visuales por sí mismas para analizar hasta dónde permiten llegar.

Resultados

La mejoría obtenida es muy, muy notable al incorporar las variables de análisis visual. Los cuatro mejores modelos de cada tipo de clasificador superan *accuracy*, incluyendo a KNN y RF, que hasta el momento no habían proporcionado buenos resultados. El mejor método absoluto es SVC con todas las variables tras aplicar doble filtrado. Se trata de la misma secuencia utilizada hasta el momento: PCA (varianza retenida 0.95) segui-

		Validación				**Entrenamiento**			
Variables	Clasif.	Métrica				Métrica			
		Acc.	AUC	Sens.	Esp.	Acc.	AUC	Sens.	Sens.
Todas	LogReg	0.782	0.874	0.816	0.742	0.894	0.957	0.913	0.871
	SVC	**0.815**	**0.879**	**0.833**	**0.794**	0.896	0.949	0.91	0.88
	KNN	0.749	0.791	0.772	0.722	0.771	0.847	0.796	0.741
	RF	0.706	0.797	0.737	0.67	**1.0**	**1.0**	**1.0**	**1.0**
Volum.	LogReg	**0.73**	**0.737**	**0.789**	0.66	0.753	0.807	0.794	0.703
	SVC	0.682	0.707	0.728	0.629	0.68	0.7	0.718	0.635
	KNN	0.673	0.674	0.702	0.639	0.767	0.845	0.779	0.753
	RF	0.664	0.706	0.632	**0.701**	**1.0**	**1.0**	**1.0**	**1.0**
Hierro	LogReg	**0.649**	0.647	**0.728**	**0.557**	0.647	0.677	0.74	0.537
	SVC	0.602	0.627	0.667	0.526	0.679	0.702	0.75	0.597
	KNN	0.645	**0.664**	0.719	**0.557**	**1.0**	**1.0**	**1.0**	**1.0**
	RF	0.578	0.612	0.658	0.485	**1.0**	**1.0**	**1.0**	**1.0**
Espesor	LogReg	0.697	**0.771**	**0.746**	0.639	0.711	0.782	0.739	0.679
	SVC	**0.701**	0.759	**0.746**	**0.649**	0.862	0.928	0.841	**0.885**
	KNN	0.645	0.731	0.684	0.598	0.688	0.781	0.715	0.656
	RF	0.621	0.696	0.675	0.557	**0.87**	**0.95**	**0.879**	0.86

Tabla 5.3: Resultados de los mejores modelos para ELA v. resto con doble filtrado de variables. Las filas con fondo rojo corresponden al mejor modelo para el grupo de variables en validación. Los valores en negrita representan las mejores métricas obtenidas para el conjunto de datos acorde. Los valores subrayados pertenecen al mejor modelo. El resto de valores se muestran grises para mejor legibilidad.

		Validación				**Entrenamiento**			
Variables	Clasif.	Métrica				Métrica			
		Acc.	AUC	Sens.	Esp.	Acc.	AUC	Sens.	Sens.
Todas	LogReg	**0.825**	**0.879**	**0.842**	0.804	0.893	0.955	0.889	0.898
	SVC	0.806	0.867	0.807	0.804	**0.932**	**0.974**	**0.926**	**0.939**
	KNN	0.806	0.839	0.728	**0.897**	0.81	0.838	0.733	0.901
	RF	0.801	0.794	0.728	0.887	0.815	0.852	0.738	0.906
Todas (doble filtrado)	LogReg	0.858	0.924	0.851	**0.866**	0.905	0.966	0.889	0.923
	SVC	**0.872**	**0.94**	**0.886**	0.856	0.935	0.981	0.941	0.928
	KNN	0.754	0.865	0.64	0.887	0.781	0.9	0.671	0.911
	RF	0.825	0.883	0.868	0.773	**1.0**	**1.0**	**1.0**	**1.0**
Visuales	LogReg	**0.806**	**0.808**	**0.763**	**0.856**	0.805	0.849	0.764	0.852
	SVC	0.787	0.794	0.728	**0.856**	0.815	0.842	0.759	0.881
	KNN	0.749	0.807	0.702	0.804	0.778	0.834	0.72	0.845
	RF	0.782	0.804	0.737	0.835	**0.819**	**0.864**	0.76	**0.889**

Tabla 5.4: Resultados de los mejores modelos para ELA v. resto incorporando los datos de análisis visual. Las filas con fondo rojo corresponden al mejor modelo para el grupo de variables en validación. Los valores en negrita representan las mejores métricas obtenidas para el conjunto de datos acorde. Los valores subrayados pertenecen al mejor modelo. El resto de valores se muestran grises para mejor legibilidad.

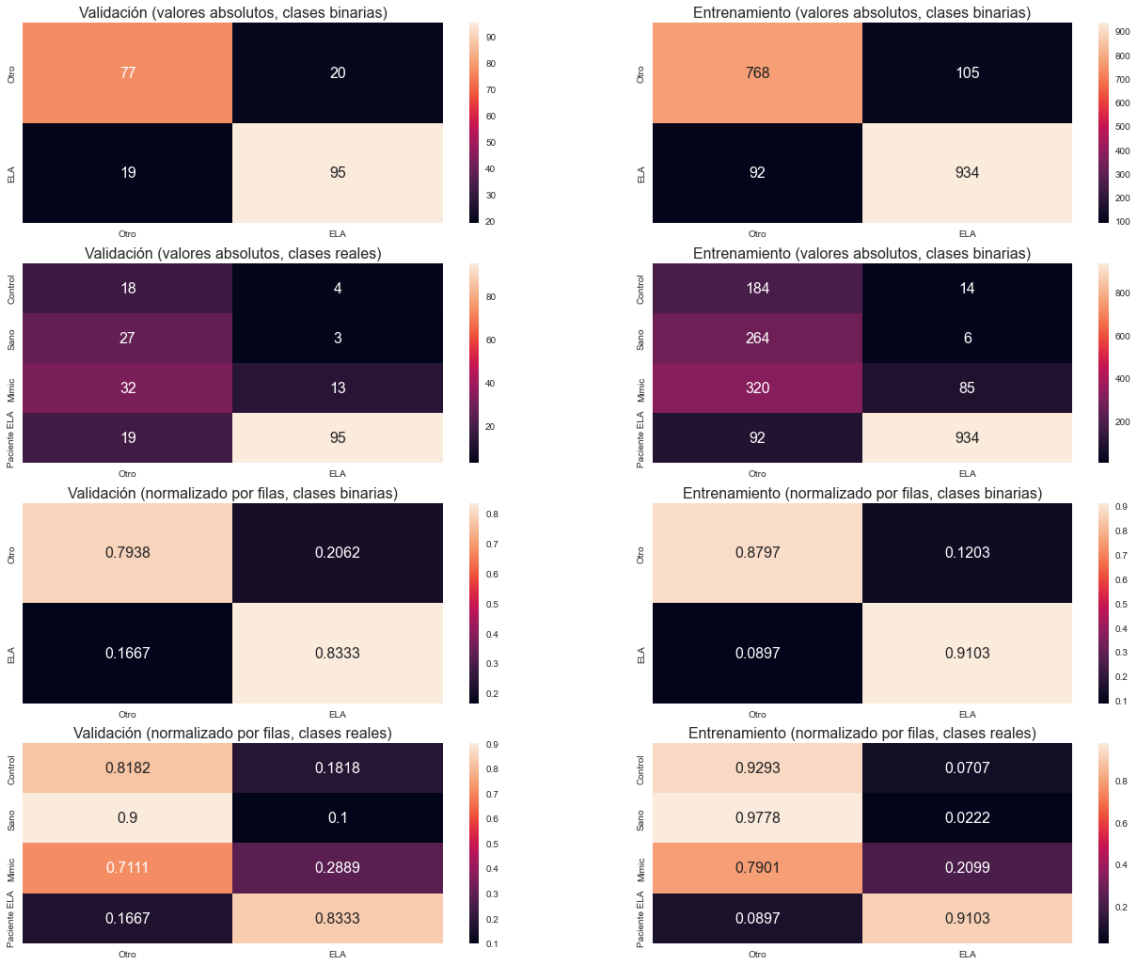

Figura 5.6: Matrices de confusión para SVC con todas las variables, el mejor clasificador encontrado con doble filtrado.

