



UNIVERSITAT OBERTA DE CATALUNYA (UOC)

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (*Data Sciencie*)

TRABAJO FINAL DE MÁSTER

ÁREA 3

Desarrollo de un modelo predictivo para el diagnóstico del Alzheimer

Autor: María Amparo Blanch Ruiz

Tutor: Antonio Ruiz Falcó Rojas

Profesor: Laia Subirats Maté

Valencia, mayo de 2025

Créditos/Copyright



Esta obra está sujeta a una licencia de Reconocimiento - NoComercial - SinObraDerivada

3.0 España de CreativeCommons,

FICHA DEL TRABAJO FINAL

Título del trabajo:	Desarrollo de un modelo predictivo para el diagnóstico del Alzheimer
Nombre del autor:	Maria Amparo Blanch Ruiz
Nombre del colaborador/a docente:	Antonio Ruiz Falcó Rojas
Nombre del PRA:	Laia Subirats Maté
Fecha de entrega (mm/aaaa):	05/2025
Titulación o programa:	Máster en Ciencia de Datos
Área del Trabajo Final:	TFM – Área 3
Idioma del trabajo:	Español
Palabras clave:	Alzheimer, diagnóstico, <i>machine learning</i> , biomarcadores, factores de riesgo

“El aprendizaje es experiencia, todo lo demás es información”.

Albert Einstein

Resumen

Actualmente, la enfermedad del Alzheimer es la principal causa de demencia y afecta alrededor de 50 millones de personas, suponiendo un coste médico importante para la sociedad. Uno de los síntomas más significativos de esta enfermedad es la pérdida de memoria y el progresivo deterioro cognitivo de los pacientes. Se han descrito diferentes factores de riesgo, pero actualmente no se conoce por completo los mecanismos patológicos que conducen a la enfermedad. Uno de los desafíos que enfrenta el Alzheimer es el diagnóstico temprano de la enfermedad, que permitiría que se pudieran emplear terapias efectivas que impidan el progreso del deterioro cognitivo. Por ello, el objetivo del trabajo es desarrollar un modelo predictivo que permita identificar la enfermedad del Alzheimer en los pacientes. Para ello, se han empleado dos conjuntos de datos: uno que incluye datos procedentes de pacientes con la enfermedad y de personas sanas, y que recoge información sobre diferentes factores de riesgo, y otro que incluye pacientes con Alzheimer, controles y pacientes de otras enfermedades y recoge información sobre biomarcadores obtenidos de un estudio de proteómica. Con estos datos, en el caso que sea necesario se ha realizado una selección de características y se han desarrollado diferentes modelos de aprendizaje automático como *Random Forest*, *Support Vector Machine* (SVM), XGBoost y CatBoost, optimizando sus respectivos hiperparámetros, y finalmente seleccionando el modelo que mejores resultados obtenga. Con el *dataset* de biomarcadores también se ha realizado una tarea de regresión para predecir la progresión de la enfermedad. Por último, para el mejor modelo en cada caso se ha realizado la interpretación del modelo y de las variables más significativas en la toma de decisiones, lo que permite obtener información sobre los mecanismos implicados en el progreso de la enfermedad.

Palabras clave: Alzheimer, diagnóstico, factores de riesgo, biomarcadores, *machine learning*.

Abstract

Alzheimer's disease is the leading cause of dementia, affecting approximately 50 million people, assuming a significant medical and economical burden on society. One of the most significant symptoms of this disease is memory loss and the progressive cognitive impairment in these patients. Different risk factors have been described, but the pathological mechanisms that lead to the disease are still not completely understood. One of the major challenges in Alzheimer's research is the early diagnosis of the disease, which would enable effective therapies to be used to prevent the progression of cognitive dysfunction. Therefore, the aim of this project is to develop a predictive model that allows the identification of Alzheimer's disease in patients. To achieve this, two datasets were used: one consisting of data from both patients with the disease and healthy individuals, including information on various risk factors; and another containing biomarker data obtained from a proteomics study, which includes patients with Alzheimer's, healthy controls, and individuals with other diseases. When necessary, feature selection was performed on the datasets, and several machine learning models were developed, including Random Forest, Support Vector Machine (SVM), XGBoost, and CatBoost, optimizing their respective hyperparameters and ultimately selecting the model that produced the best results. Additionally, a regression task was carried out using the biomarker dataset to predict disease progression. Finally, for the best-performing model in each case, model interpretation and analysis of the most relevant variables were conducted, providing insights into the mechanisms involved in disease progression.

Keywords: Alzheimer's disease, diagnosis, risk factors, biomarkers, machine learning.

Índice general

Resumen	VII
Índice	X
Índice de figuras	XIII
Índice de tablas	XV
Introducción	1
1.1. Contexto y justificación del trabajo	1
1.1.1. Epidemiología.....	1
1.1.2. Fisiopatología	3
1.1.3. Signos clínicos y síntomas	4
1.1.4. Criterios diagnósticos	5
1.1.5. Justificación del trabajo	6
1.2. Objetivos	6
1.3. Motivación personal	7
1.4. Sostenibilidad, diversidad y desafíos ético/sociales	8
1.5. Enfoque y metodología.....	10
1.6. Planificación.....	11
1.6.1. Hitos	11
1.6.2. Tareas.....	11
1.6.3. Análisis de riesgos	13
1.6.4. Planificación temporal.....	13
1.7. Resumen de los productos del proyecto	15
1.8. Breve descripción de los demás capítulos.....	15

Estado del arte	17
------------------------	-----------

Metodología	26
3.1. Descripción de los <i>datasets</i>	26
3.1.1. <i>Dataset</i> factores de riesgo	26
3.1.2. <i>Dataset</i> biomarcadores	27
3.2. Preprocesamiento de los datos	27
3.2.1. Gestión de valores nulos.....	27
3.2.2. Análisis de correlaciones.....	28
3.2.3. Desbalanceado de clases	28
3.2.4. Limpieza y transformación del <i>dataset</i> factores de riesgo.....	29
3.2.5. Limpieza y transformación del <i>dataset</i> biomarcadores	29
3.3. Métodos de selección de características.....	31
3.3.1. Métodos de filtro por p-valor	32
3.3.2. Métodos reliefF y wrapper.....	32
3.4. Modelos de predicción	33
3.4.1. Decision Tree.....	33
3.4.2. Random Forest.....	34
3.4.3. Support Vector Machine	34
3.4.4. K-nearest neighbors.....	35
3.4.5. Logistic Regression	35
3.4.6. Ridge Regression	36
3.4.7. XGBoost.....	36
3.4.8. CatBoost	36
3.5. Validación y evaluación de los modelos.....	37
3.5.1. Ajuste de hiperparámetros	37
3.5.2. Validación cruzada	38
3.5.3. Métricas de evaluación	39

3.6.	Interpretabilidad de los modelos	42
3.7.	Softwares empleados	43
Resultados		44
4.1.	Análisis exploratorio de los <i>datasets</i>	44
4.1.1.	<i>Dataset</i> factores de riesgo	44
4.1.2.	<i>Dataset</i> biomarcadores	47
4.2.	Resultados obtenidos para el <i>dataset</i> factores de riesgo	51
4.3.	Resultados obtenidos para el <i>dataset</i> biomarcadores.....	53
4.3.1.	Predicción del Alzheimer	53
4.3.2.	Predicción de la evolución del Alzheimer (MMSE)	57
4.4.	Comparación de los modelos de clasificación.....	60
4.5.	Interpretabilidad de los modelos	61
4.5.1.	<i>Dataset</i> factores de riesgo	61
4.5.2.	<i>Dataset</i> biomarcadores	62
Conclusiones y trabajo futuro		66
5.1.	Conclusiones.....	66
5.2.	Limitaciones del estudio	69
5.3.	Trabajo futuro.....	70
Glosario		71
Bibliografía		73
Anexo		79

Índice de figuras

1.	Factores de riesgo asociados con el Alzheimer	2
2.	Estado fisiológico de un cerebro sano y un cerebro con Alzheimer	3
3.	Representación esquemática de los cambios en los síntomas clínicos y en la acumulación de los marcadores del Alzheimer durante el progreso de la enfermedad.....	4
4.	Distribución de la edad, sexo, origen étnico y grado educativo de los pacientes incluidos en el <i>dataset</i> factores de riesgo y biomarcadores	9
5.	Relación entre las tareas y los hitos planteados en el Trabajo Fin de Máster	12
6.	Diagrama de Gantt de la planificación del Trabajo Fin de Máster.....	14
7.	Estructura general de los pasos a seguir en los estudios de aprendizaje automático para la predicción del Alzheimer y otras formas de demencia.....	21
8.	Matriz de confusión de un problema de clasificación binario y otro multiclase.	39
9.	Matriz de correlaciones (Spearman) de las variables incluidas en el <i>dataset</i> factores de riesgo.....	45
10.	Distribución de las variables estadísticamente significativas en función de los grupos de diagnóstico del <i>dataset</i> factores de riesgo.....	47
11.	Matriz de correlaciones (Spearman) de las proteínas incluidas en el <i>dataset</i> biomarcadores	49
12.	<i>Volcano plot</i> de las proteínas analizadas del <i>dataset</i> biomarcadores.	50
13.	Distribución de la variable MMSE del <i>dataset</i> biomarcadores para los pacientes con Alzheimer	51
14.	Representación del ranking de los modelos para el <i>dataset</i> factores de riesgo para las diferentes métricas.....	52
15.	Correlación entre las 98 variables del <i>dataset</i> biomarcadores seleccionadas con el filtro p-valor	54
16.	Representación del ranking de los modelos para el <i>dataset</i> biomarcadores con el filtro p-valor para las diferentes métricas.....	55
17.	Correlación entre las 23 variables del <i>dataset</i> biomarcadores seleccionadas con el filtro reliefF y wrapper	56
18.	Representación del ranking de los modelos para el <i>dataset</i> biomarcadores con el filtro reliefF y wrapper para las diferentes métricas.....	57

19. Correlación entre las 29 variables del <i>dataset</i> biomarcadores seleccionadas para la predicción del MMSE.....	58
20. Representación del ranking de los modelos para el <i>dataset</i> biomarcadores para la predicción del MMSE para las diferentes métricas.....	59
21. Análisis de los residuos de los valores predichos de MMSE con el modelo CatBoost con respecto a los valores reales	60
22. Comparación del <i>recall</i> y AUC de los modelos para los diferentes <i>datasets</i> y los dos filtros de selección de variables aplicados	61
23. Ranking de la relevancia en el modelo CatBoost de las variables del <i>dataset</i> factores de riesgo para la predicción del Alzheimer.....	62
24. Ranking de la relevancia de las variables del <i>dataset</i> biomarcadores seleccionadas mediante el filtro p-valor para la predicción del Alzheimer: con el modelo Random Forest	63
25. Ranking de la relevancia de las variables del <i>dataset</i> biomarcadores seleccionadas mediante el filtro reliefF y wrapper para la predicción del Alzheimer con el modelo XGBoost.....	64
26. Ranking de la relevancia en el modelo CatBoost de las variables del <i>dataset</i> biomarcadores para la predicción del valor del MMSE	65

Índice de tablas

1.	Características de los estudios que emplean el aprendizaje automático para la detección y predicción del Alzheimer	22
2.	Distribución de registros en las diferentes clases de diagnóstico del <i>dataset</i> ...	31
3.	Módulos empleados para el desarrollo del estudio y su versión utilizada.....	43
4.	Descripción de las variables del <i>dataset</i> factores de riesgo por grupo de diagnóstico.	46
5.	Características demográficas de los sujetos incluidos en el <i>dataset</i> biomarcadores.	48
6.	Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el <i>dataset</i> factores de riesgo.	52
7.	Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el <i>dataset</i> biomarcadores tras aplicar la selección de características por el filtro del p-valor.	54
8.	Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el <i>dataset</i> biomarcadores tras aplicar la selección de características por el filtro reliefF y <i>wrapper</i>	56
9.	Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de regresión y predicción del MMSE en pacientes con Alzheimer con el <i>dataset</i> biomarcadores	59

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

El Alzheimer es una enfermedad progresiva y degenerativa del cerebro, que provoca el deterioro de la memoria y la conducta. Es la principal causa de demencia y se está convirtiendo rápidamente en una de las enfermedades más caras, letales y angustiosas de este siglo [1]. Además, la enfermedad del Alzheimer está reconocida por la Organización Mundial de la Salud como una prioridad global en salud pública [2].

1.1.1. Epidemiología

La enfermedad del Alzheimer es la forma más común de demencia en todo el mundo y se estima que es responsable del 60 al 80% de todos los casos [3,4]. Esta enfermedad ha afectado entorno al 2-8% (más de 50 millones de personas) de la población mundial en las últimas décadas, suponiendo un coste médico significativo para la sociedad [3,5]. La prevalencia es mayor en mujeres que en hombres y aumenta con la edad, duplicándose aproximadamente cada 5 años hasta los 85 años [3]. Para 2050, se estima un aumento de la prevalencia de la demencia en todas las edades debido al crecimiento y al envejecimiento de la población [3].

El factor de riesgo más importante para la demencia y el Alzheimer es la edad, otros factores son la diabetes mellitus, hipertensión, obesidad y colesterol HDL bajo, la pérdida

de la audición, abuso del alcohol, tabaquismo, depresión, baja actividad física, aislamiento social, lesiones cerebrales traumáticas [6]. Sin embargo, varios de ellos, como la baja actividad física, el aislamiento social y la depresión, pueden tener un vínculo bidireccional y pueden ser parte de la fase prodrómica de la demencia [6,7]. Por otro lado, la diabetes mellitus y la hipertensión son probablemente los factores de riesgo más comunes e importantes para la demencia, además estos factores pueden influir sobre el riesgo cerebrovascular que a su vez afecta a la expresión clínica de esta patología [6,8] (**Figura 1**).