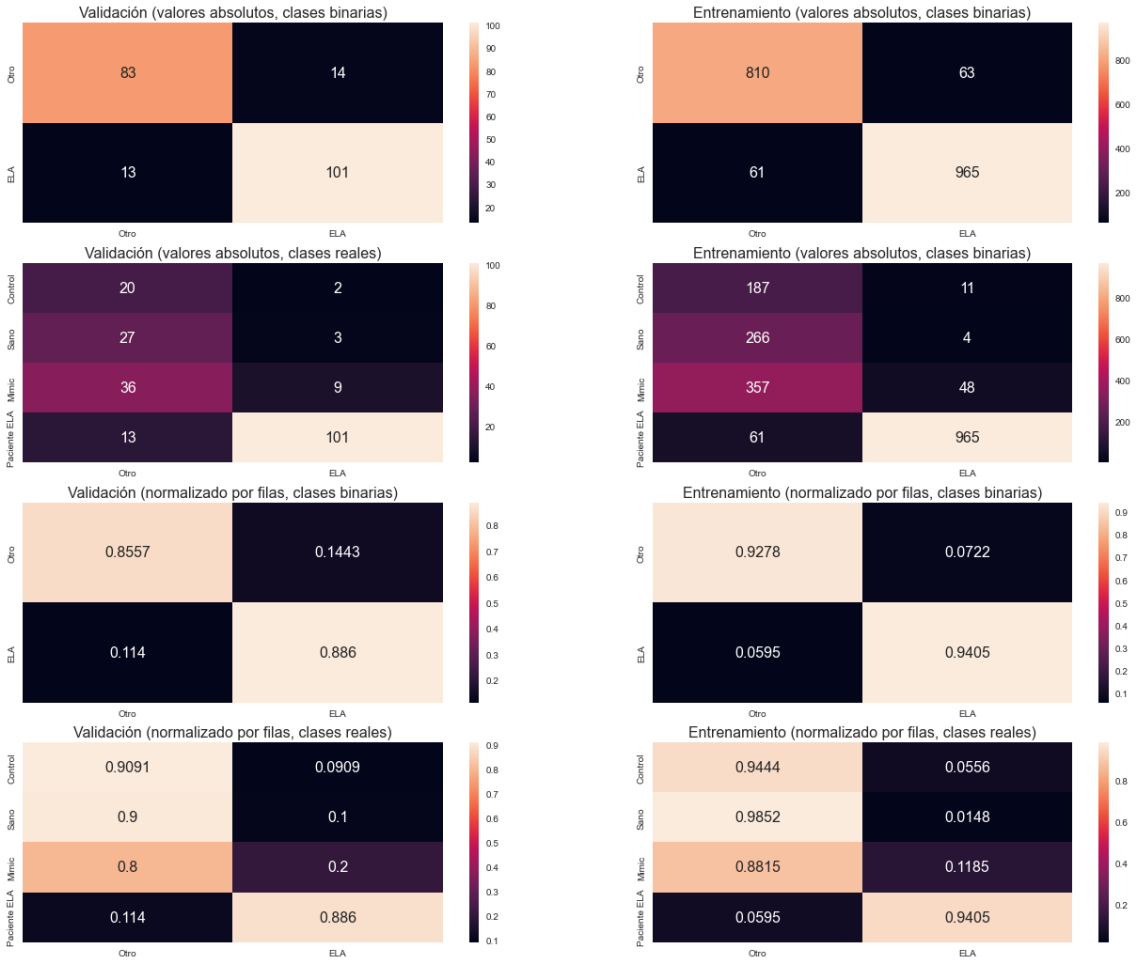

Figura 5.7: Matrices de confusión para SVC con doble filtrado sobre todas las variables, el mejor clasificador encontrado incorporando variables de análisis visual.

da de SVC (C=10, kernel lineal, gamma escalada a varianza). Pese al alto valor de C, que sugiere lo contrario, el clasificador no parece estar muy sobreajustado a los datos de entrenamiento. La figura 5.7 muestra los resultados para este modelo.

Por otra parte, las variables de análisis visual obtienen resultados muy positivos, sobrepasando con creces el rendimiento de la arquitectura base con todas las variables radiómicas: *accuracy* de 0.806 y AUC de 0.808. Su mejor clasificador es LogReg (C=0.001, sin penalización) precedido de kernel PCA (kernel radial, retiene 5 componentes). Sus resultados se muestran en la figura 5.8. Era de esperar que estas variables tuvieran un buen rendimiento, pues apenas pueden guardar variabilidad no relacionada con la enfermedad. Con todo, resulta sorprendente que un conjunto tan pequeño de variables pueda aportar una cantidad de información similar para diagnóstico que otro conjunto de más de 300.

5.5 Clasificador compuesto: votación suave

Justificación

Hasta el momento, algunos de los clasificadores parecen mostrar determinados sesgos. Particularmente, los modelos que mejor han clasificado a mimic han compensado reduciendo la sensitividad. Se especula que la combinación del juicio de algunos de los

Figura 5.8: Matrices de confusión para LogReg, el mejor clasificador encontrado con variables de análisis visual.

	Validación				**Entrenamiento**			
Variables	Métrica				Métrica			
	Acc.	AUC	Sens.	Esp.	Acc.	AUC	Sens.	Esp.
Con Visual	**0.896**	**0.929**	**0.886**	**0.907**	**0.914**	**0.971**	**0.896**	**0.935**
Sin Visual	0.815	0.872	0.825	0.804	0.873	0.938	0.894	0.848

Tabla 5.5: Resultados de los clasificadores compuestos.

mejores modelos con distintas tendencias predictivas podría contribuir a paliar este problema detectado.

Esta combinación se plantea a modo de clasificador por votación suave. Un clasificador compuesto es aquel que clasifica a partir de la salida de otros modelos. En este caso, se plantea un clasificador compuesto que vote en función a las probabilidades estimadas devueltas por otros modelos.

Para implementar esta idea, se ha seleccionado los cinco mejores modelos con todas las variables automatizables, y los cinco mejores modelos con todas las variables incluyendo análisis visual. Se diferencia, una vez más, para analizar la mejoría que permite su incorporación. Se combinará modelos que incorporen doble filtrado, pues ha sido la arquitectura propuesta con mejores resultados.

Resultados

La tabla 5.9 muestra los resultados. Los resultados para el conjunto de variables entero son los mejores hasta el momento. Todas sus métricas rozan el 0.9 en validación: *accuracy* de 0.896, AUC de 0.929, sensitividad de 0.886 y especificidad de 0.907. Además, se cumple la previsión de mejora en la capacidad predictiva de mimic sin notar una compensación en otras capacidades. La figura 5.9 contiene las matrices de confusión asociadas al modelo.

Por otro lado, y algo sorprendentemente, el subconjunto de datos sin análisis visual no mejora respecto al mejor modelo individual obtenido en las iteraciones anteriores.

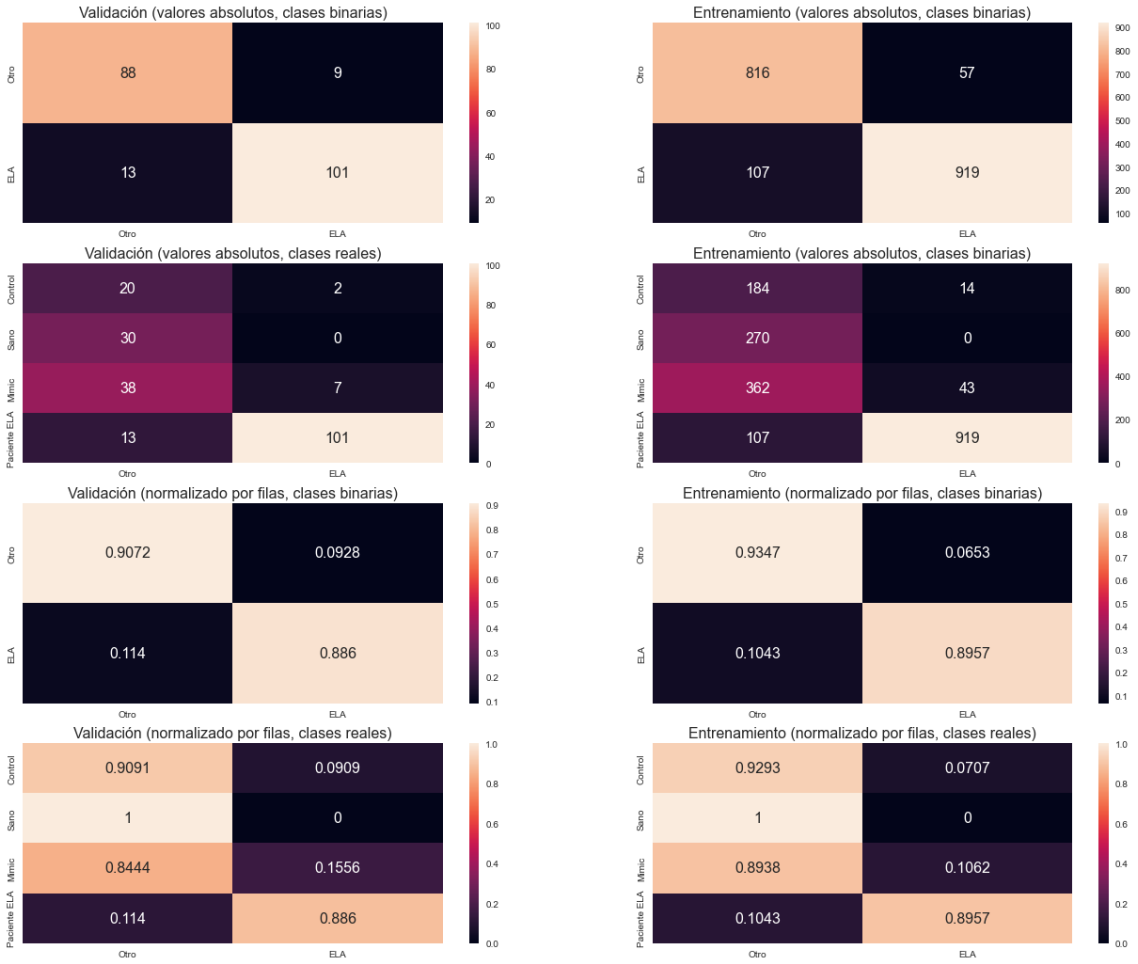

Figura 5.9: Matrices de confusión para el mejor clasificador compuesto encontrado con todas las variables radiómicas y las variables de análisis visual.

CAPÍTULO 6
Prototipo de aplicación

Para poner a disposición el trabajo de modelado desarrollado y secuenciar el conjunto de tareas, tanto las ejecutadas en este proyecto como las que le antecedieron para obtener el conjunto de datos, se propone un prototipo de aplicación explicado en este apartado.

La aplicación contiene una única interfaz de uso sencillo en la que se puede distinguir varias secciones, divididas en recuadros. De forma general, se distingue entre una sección de carga y preprocesado de datos, a la izquierda, y una sección orientada al análisis, a la derecha. De forma resumida:

- **Datos** contiene un cuadro de búsqueda que redirige al usuario al explorador de archivos de su dispositivo para que introduzca un comprimido con las imágenes fuente del paciente a estudio. Una vez cargado el archivo, se realiza la preparación de imágenes de forma automática. Para que el proceso no tenga que realizarse cada vez que se inicia la aplicación, el usuario puede guardar los resultados de la preparación de imágenes y del procesado utilizando el botón de guardar que aparece a la derecha de la barra de búsqueda una vez hay datos cargados. También se puede leer datos minados en lugar de imágenes fuente.

- **Procesado** permite realizar el procesado por partes para ir almacenando los resultados y para permitir realizar análisis en los que no sea de interés utilizar el conjunto de todos los atributos, si es que los hubiera. Se distingue entre volumetría, espesor cortical y hierro.

- **Valoración visual** redirige al usuario a una ventana *pop-up* en la que se le pide que vaya examinando y graduando las áreas del cerebro, siguiendo la misma escala planteada en este trabajo. Requiere que antes se haya ejecutado el procesado de hierro.

- **Probabilidad de ELA** muestra las probabilidades de que el paciente tenga ELA calculadas por hasta cuatro modelos distintos junto a sus intervalos de confianza 95 %. Los modelos incluidos por defecto son varios de los desarrollados en este trabajo, pero se incluye la opción de incorporar nuevos modelos de forma sencilla: al desplegar la pestaña para seleccionar modelos, aparece una opción de modelo customizado. Introducir el código de Python asociado al modelo sería suficiente para incorporarlo al análisis, aunque se necesitaría esperar a que el modelo se entrenase con los datos disponibles para poder proseguir.

  En esta parte de la interfaz se muestra dos probabilidades por modelo. La primera probabilidad, en negrita, es la calculada para los datos del paciente. La segunda, en fuente normal, es la calculada teniendo en cuenta las modificaciones a los atributos que el usuario realice en la aplicación en la sección última de la interfaz.