En cuanto a los factores de riesgo genéticos, se han analizado más de 600 genes que pueden ser factores susceptibles en el Alzheimer [6] (**Figura 1**). Mutaciones raras en los genes APP (que codifica la proteína precursora amiloide), PSEN1 y PSEN2 (que codifica la presenilina 1 y 2, respectivamente) son las responsables de casi todos los casos de Alzheimer de herencia dominante [9]. Las personas con mutaciones en estos genes son casi siempre menores de 65 años cuando empiezan a desarrollar los síntomas. Mientras que polimorfismos en el gen APOE suponen el factor de riesgo genético más importante para la enfermedad que se presentan después de los 65 años [6].



Figura 1. Factores de riesgo asociados con el Alzheimer, (Realizado con Biorender).

1.1.2. Fisiopatología

El Alzheimer es un trastorno que conduce a la disfunción sináptica y a la pérdida de la integridad neuronal, que son las causas probables del deterioro cognitivo [6,10,11]. Esta patología se define histopatológicamente por la formación y acumulación de placas de A β y de ovillos neurofibrilares (formados por la proteína tau). En este sentido, se ha teorizado que las proteínas A β y tau se pliegan incorrectamente, se autoensamblan y se propagan mediante un mecanismo endógeno similar a la agregación y la propagación de la proteína priónica [12,13]. Las placas de A β alteran los circuitos y las células cerebrales vecinas, además también forman pequeños ensamblajes oligoméricos solubles que perjudican la función de las neuronas y la glía, y también se acumula en las paredes de los vasos sanguíneos cerebrales dando lugar a angiopatías β -amiloides (**Figura 2**) [12].

A pesar del creciente conocimiento sobre los mecanismos moleculares, bioquímicos y celulares de la enfermedad, la verdadera etiología y patogénesis siguen siendo desconocidas, dificultando el desarrollo de nuevos fármacos efectivos frente a la enfermedad [2]. Además de la predisposición genética cada vez hay más evidencias que vinculan a la gliosis, la inflamación, las alteraciones en la producción y eliminación de las especies reactivas de oxígeno (ROS), la disfunción mitocondrial y la acumulación de iones metálicos con la patogénesis de la enfermedad [2,6].

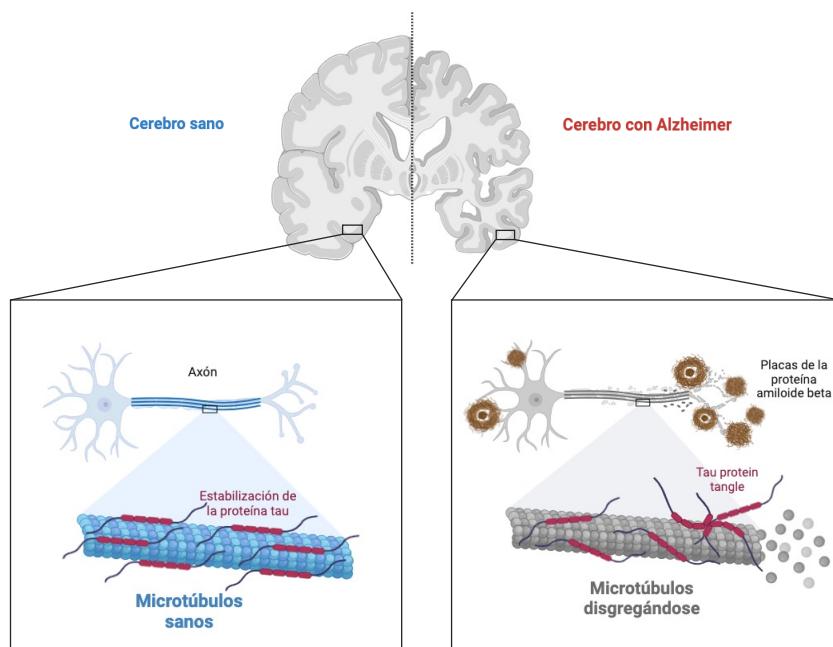


Figura 2. Estado fisiológico de un cerebro sano y un cerebro con Alzheimer, (Realizado con Biorender).

1.1.3. Signos clínicos y síntomas

Los principales dominios cognitivos que se ven afectados en el Alzheimer son la memoria, el lenguaje, la función visoespacial y la función ejecutiva. La gravedad del deterioro cognitivo causado por el Alzheimer puede variar desde la ausencia del deterioro cognitivo que correspondería con la fase asintomática, pasando por el deterioro cognitivo leve o fase prodromal y finalmente la demencia (**Figura 3**) [6].

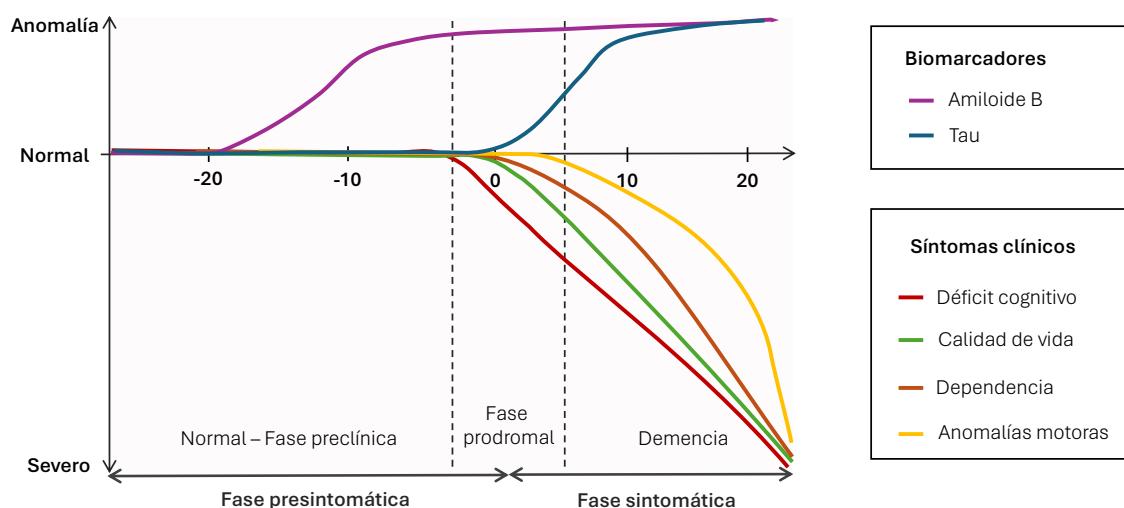


Figura 3. Representación esquemática de los cambios en los síntomas clínicos y en la acumulación de los marcadores del Alzheimer durante el progreso de la enfermedad, (Adaptado de [14]).

Un paciente típico con Alzheimer presenta un deterioro cognitivo leve con episodios amnésicos que progresa a grados variables de deterioro en el lenguaje, la cognición espacial, la función ejecutiva o la memoria de trabajo, y que va interfiriendo progresivamente en el funcionamiento diario de la persona, dando lugar a la demencia multidominio [6]. La aparición inicial y la progresión de los déficits cognitivos en el deterioro cognitivo leve típico del Alzheimer y la demencia siguen la propagación desde el lóbulo temporal medial hasta el isocórtex temporal lateral, parietal y frontal [6,15]. Los síntomas neuropsiquiátricos a menudo coexisten con los déficits cognitivos, de los cuales la depresión, la ansiedad y el aislamiento social pueden ser más evidentes en la demencia leve, mientras que en etapas más avanzadas pueden observarse delirios, alucinaciones, descontrol emocional o conductas físicamente agresivas [6,16].

Por otro lado, esta enfermedad se puede clasificar según el momento de aparición de los primeros síntomas. Aproximadamente entre el 1 y el 6 % de todos los casos se clasifican como Alzheimer de inicio temprano, que se manifiesta antes de los 65 años. Por el contrario, el Alzheimer de inicio tardío se caracteriza por la aparición de síntomas a una edad superior a los 65 años y representa alrededor del 90 % de los casos. Estos dos tipos de la enfermedad se diferencian en muchos aspectos, las presentaciones amnésicas son más comunes con la edad de inicio más avanzada (>70 años), mientras que la enfermedad de inicio temprano tiene un curso más agresivo, mayor retraso en el diagnóstico, menores reservas cognitivas, menor incidencia de diabetes, obesidad y trastornos circulatorios, déficits relativamente mayores en la atención, funciones ejecutivas, praxis y visión espacial, menor frecuencia del alelo APOE $\epsilon 4$, mayores cambios en la sustancia blanca y una mayor carga de placas neuríticas y ovillos neurofibrilares [17].

1.1.4. Criterios diagnósticos

En la actualidad, el diagnóstico del Alzheimer se realiza principalmente mediante pruebas cognitivas, neuroimagen y detección de distintos biomarcadores en el líquido cefalorraquídeo. El diagnóstico se basa en la evaluación clínica de los síntomas del paciente, un historial médico detallado y la exclusión de otras causas de demencia. Los criterios diagnósticos incluyen un deterioro cognitivo progresivo que afecta a varias áreas de la función cognitiva, especialmente la memoria, y la presencia de síntomas conductuales. Las pruebas neuropsicológicas y de imagen, como la resonancia magnética y la tomografía por emisión de positrones (PET), pueden ayudar a confirmar el diagnóstico, analizando la deposición de placas amiloideas, los ovillos neurofibrilares y la pérdida significativa de sinapsis que se observan en esta patología.

1.1.4.1. Biomarcadores

En cuanto a los biomarcadores, las pautas de diagnóstico han incluido los niveles de la proteína amiloide- β 1-42 (A β 42), la proteína tau total y tau hiperfosforilada (p-tau) en el líquido cefalorraquídeo [18,19]. Sin embargo, estos métodos son costosos y relativamente invasivos, además la sensibilidad y especificidad de A β 42 y p-tau han generado inquietudes sobre su implicación clínica, ya que la sensibilidad de A β 42 varía de 0,69 a 0,81 y la especificidad de 0,44 a 0,89 [20-22].

Además, los pacientes con Alzheimer son diagnosticados tardíamente y si esta enfermedad se pudiera detectar en etapas tempranas antes de que se desarrolle un daño cerebral importante, las terapias o tratamientos podrían ser más eficaces [23,24]. Esto remarca la importancia de identificar biomarcadores que puedan ayudar a detectar el Alzheimer de forma temprana o al inicio de la enfermedad. En este contexto, un biomarcador ideal debe ser específico, sensible, predictivo, preciso, robusto, económico e idealmente no invasivo, que pueda ser medible en fluidos biológicos comunes como suero, sangre, saliva y/o orina [17,25,26]. Para el Alzheimer, el líquido cefalorraquídeo se considera una fuente biológica óptima para la evaluación de biomarcadores, ya que su contacto directo con el líquido intersticial, donde está inmerso el cerebro, refleja los cambios fisiopatológicos de la progresión de la enfermedad en tiempo real [17,27]. Para los biomarcadores que se empleen en esta patología, se necesita una especificidad y una sensibilidad de más del 80% para ser considerados como un biomarcador confiable [17,28].

1.1.5. Justificación del trabajo

Dado lo expuesto anteriormente, la identificación de nuevos marcadores confiables resulta esencial para el diagnóstico temprano del Alzheimer, lo que a su vez permitiría desarrollar terapias más eficaces para frenar su progresión. En este sentido, la detección precoz de la enfermedad, idealmente antes de la aparición de los síntomas y mediante técnicas mínimamente invasivas y de bajo costo, ha impulsado la investigación de nuevos biomarcadores. Esto podría facilitar el desarrollo de estrategias de estratificación molecular que orienten tratamientos personalizados para los pacientes. Y por otro lado, el descubrimiento de nuevos biomarcadores no solo mejoraría el diagnóstico, sino que también contribuiría a esclarecer los mecanismos moleculares subyacentes de la enfermedad, los cuales aún no se comprenden por completo.

1.2. Objetivos

El objetivo principal del Trabajo fin de Máster es generar un modelo de aprendizaje automático que permita predecir el diagnóstico y la progresión de la enfermedad del Alzheimer. Esto permitiría la identificación temprana de esta patología con el fin de

mejorar el pronóstico, y encontrar posibles nuevas rutas implicadas en la enfermedad que puedan servir como diana para el desarrollo de tratamientos frente a esta enfermedad.

Para llevar a cabo este objetivo se plantean los siguientes objetivos específicos:

1. **Recopilación y preprocesamiento de los datos.** Esto permitirá obtener un *dataset* robusto que incluya los datos de los pacientes con Alzheimer, integrando información sobre la historia clínica y diferentes biomarcadores.
2. **Análisis y procesamiento de los datos.** Se realizará un análisis estadístico de las variables incluidas en el *dataset* a estudio con la finalidad de encontrar diferencias entre los pacientes con Alzheimer y las personas sanas. En el caso de que se disponga de un número muy grande de variables, en este paso se realizará la selección de características y también se realizará el procesamiento de los datos para que estén en el formato adecuado para poder completar el siguiente objetivo.
3. **Desarrollo de un modelo de aprendizaje automático.** Consistirá en implementar varios algoritmos de aprendizaje automático (como SVM, Random Forest, o KNN, entre otros) y determinar cuál de ellos es más efectivo en la predicción de la enfermedad del Alzheimer.
4. **Validación y evaluación del modelo.** Evaluar el rendimiento del modelo mediante técnicas de validación cruzada y métricas como la precisión, sensibilidad y área bajo la curva ROC (AUC), o el error cuadrático medio, con la finalidad de garantizar que el modelo sea preciso y generalizable.
5. **Interpretabilidad del modelo.** Utilizando técnicas de interpretabilidad de modelos se podrá fomentar que los resultados sean comprensibles para los profesionales clínicos y facilitar así su implementación.

1.3. Motivación personal

La principal motivación para la realización de este proyecto, que consiste en desarrollar un modelo de aprendizaje automático para predecir el diagnóstico de la enfermedad del Alzheimer, surge de una combinación de intereses académicos, científicos y personales. El objetivo es poder contribuir al avance en el conocimiento de esta enfermedad, que afecta en gran medida a la calidad de vida de las personas que la padecen y a su entorno. Por ello, la posibilidad de desarrollar herramientas que ayuden a predecir

y mitigar el impacto de esta enfermedad es un impulso esencial para llevar a cabo este proyecto.

Asimismo, considero que este trabajo es una excelente oportunidad para poner en práctica mis conocimientos en el campo de la biomedicina, así como los nuevos conocimientos que he ido adquiriendo a lo largo del Máster, específicamente en el ámbito del aprendizaje automático y la inteligencia artificial. Por lo que, la idea de emplear estas tecnologías para poder predecir el diagnóstico de determinadas enfermedades, y concretamente del Alzheimer, y que se pueda llegar a emplearse en el ámbito clínico me resulta especialmente interesante.

1.4. Sostenibilidad, diversidad y desafíos ético/sociales

Tal y como se puede observar en los apartados anteriores, el desarrollo de un modelo predictivo para la detección temprana del Alzheimer puede tener un impacto significativo en la sostenibilidad del sistema de salud. Un diagnóstico más temprano y preciso permitiría una intervención más eficaz, reduciendo la carga económica y social asociada con el tratamiento de la enfermedad en etapas avanzadas. Además, la optimización de los recursos médicos mediante herramientas basadas en inteligencia artificial puede disminuir la necesidad de pruebas invasivas y costosas, promoviendo un uso más eficiente de los recursos sanitarios. Sin embargo, también es importante considerar el impacto energético que va asociado con el procesamiento de grandes volúmenes de datos y el uso de infraestructura computacional avanzada.

Por otro lado, uno de los principales retos en el desarrollo de modelos predictivos en el ámbito de la salud es garantizar la equidad en su aplicación, lo que plantea diferentes desafíos éticos y sociales. La diversidad en los datos utilizados para entrenar el modelo es crucial para evitar sesgos que puedan llevar a diagnósticos menos precisos en ciertos grupos poblacionales. Factores como la edad, el género, el origen étnico y el nivel educativo pueden influir en la manifestación del Alzheimer, por lo que es fundamental garantizar que el modelo sea representativo de la población global. En el primer *dataset* empleado se recogen estas variables, y se puede comprobar como en cuanto al género, se recoge el sexo biológico de los participantes y ambos sexos están representados prácticamente en la misma proporción. El *dataset* factores de riesgo incluye datos de diferentes orígenes étnicos y niveles educativos, aunque en estos casos la proporción de

las diferentes categorías no es equitativa. Por otro lado, en cuanto a la edad, se incluyen pacientes en un rango de edad de 60-90 años, ya que este es el rango de edad en el que se suele manifestar la enfermedad (**Figura 4**). Para el *dataset* biomarcadores únicamente se incluye información sobre la edad y el sexo de los sujetos, y no sobre su origen étnico ni el nivel educativo. En cuanto al rango de edad, este segundo *dataset* tiene un rango de edad de 19-97 años, ya que también se incluyen controles jóvenes. La distribución de los sexos también se puede considerar que es equitativa (**Figura 4**).

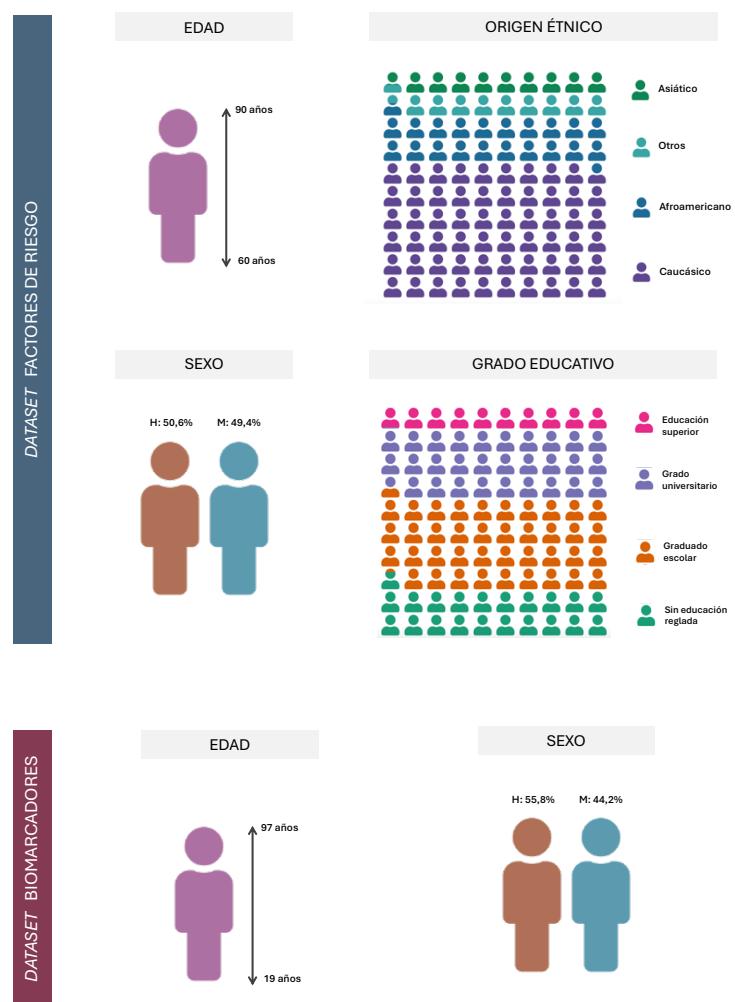


Figura 4. Distribución de la edad, sexo, origen étnico y grado educativo de los pacientes incluidos en el *dataset* factores de riesgo y biomarcadores (Realizado con Flourish).

Por último, los datos que se utilizarán en este proyecto, aunque son del ámbito de la salud, están completamente anonimizados y no contienen ninguna característica que pueda identificar a los pacientes. Este proceso de anonimización es fundamental para garantizar la privacidad y la protección de los datos, cumpliendo con el Reglamento

General de Protección de Datos (GDPR). Por ello, se ha presentado el formulario de no aplicabilidad en cuanto a las diferentes consideraciones éticas y de protección de datos, confirmando el compromiso de respetar los principios éticos y las normas legales que rigen las actividades de investigación del TFM, y declarando que en la investigación que conlleva la realización del TFM no hay participación humana ni ningún tratamiento de datos de carácter personal.

1.5. Enfoque y metodología

El proyecto planteado se basa en el análisis de diferentes datos, como pueden ser datos clínicos o datos obtenidos de estudios ómicos, obtenidos de pacientes con Alzheimer y de personas sanas. Como se puede observar esto dará lugar a lo que se conoce como Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*) [29]. Este modelo metodológico nos va a permitir, a partir de los datos, generar modelos válidos y generalizables, en este caso concreto capaces de predecir el diagnóstico de la enfermedad del Alzheimer.

Esta metodología se basa en los siguientes pasos [29]:

1. Conocer y comprender el contexto de la enfermedad para proponer soluciones viables.
2. Seleccionar los datos adecuados para realizar el estudio.
3. Preprocesamiento de los datos para garantizar la calidad de estos.
4. Procesamiento y transformación de los datos, para que se encuentren en el formato adecuado para su posterior análisis.
5. Selección de la tarea de minería de datos más apropiada según los objetivos planteados. En este caso son los métodos predictivos.
6. Selección de los algoritmos que permitan realizar la tarea escogida, en este caso la predicción. En el proyecto se plantea el uso de algoritmos como Random Forest, SVM, XGBoost o CatBoost, tanto en sus modalidades de clasificación como de regresión.
7. Aplicación de los algoritmos sobre los datos procesados.
8. Evaluación y validación de los modelos. Para ello se empleará la validación cruzada y diferentes métricas de precisión, sensibilidad, área bajo la curva ROC (AUC), o el error cuadrático medio. Si tras la evaluación del modelo los resultados no son óptimos se deberá regresar a etapas anteriores para realizar los ajustes oportunos.
9. Interpretación de los resultados con la finalidad de que la información obtenida durante el proceso pueda llegar a generar nuevo conocimiento que permita contestar a los objetivos planteados.

Durante el desarrollo del trabajo se irá especificando más concretamente la metodología empleada en cada uno de los pasos.

1.6. Planificación

1.6.1. Hitos

Según los objetivos que se plantean en este trabajo se han establecido los siguientes hitos:

- **Hito 1:** Obtención de los *datasets* de las diferentes fuentes.
- **Hito 2:** Obtención de los *datasets* preparados para su análisis.
- **Hito 3:** Conclusiones del análisis estadístico de los datos.
- **Hito 4:** Obtención de los *datasets* para el desarrollo e implementación de los modelos.
- **Hito 5:** Selección y caracterización e interpretación del mejor modelo predictivo y de las variables más relevantes.
- **Hito 6:** Presentación de los resultados y conclusiones.

1.6.2. Tareas

Las tareas planteadas para este trabajo se han establecido en función de los objetivos y diferentes hitos propuestos, y son las siguientes:

- **Tarea 1:** Obtención de los *datasets* adecuados para el objetivo del trabajo.
 - Búsqueda de *datasets* que permitan responder al objetivo del trabajo
 - Solicitar el acceso a los datos que no sean abiertos
 - Descarga de los *datasets*
 - Exploración inicial de los datos (cantidad de registros, variables incluidas, etc)
- **Tarea 2:** Procesamiento y limpieza de los datos.
 - Eliminación de registros duplicados
 - Gestión de los datos nulos
 - Eliminación de variables que no son necesarias para el resto de las tareas
- **Tarea 3:** Análisis estadísticos de los datos.
 - Análisis demográfico de los pacientes incluidos en el *dataset*
 - Correlación entre las variables incluidas en el *dataset*

- Estudio estadístico de las variables en función del diagnóstico de los pacientes
- **Tarea 4:** Desarrollo e implementación de los modelos predictivos.
 - Procesamiento de los datos para prepararlos para los modelos
 - Determinar los algoritmos que se emplearán para la implementación de los diferentes modelos
 - Selección de diferentes valores para los hiperparámetros de los modelos
 - Entrenamiento y selección de los mejores hiperparámetros para los modelos
 - Evaluación de los modelos
 - Selección del mejor modelo
 - Interpretabilidad del modelo
- **Tarea 5:** Visualización de resultados y descripción de las conclusiones.

En la figura 4 se puede observar la relación entre las tareas y los hitos planteados (**Figura 5**).

Tareas	Subtareas	Hitos
1. Obtención datasets	Búsqueda Acceso Descarga Exploración	Hito 1
2. Procesamiento de los datos	Duplicados Nulos Outliers Eliminación variables	Hito 2
3. Análisis estadístico	Demográfico Correlación Variables por diagnóstico	Hito 3
4. Modelos predictivos	Procesamiento de los datos Selección algoritmos Hiperparámetros Entrenamiento Evaluación Selección del mejor modelo Interpretabilidad del modelo	Hito 4 Hito 5
5. Presentación resultados	Visualización Conclusiones	Hito 6

Figura 5. Relación entre las tareas y los hitos planteados en el Trabajo Fin de Máster.