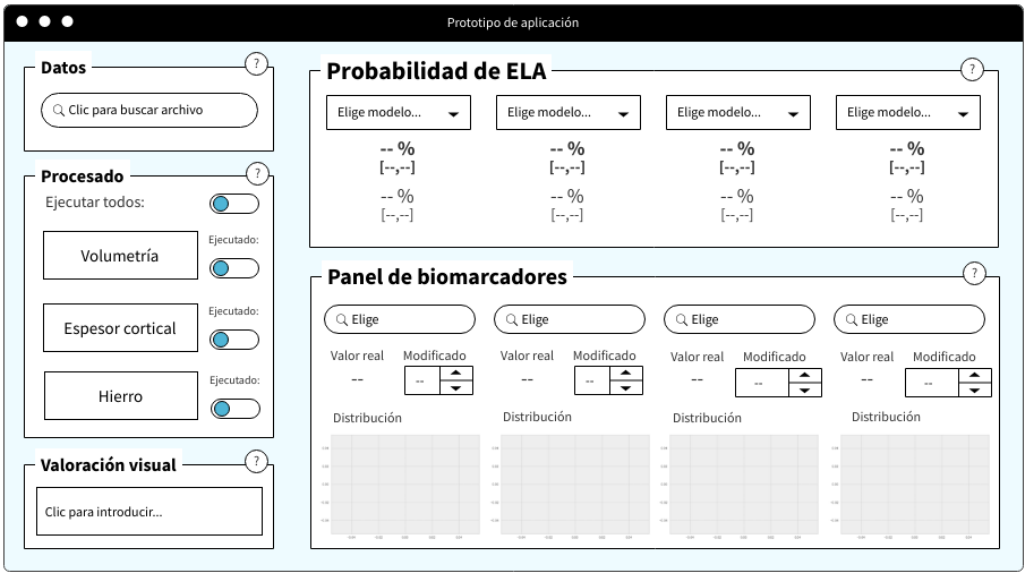

Figura 6.1: Pantalla de inicio de la aplicación.

- **Panel de biomarcadores** permite escoger hasta cuatro variables distintas para analizar cuál es su efecto en el paciente. Se muestra el valor de la variable, se permite modificar para ver cómo afectaría a las predicciones de probabilidad y se muestra dónde queda el paciente en la distribución poblacional estimada. Si la variable escogida y modificada no está incluida en el modelo, la predicción no se verá alterada.

Adicionalmente, y dado que parte de lo explicado no se muestra de forma directa en la interfaz y no permite una comprensión inmediata con un golpe de vista, se incluye secciones de ayuda que explican qué se puede esperar de cada apartado de la interfaz.

En la figura 6.1 se muestra cómo sería la pantalla de inicio de la aplicación nada más ser encendida. En la figura 6.2, durante el proceso de procesado de datos (concretamente, mientras se ejecuta el análisis asociado a espesor cortical). En la figura 6.3, se muestra la app en uso con un modelo de SVC y otro de LogReg cargados y con dos variables seleccionadas para la monitorización o el pronóstico del paciente.

La aplicación ha sido diseñada para unificar el proceso seguido hasta el momento en una sola interfaz y para brindar al usuario nuevas herramientas para uso clínico, tanto para ayudarle en el diagnóstico como para obtener información de uso pronóstico al evaluar cómo afectaría al diagnóstico el posible avance de determinada sintomatología, incluso comprendiendo el impacto conjunto de varios biomarcadores. Con todo, es un prototipo general y está sujeto a múltiples mejoras. Estas mejoras podrían pasar desde la incorporación de modelos desde una interfaz usuario frente a las técnicas de programación que harían de Python un requisito para el uso avanzado de la aplicación, hasta la distinción de las clases de pacientes en los histogramas o el uso de estimaciones kernel de densidad, para ver más claramente dónde queda el paciente a análisis. Además, contaría con dificultades evidentes en su desarrollo, pues desde la propia interfaz la aplicación debería ser capaz de hacer llamadas a varios programas específicos para poder calcular los atributos radiómicos, para entrenar modelos nuevos o para mostrar al usuario las imágenes a la hora de la valoración visual.

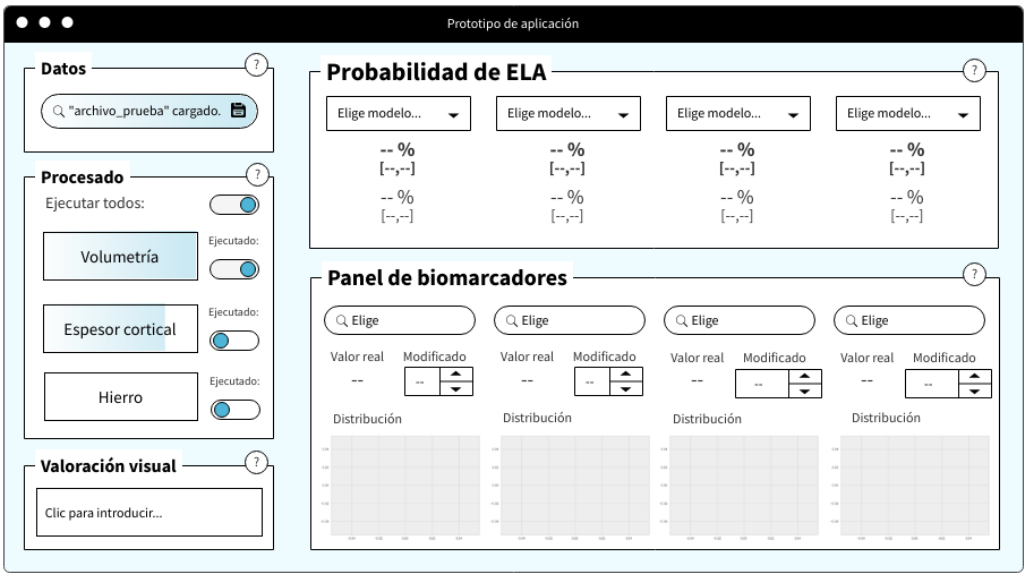

Figura 6.2: Pantalla de la aplicación durante la carga, preparación y procesado de los datos.

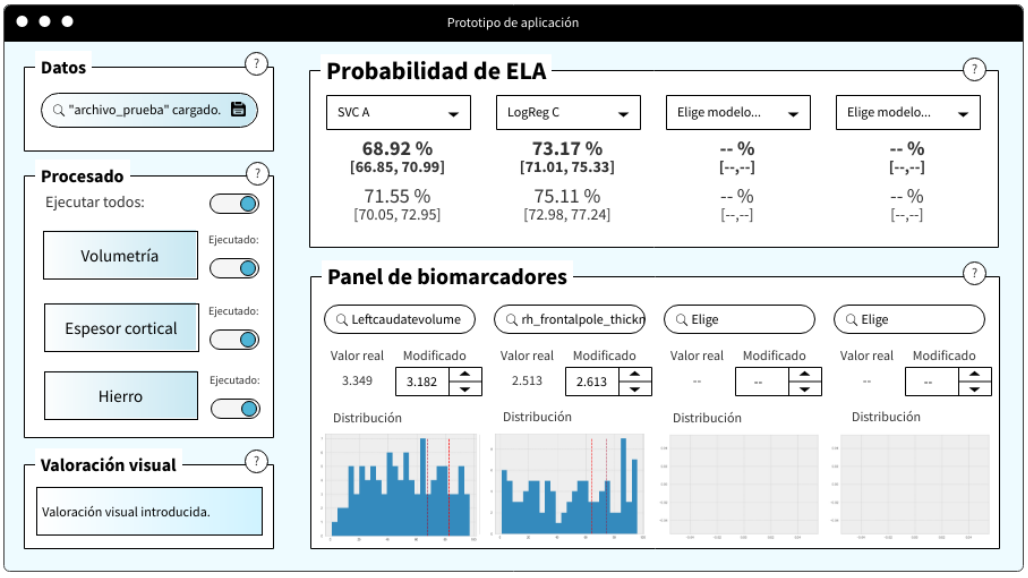

Figura 6.3: Pantalla de la aplicación en uso.