1.6.3. Análisis de riesgos

Los diferentes riesgos que puede presentar el desarrollo del proyecto y que puede suponer un retraso en la planificación del trabajo pueden ser el tiempo y algún posible fallo informático del software empleado o de otras plataformas necesarias para el desarrollo del trabajo. Otros riesgos más específicos podrían ser:

- Acceso a los datos:
 - Clasificación: Media
 - Descripción: No obtener la autorización para el acceso a los *datasets* que no son abiertos
 - Solución: Búsqueda de *datasets* alternativos de acceso abierto
- Procesamiento incorrecto de los datos:
 - Clasificación: Alta
 - Descripción: No conseguir un *dataset* robusto con datos de calidad por haber realizado de forma incorrecta la limpieza y el preprocesamiento de los datos
 - Solución: Comprobación y validación del *dataset* previo al análisis estadístico y al desarrollo de los modelos predictivos
- Errores en la implementación de los modelos:
 - Clasificación: Media
 - Descripción: Elección de los modelos adecuados y de sus hiperparámetros
 - Solución: Proponer una lista de modelos que puedan ser adecuados y entrenar los modelos con diferentes combinaciones de hiperparámetros
- Copia de seguridad:
 - Clasificación: Alta
 - Descripción: Pérdida de la memoria y del código desarrollado
 - Solución: Hacer copias de seguridad en varios dispositivos o usar la nube

1.6.4. Planificación temporal

Siguiendo las diferentes entregas propuestas por la organización del Máster para el trabajo final se propone el siguiente diagrama de Gantt con la planificación temporal del proyecto (**Figura 6**):

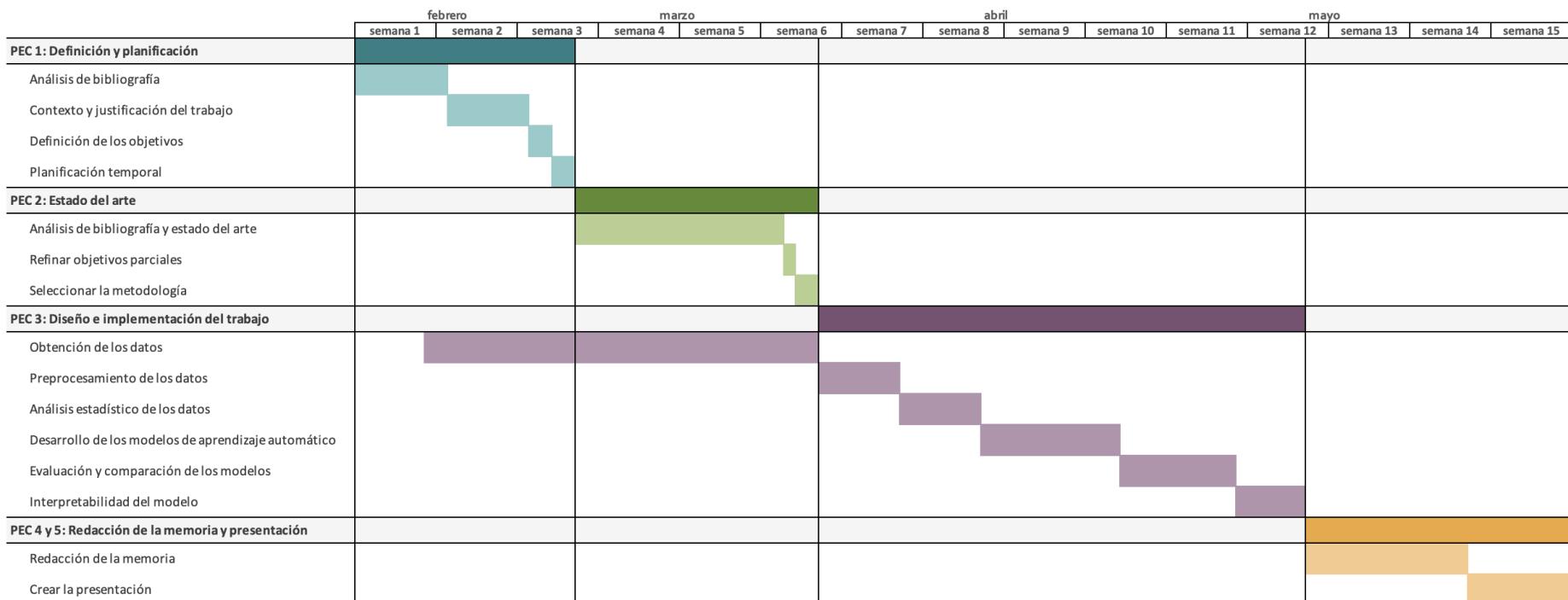


Figura 6. Diagrama de Gantt de la planificación del Trabajo Fin de Máster.

1.7. Resumen de los productos del proyecto

En cuanto a los productos esperados de este trabajo, el principal resultado será el desarrollo e implementación de un modelo predictivo capaz de detectar la enfermedad de Alzheimer a partir de un conjunto de datos específicos. Este modelo permitirá mejorar la precisión del diagnóstico temprano y contribuirá al desarrollo de herramientas computacionales para su aplicación en entornos clínicos.

Además, durante el proceso de desarrollo se generarán otros productos relevantes, como el código en lenguaje Python utilizado para el procesamiento y análisis estadístico de los datos, así como para la implementación de los modelos predictivos. Asimismo, se espera obtener conclusiones derivadas del análisis estadístico y la interpretación de los modelos, lo que podría conducir al descubrimiento de nuevas rutas implicadas en la enfermedad. Estos hallazgos no solo aportarían información valiosa sobre los mecanismos del Alzheimer, sino que también podrían contribuir al avance del conocimiento en este campo.

1.8. Breve descripción de los demás capítulos

El resto de la memoria se organizará en los siguientes capítulos:

- **Capítulo 2. Estado del arte.** En este capítulo se analizará la bibliografía de las técnicas más empleadas y que mejores resultados obtienen para la implementación de los modelos predictivos en el diagnóstico de la enfermedad del Alzheimer.
- **Capítulo 3. Metodología.** En este capítulo se detallará las técnicas empleadas durante el desarrollo del trabajo, tanto de hardware como software.
- **Capítulo 4. Resultados.** En este capítulo se analizarán los modelos implementados para la predicción del diagnóstico de la enfermedad, y se realizará la comparación de los diferentes algoritmos empleados. Además, en este apartado se analizarán las variables más importantes en la toma de decisiones de los modelos, que son las que contribuyen en mayor medida a la determinación del diagnóstico.
- **Capítulo 5. Conclusiones y líneas de trabajo futuro.** En este capítulo se presentarán las conclusiones obtenidas tras el desarrollo del trabajo, posibles limitaciones encontradas durante el mismo, así como posibles líneas de trabajo que podrían resultar prometedoras para mejorar los resultados obtenidos.

- **Capítulo 6. Glosario.** En este capítulo se incluirá un glosario de términos que se han empleado a lo largo de la memoria.
- **Capítulo 7. Bibliografía.** En este capítulo se incluirán las referencias empleadas para la elaboración del trabajo y la escritura de la memoria.
- **Capítulo 8. Anexos.** Este capítulo se incluirá en caso de que hiciera falta añadir alguna información específica que no se incluya en la propia memoria.

Capítulo 2

Estado del arte

El aprendizaje automático o *machine learning* es una de las técnicas de inteligencia artificial más empleadas que se utiliza en la clasificación, regresión, agrupamiento o modelado de datos. Los algoritmos de aprendizaje automático se pueden dividir en modelos supervisados en los que los datos están etiquetados y clasificados, en algoritmos no supervisados en los que el objetivo es separar datos no etiquetados en grupos de casos relacionados, y en algoritmos semisupervisados que incluyen datos etiquetados y no etiquetados [30]. Se han desarrollado una gran variedad de algoritmos de aprendizaje automático con diferentes características para poder abordar de forma eficaz los diferentes tipos de datos.

Estos algoritmos pueden ser útiles para la detección temprana y el análisis del Alzheimer, al igual que para otras enfermedades. El aprendizaje automático se ha empleado con éxito en el análisis de muchas modalidades de datos para enfermedades relacionadas con la demencia y para explorar diferentes biomarcadores [31]. Estos enfoques son clave para el análisis robusto de conjuntos de datos complejos y multimodales, para identificar nuevos patrones y posibles biomarcadores claves en la enfermedad [31]. De hecho, el uso de algoritmos de *machine learning* en el diagnóstico y predicción del Alzheimer ha experimentado avances significativos en los últimos años

[31,32]. Los diferentes estudios que han empleado algoritmos de aprendizaje automático en el campo del Alzheimer se pueden clasificar en función del tipo de datos que emplean para el entrenamiento de los modelos: información clínica, imágenes médicas, y datos de la voz de los pacientes [32].

Si se realiza una comparación entre los resultados obtenidos por los modelos que emplean los diferentes tipos de datos para la predicción del Alzheimer o la demencia, se observa que los modelos que obtienen un valor de precisión mayor son los que emplean como datos de entrada imágenes médicas [32]. Además, los algoritmos que han obtenido mejores resultados con este tipo de datos son *support vector machine* (SVM), *Bayesian Maximal Information Coefficient*, *Convolutional Neural Network* (CNN) y *Artificial Neural Networks* (ANN), con unos valores de precisión entorno al 97-100% [32]. En cuanto a los modelos que emplean la modalidad de datos de voz, estos son por lo general los que menor precisión consiguen (78,77-97,5%) [32]. En este caso, nos hemos centrado en los modelos que emplean variables clínicas para realizar la predicción del Alzheimer o de diferentes tipos de demencia, ya que son los datos de los que se dispone para desarrollar el trabajo. En la tabla 1 se recogen los diferentes estudios que han empleado diferentes algoritmos de *machine learning* con datos clínicos y las diferentes características de cada uno,

Tras el análisis de esta bibliografía se puede determinar que los estudios generalmente siguen la misma estructura en cuanto a los pasos a realizar (**Figura 7**). En primer lugar, se realiza la selección y obtención de los datos. Posteriormente se revisan los datos para determinar si faltan algunos valores, y en ese caso decidir cómo se procede, se pueden eliminar los registros a los que le faltan valores [33,34], o para no perder información realizar la imputación de los valores nulos [35-40] (**Figura 7**). Alguno de los algoritmos que se han empleado en las referencias analizadas para esta imputación son *Random Forest* (RF) [35,38], *k-nearest neighbors* (KNN) [38,40], *MissForest* [39], y en algunos casos también se ha empleado la imputación por el valor promedio [36].

Uno de los problemas que se plantea en los estudios de predicción del Alzheimer es el desbalance entre las clases de pacientes con la enfermedad y los individuos sanos, lo que por un lado puede complicar las tareas de predicción o también enmascarar buenos resultados de predicción, pero debido al alto porcentaje de la clase mayoritaria. Por ello, para tratar este problema se pueden emplear varias estrategias, o bien, eliminar registros de forma aleatoria de la clase mayoritaria para equilibrar el número de registros para

cada clase, o mediante técnicas de sobremuestreo (**Figura 7**) [35,41,42]. La técnica de sobremuestreo más común es SMOTE (*Synthetic Minority Over-sampling Technique*), que genera las nuevas muestras basándose en el algoritmo KNN [43].

Otro aspecto importante a tener en cuenta previo a la optimización de los modelos es la selección de variables, ya que en algunos casos se dispone de un número muy elevado de variables que pueden llegar a dificultar la tarea de los algoritmos. Por ello, se puede ver como en los estudios analizados cuando disponían de conjuntos de datos con una cantidad alta de variables se realizaba una selección de las variables más relevantes para el entrenamiento de los modelos (**Figura 7**) [34-37,39,40,44-49]. Existen diversas técnicas para realizar la selección de variables, como puede ser el algoritmo genético [39,40,48], ganancia de información [35,45], RelieF [35], CFSSubsetEval [36], Lasso [46], firma estadísticamente equivalente (SES) [47], centralidad de vector propio [49], análisis de componentes principales (PCA) [37,44] o por correlaciones [34,49].

Una vez están preparados los registros y las variables que se van a emplear para entrenar el modelo, el siguiente paso es realizar la separación de los registros para obtener los conjuntos de entrenamiento, validación y test (**Figura 7**). La proporción más común que se ha empleado en la bibliografía analizada está entorno a 70/10/20% para el entrenamiento, validación y test, respectivamente [35,40-42,46-48,50]. Posteriormente, con los datos preparados se realiza el entrenamiento de los modelos, en este paso se pueden emplear diferentes técnicas de validación que permitan optimizar el entrenamiento del modelo, la más empleada es la validación cruzada de 10 pasos (*10-fold cross-validation*) [35,38,42,50,51], aunque en algunos casos emplean la de 5, 4 o 3 pasos [34,39,41,46,48,52], en otros casos se emplea la validación dejando uno fuera (*leave-one-out validation*) [44,49] o el aprendizaje incentivando los valores predictivos positivos conocido como aprendizaje sensible al costo (*cost-sensitive learning*) [33] (**Figura 7**). Los modelos que se han empleado en la bibliografía son muy variados y generalmente en cada estudio siempre se desarrollan varios modelos para posteriormente seleccionar el que mejores resultados obtenga. Entre los algoritmos más empleados a lo largo de la bibliografía se encuentran SVM, RF, KNN, *Naïve Bayes* (NB), XGBoost, y algunas técnicas de *deep learning*. En la mayoría de estudios el algoritmo que mejor resultados obtiene es el RF, aunque el algoritmo J48 es el que consiguió una precisión más alta de todos los analizados [36]. El algoritmo J48 es un árbol de decisión simple, puede manejar datos numéricos y categóricos, puede trabajar con datos faltantes asignando probabilidades a las posibles ramas, utiliza la poda para reducir el sobreajuste y mejorar

la generalización, y permite ajustar la importancia de las diferentes características. Para generar los árboles de decisión este algoritmo selecciona el mejor atributo para dividir los datos usando la ganancia de información (basada en la entropía), posteriormente crea los nodos hijos dividiendo los datos en función del atributo seleccionado y repite estos pasos recursivamente hasta que todos los datos están clasificados o no haya más atributos que permitan dividir los datos, y finalmente realiza la poda del árbol para evitar el sobreajuste, eliminando las ramas que no sean necesarias [53,54].

Por último, una vez creado y optimizado el modelo se realiza su evaluación. Para este tipo de tareas de predicción las métricas más empleadas son la precisión, el área bajo la curva ROC (AUC), y la creación de la matriz de confusión para el cálculo de la sensibilidad y especificidad de los modelos. En cuanto a la precisión, los valores que se han obtenido están en el rango de 70,32–99,52%, y los valores de AUC entorno a 69–95% (**Tabla 1**). Además, en algunos casos también realizan la interpretación del modelo analizando cuales son las variables que tienen más relevancia en la toma de decisiones de los modelos desarrollados (**Figura 7**). Los modelos basados en árboles, como los árboles de decisión (DT), RF, XGBoost o LightBoost, son modelos fáciles de interpretar y que incorporan en su algoritmo el ajuste de la relevancia de las variables [47]. Sin embargo, hay otros modelos en los que estas tareas son más complicadas y por ello posteriormente se pueden aplicar técnicas de interpretabilidad, como puede ser el algoritmo de ganancia de información para determinar cuáles son las variables con mayor importancia [40,45].