CAPÍTULO 7
Conclusiones

La estructura de las conclusiones va a analizar cómo se ha cumplido cada uno de los objetivos.

Caracterización de tipologías de paciente y clasificación de pacientes. Se ha propuesto varias arquitecturas de modelado distintas que han combinado diferentes técnicas de filtrado, reducción de dimensionalidad y clasificadores con distintos subconjuntos de los datos, para un total de hasta 144 modelos, la mayoría multimodales, por propuesta de arquitectura. Se ha afrontado el problema como una tarea de clasificación binaria en la que se ha terminado construyendo un clasificador compuesto con potencial utilidad diagnóstica que ha obtenido 0.896 de *accuracy*, 0.929 de AUC, 0.886 de sensitividad y 0.907 de especificidad en validación. Se ha obtenido listas de potenciales biomarcadores diagnóstico y pronóstico tras comprobar la efectividad para la tarea de clasificación de las variables con mayores coeficientes de modelos de regresión logística LASSO. Los modelos de penalización han contribuido a caracterizar las tipologías de paciente, si bien desde una perspectiva binaria entre pacientes ELA y el resto de pacientes.

Accesibilidad. El enfoque metodológico iterativo ha analizado los hallazgos obtenidos progresivamente y ha justificado las decisiones tomadas para cada iteración de manera que pudiese ser seguido. Además, se ha integrado todo el procedimiento seguido dentro del diseño de la interfaz prototipada, de manera que la persona profesional externa a la ciencia de datos pueda comprender el procedimiento, los resultados y sobre todo su significado sin dificultades.

Aplicabilidad. Se ha diseñado un prototipo de interfaz pensada para uso clínico. La interfaz no solamente permite utilizar los modelos construidos en un entorno, sino que permite desarrollar tanto la parte automatizable del análisis como la valoración visual al completo. De este modo, los resultados obtenidos son trasladables a un plano clínico donde el conocimiento puede ser aplicado.

CAPÍTULO 8
Trabajos futuros

A continuación se propone varias líneas de trabajo que podrían surgir a partir del proyecto desarrollado o haberse investigado de forma paralela.

En líneas generales, el trabajo ha adoptado un enfoque eminentemente diagnóstico. Esto no quiere decir que los resultados no sean aplicables desde un punto pronóstico, pero un enfoque estrictamente pronóstico sería también interesante. En el propio conjunto de datos proporcionado se cuenta con variables clínicas de avance de ELA que permitirían este tipo de estudio, para pronosticar el avance de la enfermedad, y con variables genéticas para los pacientes sanos, para pronosticar la aparición de la enfermedad.

En lo que a metodología respecta, se podría probar nuevas arquitecturas de *pipeline*. En vista del buen funcionamiento de los métodos de filtrado, podría incorporarse métodos adicionales de este tipo. Por ejemplo, podría aplicarse filtrado tras la reducción de la dimensionalidad.

Las arquitecturas ya utilizadas también son susceptibles de mejora, y no solo con adiciones a su estructura. También puede incorporarse técnicas diferentes de clasificación, reducción de dimensionalidad o filtrado. En lo que a técnicas de remuestreo respecta, en este trabajo nos hemos limitado a balancear las clases, pero podría experimentarse generando muchos más puntos de datos sintéticos después de una selección de variables más elaborada para tratar de minimizar el ruido introducido en estos datos. No obstante, un trabajo así tendría un coste temporal no asumible con las prestaciones utilizadas en este proyecto.

Otro punto metodológico muy relevante sería la incorporación de thresholds distintos a 0.5 para la clasificación. En el trabajo actual, por una cuestión de recursos computacionales, no se ha modificado el threshold por su combinación con los hiperparámetros a optimizar. Sin embargo, seguramente esta investigación guarde un gran margen de mejora si se optimizase el punto de corte entre clases.

Los problemas de coste temporal que surgen al plantear nuevos proyectos podrían responderse mejor con otras librerías distintas a scikit-learn, que no es acelerable por GPU como sí son pytorch o keras. No obstante, este tipo de librerías se basa en arquitecturas de aprendizaje profundo, no explicables y por tanto de menos interés a la hora del desarrollo de biomarcadores.

Bibliografía

[1] Brooks BR, Miller RG, Swash M, Munsat TL; World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord.*, 1:5:293–9, 2000.

[2] Al-Chalabi A, Hardiman O, Kiernan MC, Chiò A, Rix-Brooks B, van den Berg LH. Amyotrophic lateral sclerosis: moving towards a new classification system. *Lancet Neurol*, 15:11:1182-–94, 2016.

[3] Ravits J, Paul P, Jorg C. Focality of upper and lower motor neuron degeneration at the clinical onset of ALS. *Neurology*, 68:19:1571 –1575, 2007.

[4] Richards D, Morren JA, Pioro EP. Time to diagnosis and factors affecting diagnostic delay in amyotrophic lateral sclerosis *Journal of the Neurological Sciences*, 417:117054, 2020.

[5] Esclerosis Lateral Amiotrófica (ELA). Consultado en https://hospital.vallhebron.com/es/asistencia/enfermedades/esclerosis-lateral-amiotrofica-ela.

[6] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69**, 89–95, 2001.

[7] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther.* 69:3:89–95, 2001.

[8] FDA-NIH Biomarker Working Group, National Institutes of Health. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Silver Spring (MD): Food and Drug Administration (US); 2016-. Diagnostic Biomarker, 2016 [Actualizado 2020].

[9] Rosen DR, Siddique T, Patterson D, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 392:59–62, 1993.

[10] Brown RH, Al-Chalabi A. Amyotrophic Lateral Sclerosis. *N Engl J Med* 377:162–172, 2017

[11] Taylor JP, Brown RH Jr, Cleveland DW. Decoding ALS: from genes to mechanism. *Nature* 539:197–206, 2016.

[12] Qureshi M, Schoenfeld DA, Paliwal Y, Shui A, Cudkowicz ME. The natural history of ALS is changing: improved survival. *Amyotroph Lateral Scler.* 10:5-6:324–31, 2009.

[13] Juneja T, Pericak-Vance MA, Laing NG, Dave S, Siddique T. Prognosis in familial amyotrophic lateral sclerosis: progression and survival in patients with glu100gly and ala4val mutations in Cu,Zn superoxide dismutase. *Neurology* 48:1:55–7, 1997.

[14] Mehta PR, Jones AR, et al. Younger age of onset in familial amyotrophic lateral sclerosis is a result of pathogenic gene variants, rather than ascertainment bias. *J Neurol Neurosurg Psychiatry* 90:3:268-–271, 2019.

[15] Rowland LP, Shneider NA. Amyotrophic Lateral Sclerosis. *New England Journal of Medicine* 344:22:1688-–1700, 2001.

[16] Majoor-Krakauer D, Ottman R, Johnson WG, Rowland LP. Familial aggregation of amyotrophic lateral sclerosis, dementia, and Parkinson's disease: evidence of shared genetic susceptibility. *Neurology* 44:7:1872, 1994.

[17] Vogan, K. ALS susceptibility genes. *Nat Genet* 47:311, 2015.

[18] Kassubek J, Pagani M. Imaging in amyotrophic lateral sclerosis: MRI and PET. *Curr Opin Neurol* 32:5:740–746, 2019.

[19] Agosta F, Spinelli EG, Filippi M. Neuroimaging in amyotrophic lateral sclerosis: current and emerging uses. *Expert Rev Neurother* 18:5:395–406, 2018.

[20] Kocar TD, Müller H-P, Ludolph AC, Kassubek J. Feature selection from magnetic resonance imaging data in ALS: a systematic review. *Therapeutic Advances in Chronic Disease* 12:20406223211051002, 2021.

[21] Kwan JY, Jeong SY, Van Gelderen P, Deng H-X, Quezado MM, Danielian LE, et al. Iron accumulation in deep cortical layers accounts for MRI signal abnormalities in ALS: correlating 7 tesla MRI and pathology. *PLoS One* 7:4:e35241, 2012.

[22] Petillon C, Hergesheimer R, Puy H, Corcia P, Vourc'h P, Andres C, et al. The Relevancy of Data Regarding the Metabolism of Iron to Our Understanding of Deregulated Mechanisms in ALS; Hypotheses and Pitfalls. textitFrontiers in Neuroscience 12, 2019.

[23] Discovery and validation of protein biomarkers. Horvatovich P, Bischoff R. Consultado en: https://www.europeanpharmaceuticalreview.com/article/13690/discovery-and-validation-of-protein-biomarkers/.

[24] Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. *Cancer Res.* 4:3:256-–269, 2015.

[25] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol.* 24:971—983, 2006.

[26] Taga A, Maragakis NJ. Current and emerging ALS biomarkers: utility and potential in clinical trials. *Expert Review of Neurotherapeutics* 18:11:871–886, 2018.

[27] Huynh W, Simon NG, Grosskreutz J, et al. Assessment of the upper motor neuron in amyotrophic lateral sclerosis. *Clin Neurophysiol.* 127:7:2643–2660, 2016.

[28] McDermott JE; Wang J et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opinion on Medical Diagnostics* 7:1:37—51, 2013.

[29] Lawton KA, Brown MV et al.(2014) Plasma metabolomic biomarker panel to distinguish patients with amyotrophic lateral sclerosis from disease mimics. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 15:5–6:362–370, 2014.

[30] Mitchell, RM, Simmons Z, Beard JL, Stephens HE, Connor JR. Plasma biomarkers associated with ALS and their relationship to iron homeostasis. *Muscle Nerve* 42:95–103, 2010.

[31] Schuster C, Hardiman O, Bede P. Development of an Automated MRI-Based Diagnostic Protocol for Amyotrophic Lateral Sclerosis Using Disease-Specific Pathognomonic Features: A Quantitative Disease-State Classification Study. *PLoS ONE* 11:12:e0167331, 2016.

[32] Taguchi YH, Wang H. Exploring microRNA Biomarker for Amyotrophic Lateral Sclerosis. *International Journal of Molecular Sciences* 19:5:1318, 2018.

[33] Imamura K, Yada Y et al. Prediction Model of Amyotrophic Lateral Sclerosis by Deep Learning with Patient Induced Pluripotent Stem Cells. *Annals of Neurology* 89:6:1226–1233.

[34] Morello G, Salomone S, D'Agata V, Conforti FL, Cavallaro S. From Multi-Omics Approaches to Precision Medicine in Amyotrophic Lateral Sclerosis. *Front. Neurosci.* 14:577755, 2020.

[35] Calvo A, Moglia C, et al. Factors predicting survival in ALS: a multicenter Italian study. *J Neurol.* 264:1:54–63, 2017.

[36] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54:3:2033–2044, 2011.

[37] Penny W, Friston K, Ashburner J, Kiebel S, Nichols T. Statistical Parametric Mapping: The Analysis of Functional Brain Images. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. 2007.

[38] Artifact Detection Tools (RRID:SCR_005994).

[39] Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV, et al. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophr Res.* 83:2–3:155–171, 2006.

[40] Fischl B. FreeSurfer. *Neuroimage* 62:2:774–781, 2012.

[41] Hospital Universitario y Politécnico La Fe. Imagen médica en ELA. Working Paper. Grupo GIBI30. 2022

[42] Schölkopf, Bernhard, Smola A, Müller KR. Kernel principal component analysis. *International conference on artificial neural networks.* Springer, Berlin, Heidelberg, 1997.

[43] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *MIT Press*, 2000.

[44] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46:3:175-–185, 1992.

[45] Breiman L. Random Forests. *Machine Learning* 45:1:5–32, 2001.

[46] Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA. CreateSpace, 2009.

[47] McKinney W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* 445:56–61, 2010.

[48] Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature* 585:357–362, 2020.

[49] Pedregosa F, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830, 2011.

[50] Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9:3:90–95, 2007.

[51] Waskom ML. seaborn: statistical data visualization. *The Open Journal* 6:60:3021, 2021.

[52] Plotly Technologies Inc. *Collaborative data science*. Plotly Technologies Inc., Montréal, QC, 2015.

[53] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18:17:1–5, 2017.

[54] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena, Austria. Disponible en https://www.R-project.org/.

[55] REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) Consultado en https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A02016R0679-20160504.

[56] ¿Qué datos personales se consideran sensibles? Consultado en https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_es.

APÉNDICE A
Enfoque multiclase

El enfoque multiclase comenzó utilizando una *pipeline* con escalado y centrado, reducción de dimensionalidad y clasificadores. Concretamente:

- Reducción de dimensionalidad.
  - PCA (varianza retenida 0.95).
  - LDA (análisis discriminante lineal).
- Clasificadores. Salvo que se indique lo contrario, los hiperparámetros utilizados son por defecto de scikit-learn.
  - LogReg. Con y sin balanceo de clases (pesos).
  - SVC. Kernel lineal y radial, con y sin balanceo de clases.
  - KNN. 3, 5 y 10 vecinos.
  - RF. Con y sin balanceo por clase.

Los resultados de *accuracy* y *accuracy* balanceada para todas las combinaciones se muestran en la figura A.1. Ningún resultado llega a un 45 % de *accuracy* balanceada, y los resultados con mayor *accuracy* obtienen mala *accuracy* balanceada. A priori, en vista de unos resultados iniciales tan malos, parece muy optimista pensar que podría obtenerse resultados sustancialmente mejores con el enfoque de clasificación multiclase.