En total, se han revisado 19 estudios, algunos de ellos se han centrado en predecir la demencia en los pacientes, mientras que otros en clasificar los diferentes estados del deterioro cognitivo (**Tabla 1**). Sin embargo, en muchos de estos estudios no se hace una distinción entre la enfermedad del Alzheimer y otros tipos de demencia, lo que es un aspecto clave para proporcionar el tratamiento más adecuado y mejorar la calidad de vida de los pacientes. Idealmente, se debería desarrollar un método de diagnóstico preciso, accesible y mínimamente invasivo. Como se ha visto en la bibliografía revisada, el uso de imágenes cerebrales de los pacientes ha permitido desarrollar modelos de alta precisión. No obstante, estas técnicas implican un coste alto y una elevada demanda computacional para el procesamiento de las imágenes y el entrenamiento de los modelos.

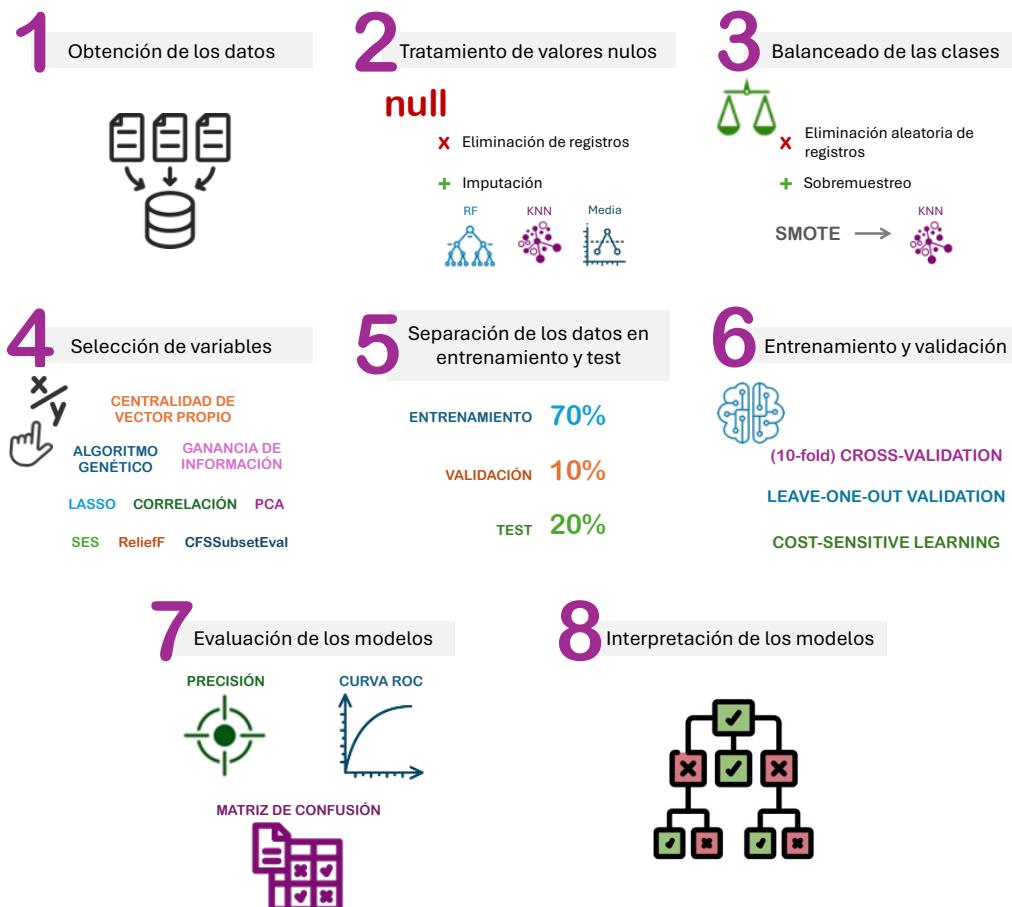


Figura 7. Estructura general de los pasos a seguir en los estudios de aprendizaje automático para la predicción del Alzheimer y otras formas de demencia.

Por ello, teniendo en cuenta todo lo mencionado anteriormente, este Trabajo Fin de Máster tiene como objetivo desarrollar modelos capaces de identificar la enfermedad del Alzheimer a partir de datos clínicos, ya que estos se pueden obtener de forma poca invasiva, a bajo coste y nos pueden permitir desarrollar modelos de bajo coste computacional. Además, el desarrollo de estos modelos no solo contribuiría a mejorar el proceso de diagnóstico del Alzheimer, sino que también podría permitir la identificación de nuevos factores de riesgo y, en última instancia, un mejor conocimiento de esta patología.

Tabla 1. Características de los estudios que emplean el aprendizaje automático para la detección y predicción del Alzheimer.

Ref	Año	Nº registros	Nº Variables	Objetivo	Algoritmo	Entrenamiento/test	Entrenamiento	Evaluación
[44]	2017	C = 88 / AD = 70	No se especifica	Predicción diagnóstico y clasificación	SVM	No se especifica	<i>Leave-one-out validation</i>	ACC = 82%
[35]	2018	C = 617 / MCI = 886 / Demencia= 348	49	Predicción diagnóstico y clasificación	RF, SVM, <i>Gaussian Processes, Stochastic Gradient Boosting, XGBoost</i>	75 / 25 %	<i>10-fold cross-validation</i>	XGBoost ACC = 91% (<i>Monte Carlo simulation</i>)
[36]	2018	T = 416	373	Predicción	J48, NB, RF, DL	No se especifica	No se especifica	J48, NB, DL ACC = 99,52%
[37]	2018	T = 18165	92	Clasificación (no supervisado)	Clustering jerárquico	No aplica	No aplica	Clasificación del cluster 3 ACC = 93,1% (AUC = 91%)
[45]	2019	C = 53 / MCI = 91 / VMD = 108 / Demencia= 386 (AD = 202)	45	Predicción	<i>Information Gain</i>	53 / 47 %	No se especifica	AUC = 95% (95–98%)

Ref	Año	Nº registros	Nº Variables	Objetivo	Algoritmo	Entrenamiento/test	Entrenamiento	Evaluación
[38]	2019	C = 242 / AD = 115	883	Predicción	DL, XGBoost	No se especifica	<i>10-fold cross-validation</i>	XGBoost ACC = 88% (<i>Monte Carlo simulation</i>)
[46]	2019	C = 760646 / AD = 44945	10000	Predicción	RF	80 / 20 %	<i>4-fold cross-validation</i>	AUC = 69,3%
[51]	2019	C = 101 / MCI = 68	14	Predicción	LR, SVM, KNN, NB, RF, DL	No se especifica	<i>10-fold cross-validation</i>	<i>Gradient Boosting</i> ACC= 93%, RF ACC = 92%
[41]	2019	C = 4547 / MCI = 1376	37	Predicción	LSTM <i>recurrent neural network</i> , RF	70 / 10 / 20 %	<i>5-fold cross-validation</i>	RF <i>Over-sampled</i> ACC = 79 % (AUC = 69 %) / LSTM <i>Over-sampled</i> ACC = 71% (AUC = 75%)
[47]	2020	mRNA (C = 22 / AD = 48; C = 100 / AD = 134), Proteómica (C = 37 / AD = 25)	mRNA (506; 38327) / PROTEOMICA (9483)	Predicción	RF, SVM, DT, <i>Ridge Logistic Regression</i>	70 / 30 %	No se especifica	AUC = 97,5% (mRNA SVM), AUC = 84,6% (mRNA RF), AUC = 92,1% (Proteómica <i>Ridge Logistic Regression</i>)
[33]	2020	C = 544759 / AD = 136189	7391	Predicción	LightGBM	40 / 20 / 40 %	Aprendizaje por valores predictivos positivos	AUC = 94%

Ref	Año	Nº registros	Nº Variables	Objetivo	Algoritmo	Entrenamiento/test	Entrenamiento	Evaluación
[34]	2020	T = 566	33	Predicción	XGBoost	No se especifica	<i>5-fold cross-validation</i>	ACC = 85,61%
[48]	2021	C = 16 / AD = 21	1492	Predicción	DL, SVM	70 / 15 / 15 %	<i>5-fold cross-validation</i>	SVM AUC = 90%
[50]	2021	T = 4096; DD = 2336 / LD = 192	No se especifica	Predicción	<i>Stochastic Gradient Boosting, RF, SVM, Elastic net, Multivariate Adaptive Regression Splines, DL</i>	70 / 30 %	<i>10-fold cross-validation</i>	ACC = 95-96%
[42]	2021	C = 802 / Demencia = 200	No se especifica	Predicción	LR, DT, RF, KNN, SVM	80 / 20 %	<i>10-fold cross-validation</i>	Balanced RF AUC = 87,6%
[40]	2021	C = 38 / MNCD = 46	37	Predicción	SVM, RF, ANN, AdaBoost, LDA, LR	60 / 20(10/10) / 20 %	No se especifica	RF ACC = 88% (AUC = 92%)
[39]	2022	C = 87 / AD = 170 / bvFTD = 72	No se especifica	Predicción diagnóstico y clasificación	DT, RF, SVM, NB, AdaBoost, <i>Gradient Boost</i>	No se especifica	<i>5-fold cross-validation</i>	ACC = 86,8% (DT: AD vs C)

Ref	Año	Nº registros	Nº Variables	Objetivo	Algoritmo	Entrenamiento/test	Entrenamiento	Evaluación
[52]	2022	C = 46 / AD = 77	No se especifica	Predicción	DT, RF, SVM, NB, LR, LDA, KNN, DL	No se especifica	<i>3-fold cross-validation</i>	DL ACC = 70,32%
[49]	2022	C = 30 / MCI = 30	23	Predicción	DT, RF, SVM, DL	No se especifica	<i>Leave-one-out validation</i>	SVM ACC = 71,67%

C: control; AD: Alzheimer; MCI: *Mild cognitive impairment*; T: totales; VMD: *very-mild dementia*; DD: Diagnóstico con demencia; LD: Viviendo con demencia; MNCD: *Major neurocognitive disorder*; bvFTD: *behavioural variant frontotemporal dementia*; SVM: *Support vector machine*; RF: *Random Forest*; NB: *Naïve Bayes*; DL: *Deep learning*; LR: Logistic Regresion; KNN: *K-nearest neighbors*; LSTM: *Long Short Term Memory*; DT: *Decision Tree*; ANN: Artifical neural network; LDA: *Linear Discriminant Analysis*; ACC: Precisión; AUC: área bajo la curva ROC.

Capítulo 3

Metodología

3.1. Descripción de los *datasets*

Se han empleado dos *datasets* para el desarrollo del proyecto, uno en el que se incluyen diferentes factores de riesgo que pueden ayudar a predecir el desarrollo del Alzheimer, y el segundo con datos ómicos de biomarcadores, concretamente de un estudio de proteómica.

3.1.1. Dataset factores de riesgo

Esta base de datos está disponible públicamente y se ha obtenido de (<https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset>). Este conjunto de datos contiene información clínica sobre la salud de 2149 pacientes, con o sin diagnóstico de Alzheimer, con una media de edad de $74,91 \pm 8,99$ (60 – 90) años, y un 50,6% de mujeres frente al 49,4% de hombres (**Figura 4**). Se dispone de 35 variables con información demográfica, factores de estilo de vida, historial médico, evaluaciones cognitivas y funcionales, síntomas y si han sido diagnosticados de Alzheimer (variable objetivo).

3.1.2. *Dataset* biomarcadores

Este *dataset* está disponible públicamente y procede del repositorio Gene Expression Omnibus (GEO) con el identificador GSE29676. Este conjunto de datos se ha generado a partir de muestras de sangre de pacientes con Alzheimer, Parkinson o cáncer de mama y también con muestras de personas sanas de diferentes edades. Hay un total de 609 registros o pacientes con una media de edad de $64,91 \pm 20,48$ (19 – 97) años, y un 55,8% de hombres frente a un 44,2% de mujeres (**Figura 4**). El *dataset* extraído dispone de 9486 variables, en las que se incluye el ID de los pacientes, el diagnóstico, la edad, el sexo, el MMSE, y 9480 proteínas analizadas.

3.2. Preprocesamiento de los datos

Las tareas o técnicas de preparación de datos están orientadas a la adecuación del conjunto de datos para que posteriormente pueda ser empleado en los modelos de predicción. Estas tareas son importantes ya que la eficacia de los modelos va a depender en gran medida de la calidad de los datos con los que se entrene.

Con el fin de obtener unos conjuntos de datos de calidad hay que tener en cuenta una serie de aspectos que se detallan en los siguientes apartados.

3.2.1. Gestión de valores nulos

Los valores nulos son datos ausentes o faltantes, por lo que es uno de los problemas más habituales en la minería de datos. Dependiendo de la proporción de valores nulos que contenga cada variable se puede proceder de diferentes formas. Si una variable incluye en su mayoría valores nulos esa variable debería ser excluida para el desarrollo de los modelos. Si hay registros que incluyen valores nulos para diversas variables se opta por eliminar esos registros del conjunto de datos. Por el contrario, cuando hay variables con valores nulos, pero estos están dentro de una proporción razonable se pueden aplicar técnicas de imputación para que no se produzca la pérdida de datos [55-57].

Los *datasets* empleados en este trabajo no incluían valores nulos en las variables que se querían emplear para el desarrollo de los modelos de predicción.

3.2.2. Análisis de correlaciones

La información redundante también puede llegar a ser un problema, ya que incrementa el tamaño del conjunto de datos, lo que conlleva una mayor carga computacional en la fase de construcción de los modelos y un incremento del riesgo de sobreajuste [58,59].

Para detectar si hay variables con información redundante se ha realizado un análisis de correlaciones. Para ello, se ha aplicado el coeficiente de correlación de Spearman, que mide la fuerza y dirección de una relación monotónica entre dos variables. No asume una relación lineal ni la normalidad de los datos. Una vez realizado el análisis, las variables altamente correlacionadas ($|\rho| \geq 0,80$) han sido eliminadas.

3.2.3. Desbalanceado de clases

El desbalanceado de clases se refiere a la situación en la que las categorías objetivo no están representadas de forma equitativa. Este desbalance puede generar varios problemas, uno de ellos es que los algoritmos de clasificación tienden a favorecer la clase mayoritaria, ya que clasificando las instancias simplemente como la clase mayoritaria minimizan el error global. Por ello, las métricas tradicionales como la precisión del modelo pueden no ser representativas del rendimiento real del modelo. Por otro lado, las clases minoritarias no se clasifican correctamente e incluso puede producirse un *overfitting* en estas categorías por no disponer de suficientes datos para generalizar. Por norma general se considera que un conjunto de datos está desbalanceado cuando una clase representa menos del 30%, o si hay ratios de 1:4 o mayores [60,61].

Para abordar este problema se pueden aplicar diferentes técnicas, como el sobremuestreo (oversampling) de las clases minoritarias, el submuestreo de las clases mayoritarias, la generación de datos sintéticos, y el uso de métricas adecuadas. Como en este caso no se dispone de un número muy alto de registros, no se puede perder información, por lo que se va a proceder a la generación de datos sintéticos para realizar el sobremuestreo de las clases minoritarias. Para ello, se ha empleado la técnica de Synthetic Minority Over-Sampling Technique (SMOTE). Esta es una técnica que genera registros sintéticos a partir de los registros existentes, mediante la interpolación de los registros de las clases minoritarias y sus vecinos más cercanos en el espacio de características ($k=5$) [43,58,62].

3.2.4. Limpieza y transformación del *dataset* factores de riesgo

Este *dataset* está bastante limpio y completo, sin embargo, a continuación, se muestran los pasos que se han seguido para obtener el conjunto de datos preparado para el entrenamiento de los modelos de predicción:

1. Se comprueba si hay valores nulos en alguna de las variables incluidas.
2. Se determina que los datos que incluye cada variable estén en el formato adecuado según se precise.
3. Comprobar si hay registros duplicados.
4. Se eliminan las columnas con información innecesaria como “PatientID” y “DoctorInCharge”.
5. Se comprueba que todas las variables incluyen valores dentro del rango esperado para cada una de ellas.
6. Se comprueba si hay variables que no aporten información, por ejemplo si tienen varianza cero o próxima a 0 ($< 0,01$).
7. Se calcula el coeficiente de correlación de Spearman, y se comprueba que no haya una correlación fuerte entre las variables incluidas en el *dataset*, descartando así que no se tenga información redundante.
8. Se comprueba la proporción de registros clasificados con Alzheimer y sin la enfermedad, y se obtiene una proporción de 64,62% sin la enfermedad y 35,38%. En este caso como se obtiene una proporción de 1,82:1, se considera que la muestra está un poco desbalanceada pero es una proporción asequible para poder trabajar con ella [60,61].
9. Finalmente, previo a la partición de los datos en los conjuntos de entrenamiento y test se realiza la estandarización de las variables, para que estén en un formato más adecuado para el entrenamiento de los modelos y no interfiera las diferentes escalas que pueda haber entre las diferentes variables y que puedan afectar a los resultados obtenidos con los modelos.

3.2.5. Limpieza y transformación del *dataset* biomarcadores

En este caso el archivo de texto que contiene los datos no está preparado para leerlo directamente, sino que hay que extraer la parte donde se encuentran las variables y los registros que se quiere incluir. El archivo de texto original contiene un encabezado en el que se incluye información sobre el estudio que ha producido esa base de datos y

posteriormente se incluye en la primera columna el nombre de las variables incluidas en la base de datos y en el resto de las columnas los diferentes valores para esa variable, que corresponde a los valores de los diferentes registros. Por lo tanto, para la obtención de la base de datos se han realizado los siguientes pasos:

1. Se selecciona el índice a partir del cual se encuentran nuestros datos de interés ("!Sample_title") y se realiza la transposición de las columnas con las filas, para obtener de esta forma el nombre de las variables en las columnas de nuestro *dataset* y en las filas los diferentes registros.
2. Para preprocesar o analizar los datos es importante saber cuáles son las variables con las que se está trabajando. Por lo tanto, el siguiente paso fue renombrar las variables de nuestro interés que no tenían nombres descriptivos (ej: "Sample_characteristics_ch1" pasará a "Sample_age"), y depurar el contenido de las mismas (ej: "age: 83" pasará a "83").
3. Eliminar las columnas que no son relevantes para el estudio (ej: "Sample_status", "Sample_tissue", "Sample_contact_country").
4. Se comprueba que los datos que se incluyen en cada variable son consistentes entre sí.
5. Modificación de los campos sin datos, de modo que todos los datos nulos aparezcan como valores NaN. Para ello, se reemplazan los valores "UNK" en la columna "Sample_age" por valores nulos.
6. Se comprueba que los valores únicos para los identificadores de las muestras coinciden con los registros incluidos en el *dataset*, y que por tanto no hay duplicados.
7. El siguiente paso fue factorizar las columnas del género y el diagnóstico.

Género: {"Male": 0, "Female": 1}

Diagnóstico: {"Older Control": 0, "Younger Control": 1, "Alzheimer's Disease": 2, "Breast Cancer": 3, "Parkinson's Disease": 4}

8. Se comprueba que columnas incluyen valores nulos.
9. Se comprueba que los datos para las columnas de la edad y el MMSE se encuentren dentro del rango adecuado.
10. Finalmente se comprueba si hay columnas que tengan varianza cero o próxima a cero ($< 0,01$), para en ese caso eliminarlas ya que estas no aportan información.
11. Se comprueba la proporción de registros que se disponen para cada clase (**Tabla 2**). Puesto que en este caso hay un desbalance considerable de las distintas clases de proporción 2,9:1 a 12:1, se realizará previo al entrenamiento de los modelos un *oversampling* de las clases menos representadas mediante la técnica SMOTE.

Tabla 2. Distribución de registros en las diferentes clases de diagnóstico del *dataset*.

Diagnóstico	% (valor)
Alzheimer	57,47 (350)
Controles mayores	19,70 (120)
Controles jóvenes	13,14 (80)
Cáncer de mama	4,93 (30)
Parkinson	4,76 (29)

En este *dataset* también se aplica la estandarización de los datos, sin embargo, en función del método de selección de características que se utilice se realiza previo a la selección d características con los métodos reliefF y wrapper, o posterior a la selección de las mismas, mediante el método del filtro por el p-valor.

3.3. Métodos de selección de características

Otro de los problemas a los que se enfrentan las tareas de minería de datos es la cantidad de variables disponibles para la construcción de los modelos. Trabajar con un número de variables demasiado alto puede generar ruido y hacer que los modelos reduzcan su rendimiento, suponen un coste computacional elevado y hay un mayor riesgo de sobreajuste, además cuantas más variables se dispongan se necesitan más registros para que los modelos puedan aprender correctamente, ya que si no aumenta el riesgo de que los modelos no generalicen. Por lo tanto, la reducción del número de variables para la construcción de los modelos es fundamental, ya que permite reducir la complejidad de los modelos, disminuye el tiempo de entrenamiento y los recursos necesarios, y ayuda a mejorar la precisión de los modelos eliminando variables irrelevantes o redundantes, y, por tanto, permitiendo que el modelo se centre en la información realmente útil y que se reduzca, de la misma manera, el riesgo de sobreajuste [58,59,63].

En este trabajo se han aplicado dos métodos para la selección de características sobre el *dataset* de biomarcadores, el filtrado de las variables que son estadísticamente significativas entre los grupos que se quieren clasificar y la selección de características mediante el uso de reliefF y wrapper, que se detallan a continuación.

3.3.1. Métodos de filtro por p-valor

Este método se basa en la idea de que no todas las variables presentan diferencias entre los distintos grupos que se quieren clasificar, en este caso para poder diferenciar los pacientes con Alzheimer del resto de grupos incluidos en el *dataset*.

Para llevar a cabo este método en primer lugar se ha realizado una prueba estadística (Shapiro-Wilk test) para comprobar la normalidad de los datos que incluyen las diferentes variables. Debido a los resultados obtenidos con este test, para el análisis de las diferencias significativas entre los dos grupos establecidos se emplea el test de Mann-Whitney, mediante el que se obtiene el p-valor para cada variable, de forma que si este es menor a 0,05 se considera que hay diferencias significativas para esa variable entre los dos grupos de pacientes analizados [55,64]. Posteriormente, se aplica la corrección de Bonferroni, que se emplea para contrarrestar el problema de las comparaciones múltiples. Este problema radica en que al realizar múltiples tests de hipótesis, se aumenta la probabilidad de obtener al menos un resultado estadísticamente significativo que sea por casualidad. Esta corrección consiste en modificar el umbral del p-valor por el que se considera que es estadísticamente significativo dividiendo el valor original por el número de pruebas que se realiza, o lo que es lo mismo, multiplicar el p-valor obtenido en cada test por el número total de tests realizados y emplear el umbral original [65].

Para finalizar la selección de características además se ha aplicado un umbral de *fold-change*, que representa la magnitud de la diferencia que hay entre ambos grupos analizados, ya que cambios muy tenues, aunque significativos, también podrían aportar información irrelevante. Para ello, en este caso se ha empleado un $\log_2 \text{fold-change} > 0,5$ [66].

3.3.2. Métodos reliefF y wrapper

Otro método que se ha empleado para la selección de características es la aplicación consecutiva de la técnica reliefF y posteriormente el wrapper. Este método es útil y complementario al del filtro del p-valor ya que en el caso del filtro por el p-valor se podrían estar seleccionando variables que actúen de forma similar entre los diferentes grupos, por lo que, aunque se consideren variables relevantes no tienen por qué ser las variables más útiles para los modelos de predicción [55,59,63].

Los algoritmos de la familia reliefF son ampliamente reconocidos y empleados. Estos algoritmos son capaces de detectar dependencias entre características, utilizando el concepto de vecinos más cercanos para evaluar la relevancia de cada variable en función

de su capacidad para distinguir entre instancias cercanas pertenecientes a clases diferentes. Para ello el algoritmo reliefF recorre un número concreto de instancias de forma aleatoria, busca el vecino más cercano de su misma clase (*Nearest hit*) y el vecino más cercano de una clase diferente (*Nearest miss*). Para cada característica, el algoritmo evalúa si su valor es más similar al del vecino de la misma clase o al de la clase diferente. Cuando una variable tiene valores más consistentes entre instancias de una misma clase y diferente entre instancias de distintas clases, se considera discriminativa, y se le asigna una mayor puntuación de relevancia. Por lo tanto, el resultado obtenido con este algoritmo es una lista de características puntuadas, que pueden ser seleccionadas según un umbral o por orden de importancia. En este caso, tras aplicar el algoritmo reliefF se han seleccionado las 500 variables más relevantes, para usar estas como entrada del siguiente paso en la selección de características, que será el método wrapper [59,63].

El método wrapper se basa en el uso de un algoritmo de *machine learning* para evaluar distintos subconjuntos de características y seleccionar el que obtiene el mayor rendimiento del modelo. Para evaluar las diferentes combinaciones de variables que se podrían obtener, puesto que evaluar todas las posibles combinaciones sería muy costoso, se pueden usar varias estrategias, la selección hacia delante (*forward selection*), la selección hacia atrás (*backward elimination*) y la eliminación recursiva. En este caso, se ha aplicado como modelo el RandomForest Classifier con eliminación recursiva de las características menos importantes empleando la validación cruzada estratificada. Este modelo elimina una característica en cada interacción reentrena el modelo y evalúa el rendimiento, la precisión, mediante validación cruzada. Este proceso se repite hasta encontrar un subconjunto óptimo de variables que maximiza la precisión del modelo. Por lo tanto, el número óptimo de variables lo selecciona el modelo en función de los resultados y no es una decisión previa o un parámetro definido [55,59,63,67].

3.4. Modelos de predicción

3.4.1. Decision Tree

Los árboles de decisión son uno de los modelos más empleados, debido a su capacidad explicativa y que son modelos fáciles de interpretar. Este modelo se puede usar tanto en problemas de clasificación como de regresión, supervisados.

Los árboles de decisión tratan de subdividir el espacio de datos de entrada al modelo de forma que en las subdivisiones generadas todas las muestras pertenezcan a la misma clase. Si una subdivisión contiene muestras de más de una clase se irán haciendo particiones para ir separando las muestras de las diferentes clases. El modelo finalizará

cuando todas las particiones creadas contengan muestras de una sola clase. Los árboles de decisión constan de nodos hoja o terminales y nodos internos o splits. El árbol empieza por el nodo raíz que contiene una condición la cual determinará por qué rama del árbol debe ir la muestra. Una vez determinada la rama, la muestra llegará a un nodo con otra condición (nodo interno) o a un nodo terminal asignándole la etiqueta de este último nodo. Un aspecto importante del modelo es determinar cuál es la mejor secuencia de las variables para realizar las particiones [54,68-70].