Analizando un poco más estos resultados con la matriz de confusión del mejor modelo y con una partición 70/30 entrenamiento/test (figura A.2, se observa claramente lo que sucede: overfitting. Prácticamente la totalidad de los datos se clasifican como la clase

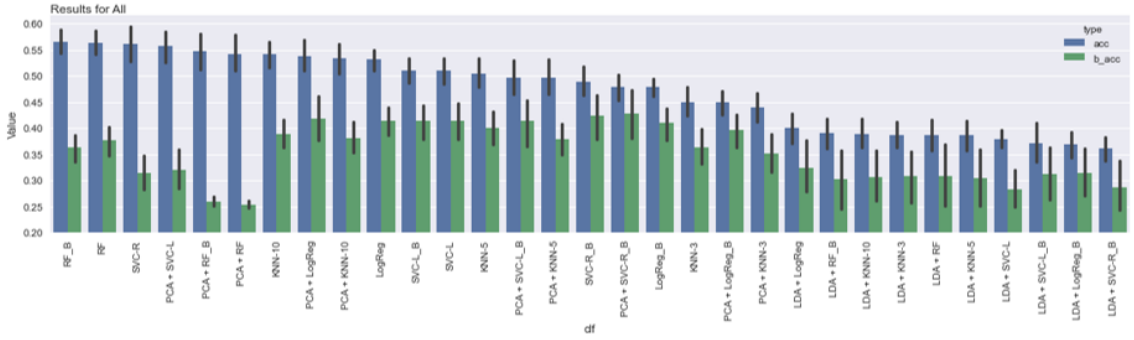

Figura A.1: Primeros resultados para clasificación multiclase.

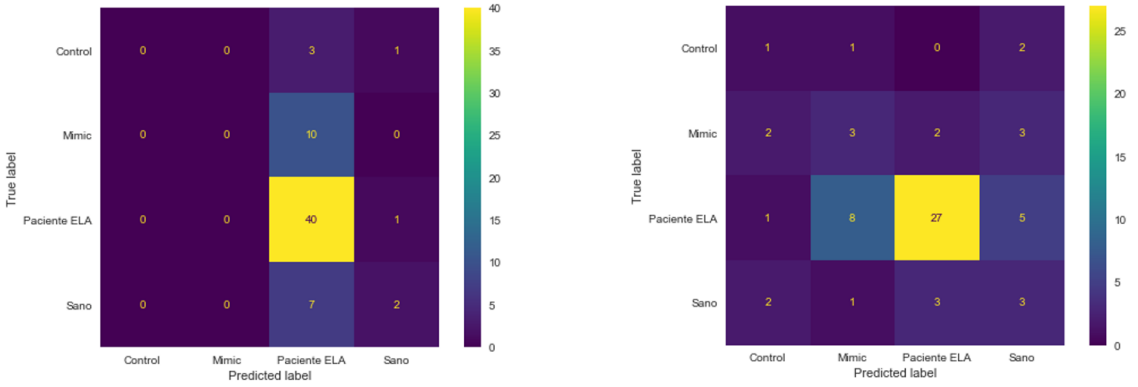

Figura A.2: Matrices de confusión para SVC radial; entrenamiento izquierda y test derecha.

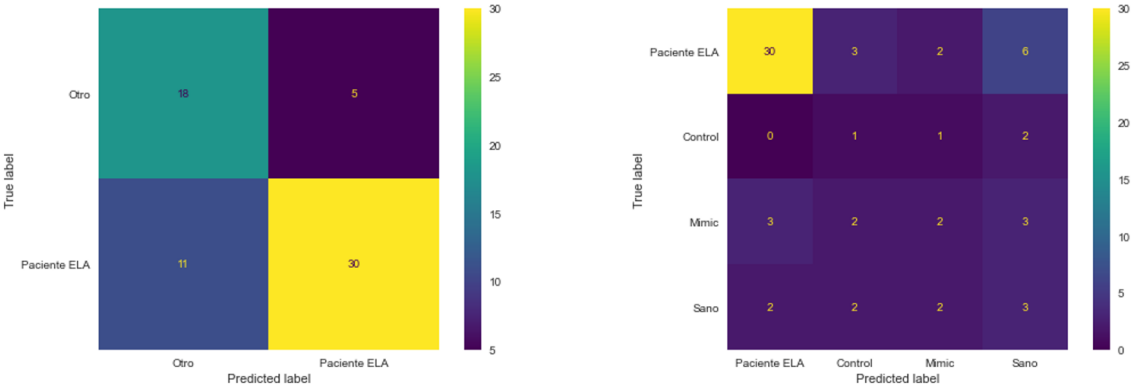

Figura A.3: Matrices de confusión de test; tarea binaria izquierda (SVC radial) y tarea binaria más tarea multiclase derecha (SVC radial con balanceo).

mayoritaria en entrenamiento, y en test, además, se confunde una proporción significativa de casos ELA con casos mimic. Una explicación plausible para esto podría ser que la introducción de métodos para el balanceo de clases dentro de los clasificadores no es suficiente, y que en su lugar debería tratar de balancearse directamente el conjunto de datos.

Para lo anterior, se propone partir la tarea de clasificación multiclase en dos partes: una primera parte binaria de ELA versus resto, y una segunda parte multiclase con las tres clases restantes. Los primeros resultados de este enfoque, que se muestran en la figura A.3 en forma de matrices de confusión con partición 70/30, no son mucho mejores. Si bien es cierto que en la tarea binaria el modelo predice ambas categorías con cierto grado de éxito, al llegar a la tarea multiclase el modelo parece completamente incapaz de discernir entre las tres clases, asignando las etiquetas similar a un clasificador aleatorio.

Se decide probar a ejecutar un PLS-DA para ver realmente cuánta variabilidad de la etiqueta de clase, para las tres clases seleccionadas, se puede explicar con los datos de radiómica disponibles. PLS-DA es una técnica que se basa en el principio de que las variables que influyen en una variable dependiente pueden ser identificadas a través de la correlación entre estas variables y la variable dependiente. PLS-DA permite conocer cuánta variabilidad de una variable se explica con un conjunto de variables. El resultado de la ejecución se muestra en la figura A.4.

Los resultados muestran que, como se podía sospechar, apenas se puede explicar parte de la variabilidad de la etiqueta con el conjunto de datos. Es tan poca la capacidad explicativa que incluso se empeora la calidad del PLS-DA a partir de la introducción de la segunda variable en el modelo. De este modo, se abandona la idea multiclase y se