3.4.2. Random Forest

Los árboles de decisión son modelos sencillos, pero tienen un riesgo alto de sobreajuste, para solucionar este problema se desarrollaron los algoritmos Random Forest. Este modelo pertenece a los métodos de ensemble que combinan las predicciones de varios modelos base creados con un mismo algoritmo de aprendizaje automático para mejorar la generalización y robustez de un solo modelo. El Random Forest consiste en una colección de árboles de decisión que se han entrenado con subconjuntos de registros y variables seleccionados aleatoriamente. Este modelo realiza las predicciones devolviendo la media de las predicciones de los árboles que lo constituyen [69-71].

3.4.3. Support Vector Machine

El algoritmo Support Vector Machine (SVM) está enfocado a tareas supervisadas y es ampliamente utilizado para problemas de clasificación y, aunque en menor medida, en problemas de regresión. El objetivo de este algoritmo es encontrar un hiperplano que pueda separar las diferentes clases de los datos. Este hiperplano también busca que se dé la distancia máxima entre el límite de decisión (hiperplano) y las muestras de cada clase. Cuando hay más de dos categorías, el problema se resuelve mediante la reducción a varios problemas binarios más simples [69,70,72].

En los problemas de regresión este modelo funciona diferente, ya que su objetivo no es encontrar un hiperplano de separación, sino crear una función que prediga los valores dentro de un margen de tolerancia definido (ϵ). En este caso se busca ajustar una función lo más plana posible que mantenga la mayor cantidad de puntos dentro del margen de tolerancia definido alrededor de la función [69,70,72].

Los modelos basados en SVM funcionan bien con conjuntos de datos de alta dimensión, y en los casos en los que el número de muestras es limitado en comparación con el número de características. Además, este modelo tiene la ventaja de que puede

trabajar también en problemas no lineales, mediante el uso de *kernel*. Los *kernel* son funciones que permiten transformar los datos originales en un espacio de mayor dimensión donde pueden ser separados linealmente [72]. Los *kernel* que se han evaluado en este *estudio* son el radial, el polinomial, y el sigmoideo.

3.4.4. K-nearest neighbors

El algoritmo k-nearest neighbors (KNN) o k-vecinos más cercanos es uno de los algoritmos más simples, aunque suele tener mejores resultados que otros algoritmos más complejos, y se puede usar para tareas de clasificación y de regresión supervisadas [68-70].

Este algoritmo no incluye una fase de entrenamiento, ya que para cada muestra nueva por clasificar calcula su distancia con todas las muestras de entrenamiento y selecciona las k muestras más cercanas, posteriormente para realizar la clasificación escoge la etiqueta que más se repite entre sus k vecinos seleccionados. Para determinar cuáles son las k muestras más cercanas se pueden aplicar diferentes métricas de distancia, en este caso se ha utilizado la distancia euclídea estándar. Teniendo en cuenta las características de este algoritmo no se genera un modelo final, sino que cada vez que se requiere clasificar una muestra el algoritmo tiene que recorrer todo el resto de los datos para seleccionar los k vecinos, y por tanto, esto computacionalmente puede llegar a ser muy costoso [68-70].

3.4.5. Logistic Regression

El modelo Logistic Regresion es un modelo estadístico ampliamente utilizado en tareas de clasificación binaria, pero también puede ampliarse a problemas de clasificación multiclase. Este algoritmo estima la probabilidad de que un registro pertenezca a una determinada clase, a partir de un conjunto de variables independientes. Emplea la función sigmoide o logística buscando la máxima verosimilitud, ajustando los valores de los coeficientes que maximicen la probabilidad de observar los datos reales. Una vez se ha entrenado el modelo, clasifica las muestras en función de si la probabilidad obtenida es superior al umbral determinado, generalmente 0,5. Este modelo presenta problemas cuando los datos no son lineales, son complejos, o existe una alta multicolinealidad entre variables [69,70].

3.4.6. Ridge Regression

La regresión de Ridge es una técnica de regresión lineal que incorpora el término de penalización o regularización, con la finalidad de reducir el sobreajuste y mejorar la generalización del modelo. Este algoritmo no busca minimizar la suma de los errores de la predicción sino una función de coste que incorpora una penalización según la magnitud de los coeficientes. La fuerza de la penalización se controla con el parámetro λ . El objetivo principal es restringir el tamaño de los coeficientes para evitar que el modelo se adapte en exceso a los datos de entrenamiento, reduciendo la varianza del modelo sin aumentar el sesgo. Esta modificación de la regresión lineal es especialmente útil para conjuntos de datos que presentan multicolinealidad o cuando el número de variables es grande en comparación con el número de registros [69,73].

3.4.7. XGBoost

El algoritmo Extreme Gradient Boosting (XGBoost) es otro método de ensemble, como el Random Forest, pero en este caso está basado en el enfoque de *boosting* por gradiente para construir el modelo predictivo final, haciéndolo más potente y eficaz que los modelos individuales. El algoritmo consiste en realizar múltiples árboles de decisión poco profundos, pero en este caso en lugar de entrenarlos de forma paralela los entrena de forma secuencial, y así cada nuevo árbol corrige los errores del anterior. En cada iteración, el algoritmo de XGBoost trata de minimizar una función de pérdida que representa como se ajusta el modelo a los datos, y un término de regularización que penaliza la complejidad del modelo para evitar el sobreajuste. Este modelo se está empleando mucho actualmente, debido a su alto rendimiento, y a su capacidad de generalización y velocidad de entrenamiento respecto a otros algoritmos de *boosting*, además de que proporciona métricas sobre la importancia de las variables, permitiendo así su interpretabilidad [74].

3.4.8. CatBoost

El algoritmo Categorical Boosting (CatBoost) es otro método de ensemble de *boosting* por gradiente basado en árboles de decisión. Este algoritmo ha sido optimizado para trabajar de forma más eficiente con variables categóricas y para realizar el boosting ordenado, mejorando así su rendimiento y su capacidad de generalización. Por lo tanto, al igual que XGBoost, este algoritmo construye árboles de decisión de manera secuencial

de forma que se van corrigiendo los errores de los árboles anteriores, pero en este caso CatBoost convierte internamente las variables categóricas mediante técnicas basadas en estadísticas de orden, permitiendo conservar la información y reduciendo la dimensionalidad. Para evitar el sobreajuste derivado del uso de la variable objetivo en la transformación de las variables categóricas, CatBoost emplea un orden aleatorio y calcula promedios acumulativos que mitigan la fuga de información. Sin embargo, no solo incluye esta implementación, sino que también aplica técnicas para mejorar la generalización y reducir el coste computacional durante el entrenamiento. Para ello, este modelo construye árboles de decisión simétricos, donde cada nivel de profundidad del árbol utiliza el mismo conjunto de reglas para todas las ramas. Por otro lado, emplea el método de *boosting* ordenado, de forma que impide que los árboles posteriores utilicen los mismos datos para el cálculo del error y la predicción del modelo, lo que ayuda a reducir el riesgo de sobreajuste, sobre todo en conjuntos de datos pequeños [75].

3.5. Validación y evaluación de los modelos

3.5.1. Ajuste de hiperparámetros

Para la construcción y optimización de los modelos hay que establecer una serie de parámetros en cada modelo, denominados hiperparámetros, que no se aprenden durante el proceso de entrenamiento, sino que tienen que ser definidos previamente. El ajuste de estos hiperparámetros es crucial para maximizar el rendimiento de los modelos, y para ello se ha empleado la técnica Grid Search (búsqueda en cuadrícula), que consiste en entrenar los modelos con todas las combinaciones posibles del conjunto de hiperparámetros de cada modelo. Para cada combinación se entrena el modelo y se evalúa mediante la validación cruzada, para finalmente seleccionar la combinación de hiperparámetros que mejor resultado ha obtenido en cuanto a la precisión global del modelo [69,71].

A continuación, se muestran los diferentes valores de hiperparámetros que se han analizado para los modelos empleados:

- Decision Tree:
 - 'max_depth' = [3, 4, 5, 6, 7, 12, None]
- Random Forest:
 - 'n_estimators': [50, 100, 200]
 - 'max_depth': [3, 5, 7, 12, None]

- SVM:
 - 'C': [0.1, 1, 10]
 - 'gamma': [0.1, 1, 'scale', 'auto']
 - 'kernel': ['linear', 'rbf', 'sigmoid', 'poly']
- KNN:
 - 'n_neighbors': [3, 5, 6, 7, 8, 9, 10]
- Logistic Regresion:
 - 'C': [0.1, 1, 10]
 - 'max_iter': [500, 1000, 1500]
- Ridge Regresion:
 - 'alpha': [0.01, 0.1, 1, 10, 100]
- XGBoost:
 - 'n_estimators': [50, 100, 200]
 - 'learning_rate': [0.01, 0.1, 1]
 - 'max_depth': [3, 5, 7]
- CatBoost:
 - 'iterations': [50, 100, 200]
 - 'learning_rate': [0.01, 0.1, 1]

3.5.2. Validación cruzada

La validación cruzada es una técnica que se emplea para el entrenamiento y la evaluación de los modelos de aprendizaje automático. Su finalidad principal es evitar el *overfitting* de los modelos, es decir que no se ajusten excesivamente a los datos de entrenamiento y que sean generalizables [69-71,76].

Una de las formas más comunes y empleadas de validación cruzada es *k-fold*. Esta técnica consiste en dividir el conjunto de datos de entrenamiento en k divisiones (*folds*) de tamaño similar. De esta forma el modelo se entrenará k veces, con k-1 conjuntos de entrenamiento y realizará la validación del modelo con la partición restante, en cada una de las diferentes combinaciones. Una vez completado, se calcula la media de las métricas obtenidas en cada iteración, de forma que se obtiene una estimación más estable y confiable del rendimiento del modelo. Además, también permite aprovechar de forma más eficiente los datos de los que se dispone, y reduce la varianza asociada a una única partición de entrenamiento/test [69-71,76].

Para el desarrollo del trabajo se ha empleado la validación cruzada con k = 5, tanto para la evaluación del rendimiento como para la optimización de hiperparámetros mediante Grid Search.

3.5.3. Métricas de evaluación

En este proyecto se han empleado diferentes métricas de evaluación según la tarea de predicción que se ha llevado a cabo. Para las tareas de clasificación se han empleado la matriz de confusión, la precisión general del modelo, la precisión y la sensibilidad para la clase de interés (pacientes con Alzheimer) y el área bajo la curva ROC de la clase de interés. En la tarea de regresión se ha empleado el error absoluto medio (MAE), el error cuadrático medio (MSE), y su raíz cuadrada. Estas métricas se explican en detalle a continuación.

3.5.3.1. Matriz de confusión y sus métricas derivadas

La matriz de confusión es una forma de visualizar los errores y los aciertos de un modelo, en función de las diferentes clases de las que se dispone. En la matriz de confusión se representa la clase a la que pertenecen los registros frente a la clase que se ha predicho con el modelo (**Figura 8**). De forma que se pueden diferenciar diferentes parámetros: los verdaderos positivos (VP) son los registros que pertenecen a la clase positiva que se han clasificado como tal con el modelo, los verdaderos negativos (VN) son los registros de la clase negativa que se han clasificado como tal, los falsos negativos (FN) con los registros de la clase positiva que se han clasificado como negativos, y los falsos positivos (FP) que son los registros de la clase negativa que se han clasificado como positivos [69,70,76].

En este caso, en uno de los *datasets* se dispone de dos clases mientras que el segundo contiene 5 clases diferentes (**Figura 8**).

		Valores predichos				
		0	1	2	3	4
Valores reales	Positivo	Verdaderos Positivos	Falsos Negativos			
	Negativo	Falsos Positivos	Verdaderos Negativos			
		0	VN	VN	FP	VN
		1	VN	VN	FP	VN
		2	FN	FN	VP	FN
		3	VN	VN	FP	VN
		4	VN	VN	FP	VN

Figura 8. Matriz de confusión de un problema de clasificación binario y otro multiclas.

La **precisión global** del modelo corresponde al número de registros que han sido clasificados en la clase correcta con respecto al total de registros [69,70,76].

$$\text{precisión global} = \frac{\text{nº de registros clasificados correctamente}}{\text{total de registros}}$$

La **precisión** para la clase de interés se calcula teniendo en cuenta los verdaderos positivos para esa clase con respecto al total de registros que se han clasificado para esa clase, es decir, tiene en cuenta los falsos positivos [69,70,76].

$$\text{precisión} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos positivos}}$$

La **sensibilidad o recall** para la clase de interés se calcula teniendo en cuenta los verdaderos positivos para esa clase con respecto al total de registros de esa clase, en este caso tiene en cuenta los falsos negativos [69,70,76].

$$\text{sensibilidad o recall} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}$$

3.5.3.2. Curvas ROC

Las curvas ROC (*Receiver Operating Characteristic*) son una representación gráfica de la tasa de falsos positivos con respecto a los verdaderos positivos. En la representación de la curva ROC se representa la diagonal que se interpreta como un modelo generado aleatoriamente, valores inferiores se consideraría que el modelo no mejora una estimación aleatoria, mientras que valores superiores que el modelo consigue predecir las clases mejor que si se realizara de forma aleatoria. A partir de la curva ROC se calcula el área bajo la curva (AUC) que permite caracterizar el rendimiento del modelo, de forma que se podrían tener en cuenta los siguientes valores en función del rendimiento del modelo: AUC 0,5-0,6 corresponde a un modelo malo, AUC 0,6-0,75 modelo regular, AUC 0,75-0,9 modelo bueno, AUC 0,9-0,97 modelo muy bueno, AUC 0,97-1 modelo excelente [69,70,76].