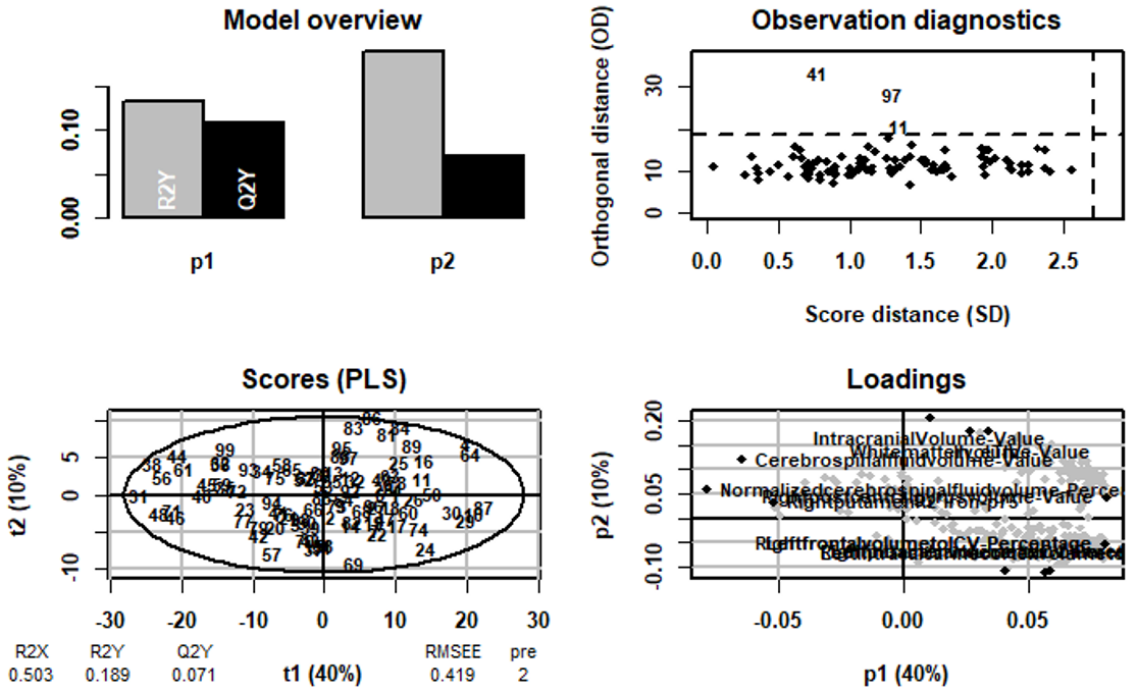

Figura A.4: Resultados de PLS entre las clases control, mimic y sano.

cambia el enfoque a la primera parte del modelado secuenciado propuesto durante este proceso: ELA versus resto. Esta terminará siendo la base de la arquitectura del trabajo.

APÉNDICE B
Enfoque alteración-no alteración

Se planteó el enfoque alteración-no alteración como contrapropuesta al enfoque binario ELA-no ELA utilizado en el grueso del trabajo. El objetivo fue agrupar los pacientes mimic con los pacientes ELA para después realizar una segunda tarea binaria en la que clasificar con más sencillez la clase mimic. Los resultados, sin embargo, no fueron los esperados, pues el problema de predicción de falsos positivos se extendió de la clase mimic a las clases sano y control. La figura B.1 muestra la matriz de confusión relativa con clases reales para el mejor modelo en este enfoque con parámetros hiperoptimizados.

A pesar de que el enfoque se incluyó de forma paralela en el trabajo durante las iteraciones de sobremuestreo y de datos filtrados, eventualmente se descartó por las complicaciones que suponía la doble tarea de clasificación –especialmente de cara a la optimización de hiperparámetros– sin llegar nunca a mejorar los resultados del enfoque ELA-resto.

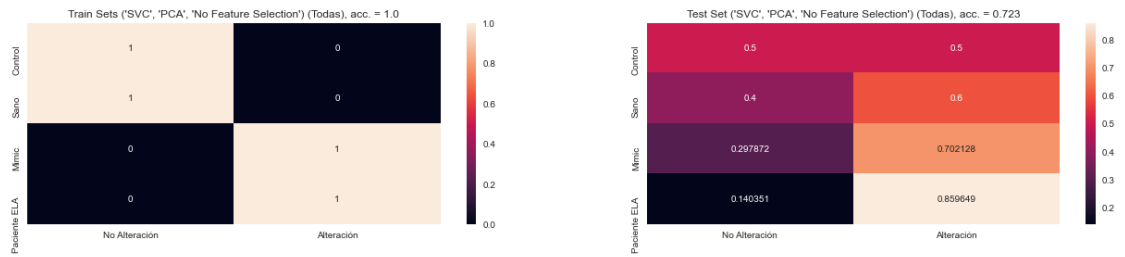

Figura B.1: Matrices de confusión para el mejor modelo en el enfoque alteración-no alteración.

APÉNDICE C
Objetivos de Desarrollo Sostenible

Los Objetivos de Desarrollo Sostenible (ODS) son una agenda global para el desarrollo sostenible que tiene como objetivo erradicar la pobreza, proteger el planeta y garantizar la prosperidad para todas las personas. Los 17 objetivos fueron acordados por 193 países en septiembre de 2015 en la Asamblea General de las Naciones Unidas. Estos son:

1. Fin de la pobreza.
2. Hambre cero.
3. Salud y bienestar.
4. Educación de calidad.
5. Igualdad de género.
6. Agua limpia y saneamiento.
7. Energía asequible y no contaminante.
8. Trabajo decente y crecimiento económico.
9. Industria innovación e infraestructura.
10. Reducción de las desigualdades.
11. Ciudades y comunidades sostenibles.
12. Producción y consumos responsables.
13. Acción por el clima.
14. Vida submarina.
15. Vida de ecosistemas terrestres.
16. Paz, justicia e instituciones.
17. Alianzas para lograr objetivos.

Los ODS son una hoja de ruta para alcanzar el desarrollo sostenible, es decir, el desarrollo que satisface las necesidades de las generaciones presentes sin comprometer la capacidad de las generaciones futuras para satisfacer las propias; son una herramienta para coordinar y orientar las acciones de los gobiernos, las empresas y la sociedad civil

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.	X			
ODS 4. Educación de calidad.			X	
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico.				X
ODS 9. Industria, innovación e infraestructuras.		X		
ODS 10. Reducción de las desigualdades.			X	
ODS 11. Ciudades y comunidades sostenibles.			X	
ODS 12. Producción y consumo responsables.				X
ODS 13. Acción por el clima.				X
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.				X

en la consecución del desarrollo sostenible. Cada objetivo tiene metas específicas que se pueden medir para saber si se está avanzando o no hacia el objetivo global.

El trabajo realizado se adecúa a los Objetivos de Desarrollo Sostenible de varias maneras.

En primer lugar, en lo referido al tercer objetivo, salud y bienestar, el trabajo tiene una aplicación evidente. El desarrollo de biomarcadores de ELA ayuda a mejorar y adelantar el diagnóstico y el tratamiento de la enfermedad, a diseñar nuevas estrategias terapéuticas y, en definitiva, a mejorar la calidad de vida de las personas afectadas por la ELA. El conocimiento adquirido también permite ayudar a pronosticar la evolución de la enfermedad. Además, con este trabajo en particular, se contribuye al desarrollo de técnicas no invasivas, minimizando el impacto psicológico de las pruebas de diagnóstico.

En segundo lugar, el trabajo contribuye al objetivo de investigación y desarrollo. El desarrollo de biomarcadores de ELA es una necesidad clínica y, al mismo tiempo, una oportunidad de investigación. El desarrollo de este tipo de biomarcadores es una tarea compleja y, por lo tanto, requiere la colaboración de un equipo multidisciplinar para la obtención de una gran cantidad de datos de calidad. El trabajo presentado aquí es un buen ejemplo de esta colaboración, pues el equipo involucrado en la investigación está compuesto por profesionales de áreas distintas. El desarrollo de biomarcadores también requiere el uso de técnicas de análisis de datos de última generación, y el trabajo presentado aquí es un buen ejemplo de esto, ya que se utilizaron técnicas de modelado poco trabajadas en el estado del arte para el análisis de datos.

En tercer lugar, el trabajo también contribuye al objetivo de la sostenibilidad. El desarrollo de biomarcadores de la ELA es una tarea compleja y, por lo tanto, requiere un gran número de recursos. El trabajo presentado aquí demuestra que es posible desarrollar biomarcadores de ELA con un impacto ambiental muy reducido, incluido a nivel computacional en una era donde cada vez los modelos son más complejos y requieren de más consumo de energía.