3.5.3.3. Métricas de evaluación para los modelos de regresión

Para evaluar el rendimiento de los modelos de regresión empleados en este estudio, se han empleado diferentes métricas que permiten cuantificar la calidad de las predicciones en relación con los valores reales. Las métricas empleadas fueron:

- El **error absoluto medio (MAE)**: se calcula restando al valor real el valor predicho y realizando la media para todos los valores obtenidos por el modelo. La ventaja de esta métrica es que está en las unidades del valor predicho, y mide en promedio cuanto se equivoca el modelo [69,70,76].

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

n = número de registros

Y_i = valores observados

\hat{Y}_i = valores predichos

- El **error cuadrático medio (MSE)**: es una de las métricas más empleadas en la evaluación de este tipo de modelos. Representa la media de los errores al cuadrado, lo que permite penalizar en mayor medida los errores más grandes [69,70,76].

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

n = número de registros

Y_i = valores observados

\hat{Y}_i = valores predichos

- La **raíz cuadrada del error cuadrático medio (RMSE)**: se calcula como la raíz cuadrada del MSE y permite interpretar la métrica en las mismas unidades que el valor original [69,70,76].
- **RMSE relativo**: dado que las anteriores métricas van a depender de la escala en la que se encuentre la variable que se quiere predecir, se puede emplear esta métrica que se calcula como el cociente entre el RMSE y el rango de valores o la media de la variable objetivo. Se considera un RMSE relativo bajo y que por tanto el modelo tiene un buen rendimiento si es menor al 10%, un RMSE relativo

entre el 10% y el 30% se considera un rango medio y que por tanto el modelo podría tener un buen funcionamiento pero que podría mejorarse, y por el contrario y el RMSE es superior al 30% se considera que el modelo no está funcionando correctamente, y que probablemente se haya producido un sobreajuste [69,70,76].

- El **coeficiente de determinación (R^2)**: indica el porcentaje de varianza explicada por el modelo respecto a la varianza total. Por lo tanto, un valor de R^2 cercano a 1 indica que se ha realizado un buen ajuste del modelo, mientras que un valor cercano a 0 indica poca capacidad predictiva del modelo [69,70,76].

3.6. Interpretabilidad de los modelos

Con la finalidad de analizar el comportamiento de los modelos y la relevancia de las variables predictoras, se ha llevado a cabo el análisis de importancia de características de los modelos que mejores resultados han obtenido en cada uno de los casos. Los algoritmos basados en árboles de decisión proporcionan de forma nativa medidas de importancia para cada variable, a través del atributo '*feature_importances_*' de los modelos entrenados. De esta forma se obtiene una lista de las variables que se emplean en el entrenamiento con una puntuación de importancia para las decisiones del modelo [69].

Con este método se obtienen listas con la importancia de cada variable, en el caso de los modelos que se han entrenado con pocas variables (*dataset* factores de riesgo y filtro releifF y wrapper del *dataset* biomarcadores, y para la predicción del MMSE) se ha analizado la importancia de todas las variables que participan en el modelo, en el resto de casos (filtro p-valor del *dataset* biomarcadores) se ha analizado las 20 variables con mayor importancia.

Este análisis permite tanto obtener modelos más transparentes como analizar la coherencia de los resultados obtenidos con el conocimiento previo sobre la enfermedad y los factores implicados en el diagnóstico. Además, este estudio también puede permitir el descubrimiento de nuevas proteínas o rutas de señalización implicadas en la enfermedad, lo que puede tener impacto en el ámbito clínico y también impulsar nuevos estudios enfocados en estas rutas.

3.7. Softwares empleados

Para el desarrollo del código necesario para el presente trabajo se ha utilizado el lenguaje de Python (versión 3.11.5), debido a su amplia disponibilidad de bibliotecas orientadas al análisis de datos y *machine learning*. La implementación del código se ha realizado en el entorno de Jupyter Notebook (versión 6.5.4), que permite combinar código, visualizaciones y texto, facilitando así el desarrollo y reproducibilidad del trabajo. Además, en la **tabla 3** se muestran las diferentes bibliotecas que se han utilizado para el desarrollo del estudio y la versión que se ha empleado.

Con el objetivo de garantizar la transparencia, reproducibilidad y facilitar futuras investigaciones, el código desarrollado durante el presente trabajo ha sido organizado y depositado en un repositorio público de GitHub (https://github.com/mblanchruiz/TFM_Prediccion_Alzheimer.git). Las notebooks contienen el flujo completo del análisis, incluyendo la preparación de los datos, la selección de características, el entrenamiento y evaluación de los modelos, así como los análisis de interpretabilidad, y las visualizaciones realizadas.

Tabla 3. Módulos empleados para el desarrollo del estudio y su versión utilizada.

Módulo	Versión
IPython	8.15.0
ipykernel	6.25.0
jupyter_core	5.3.0
pandas	1.5.3
numpy	1.24.0
scipy	1.11.1
matplotlib	3.7.2
seaborn	0.12.2
scikit-learn	1.2.2
imblearn	0.8.0
xgboost	2.1.3
catboost	1.2.7

Capítulo 4

Resultados

4.1. Análisis exploratorio de los *datasets*

4.1.1. *Dataset* factores de riesgo

Tras el preprocesamiento de los datos, en cuanto a la limpieza de las variables, una vez aplicado el filtro de la varianza 0 y el coeficiente de correlación $> 0,8$ (**Figura 9**) no se han detectado variables que cumplan estos filtros y por lo tanto el número de variables que se van a incluir en el *dataset* final para la construcción de los modelos es 32. En cuanto a la limpieza de los registros, no se han detectado valores duplicados ni valores nulos por lo que para la construcción de los modelos se ha empleado la cantidad de registros del conjunto de datos original, 2149.

De las variables disponibles para el estudio, se incluyen tanto numéricas como categóricas y todas ellas se describen en el Anexo. Ninguna de las variables presenta una distribución normal, y en la **tabla 4** se muestra cómo se comportan las diferentes variables según el diagnóstico. Además, se ha observado que las variables ‘MMSE’, ‘FunctionalAssessment’, ‘MemoryComplaints’, ‘BehavioralProblems’ y ‘ADL’ son estadísticamente diferentes entre el grupo de individuos sanos y los pacientes con Alzheimer (**Figura 10**).

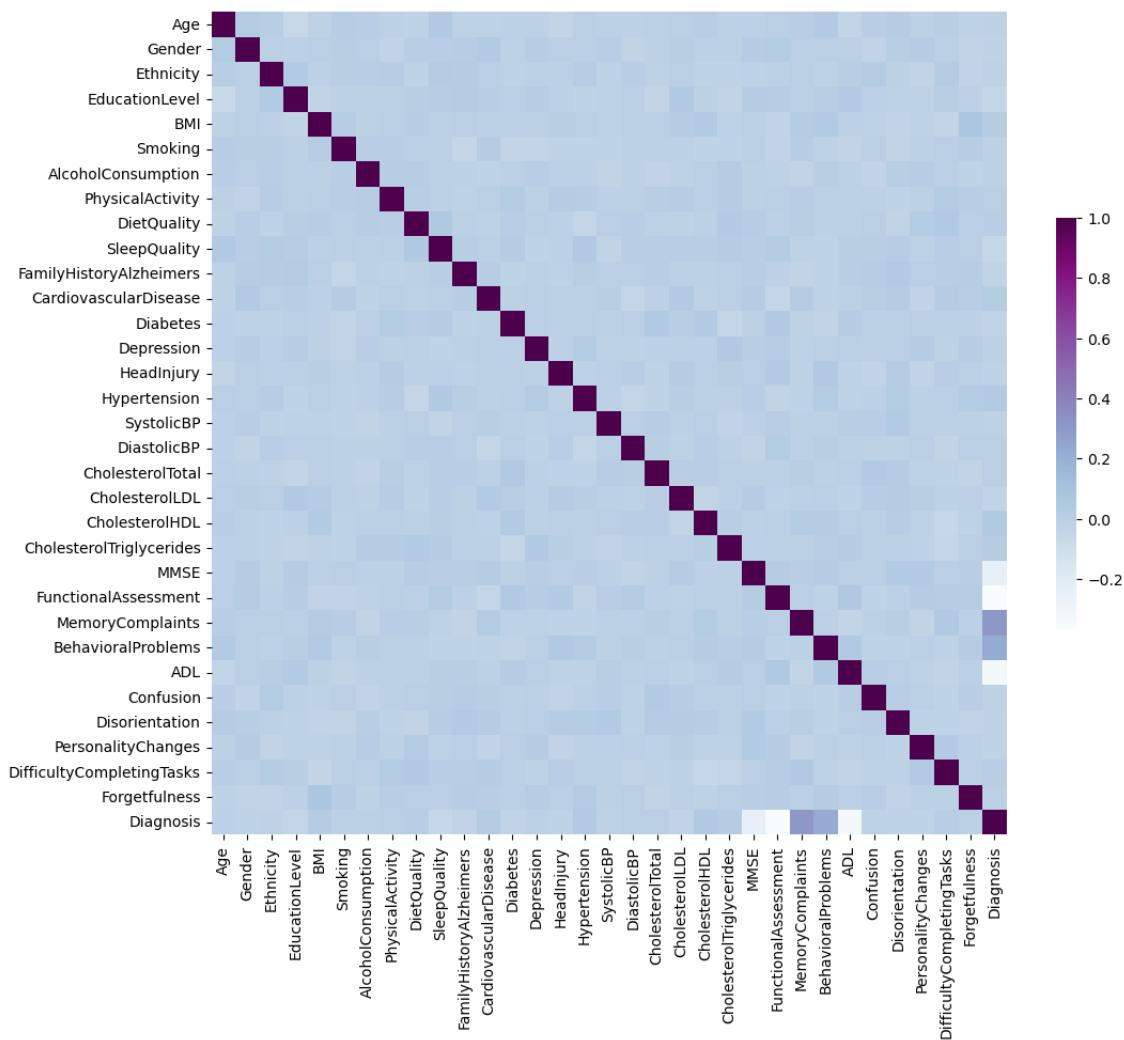


Figura 9. Matriz de correlaciones (Spearman) de las variables incluidas en el *dataset* factores de riesgo.

Tabla 4. Descripción de las variables del *dataset* factores de riesgo por grupo de diagnóstico.

Variable	General	Controles	Alzheimer	p-valor
Age (años)	74,9 ± 9,0	74,9 ± 8,9	74,8 ± 9,1	
Gender (Hombre/Mujer)	51/49	51/49	51/49	
Ethnicity (Caucásico/ Afroamericano/ Asiático/ Otro)	59/ 21/ 10/ 10	59/ 22/ 10/ 9	61/ 19/ 11/ 9	
EducationLevel (Graduado escolar /Grado Universitario/ Sin estudios reglados/ Educación Superior)	40/ 30/ 21/ 10	40/ 30/ 20/ 11	40/ 29/ 23/ 9	
BMI	27,7 ± 7,2	27,5 ± 7,2	27,9 ± 7,3	
Smoking (No/Sí)	71/29	71/29	71/29	
AlcoholConsumption	10,0 ± 5,8	10,1 ± 5,8	10,0 ± 5,8	
PhysicalActivity	4,9 ± 2,9	4,9 ± 2,9	4,9 ± 2,8	
DietQuality	5,0 ± 2,9	5,0 ± 2,9	5,0 ± 2,9	
SleepQuality	7,1 ± 1,8	7,1 ± 1,8	6,9 ± 1,8	
FamilyHistoryAlzheimers (No/Sí)	75/25	74/26	77/23	
CardiovascularDisease (No/Sí)	86/14	86/14	84/16	
Diabetes (No/Sí)	85/15	84/16	86/14	
Depression (No/Sí)	80/20	80/20	80/20	
HeadInjury (No/Sí)	91/9	90/10	92/8	
Hypertension (No/Sí)	85/15	86/14	83/17	
SystolicBP (mm Hg)	134,3 ± 25,9	134,6 ± 25,9	133,7 ± 26,0	
DiastolicBP (mm Hg)	89,8 ± 17,6	89,8 ± 17,7	90,0 ± 17,5	
CholesterolTotal (mg/dL)	225,2 ± 42,5	225,0 ± 42,2	225,6 ± 43,2	
CholesterolLDL (mg/dL)	124,3 ± 43,4	125,4 ± 43,4	122,5 ± 43,2	
CholesterolHDL (mg/dL)	59,5 ± 23,1	58,7 ± 23,1	60,8 ± 23,2	
CholesterolTriglycerides (mg/dL)	228,3 ± 102,0	226,6 ± 101,9	231,4 ± 102,1	
MMSE	14,8 ± 8,6	16,3 ± 8,9	12,0 ± 7,2	*
FunctionalAssessment	5,1 ± 2,9	5,9 ± 2,8	3,7 ± 2,6	*
MemoryComplaints (No/Sí)	79/21	88/12	62/38	*
BehavioralProblems (No/Sí)	84/16	90/10	73/27	*
ADL	5,0 ± 2,9	5,7 ± 2,8	3,7 ± 2,7	*
Confusion (No/Sí)	79/21	79/21	81/19	
Disorientation (No/Sí)	84/16	84/16	85/15	
PersonalityChanges (No/Sí)	85/15	84/16	86/14	
DifficultyCompletingTasks (No/Sí)	84/16	84/16	84/16	

Forgetfulness (No/Sí)	70/30	70/30	70/30	
Diagnosis (Control/Alzheimer)	65/35	-	-	

Los resultados se muestran como la media \pm sd, o como el porcentaje para cada clase, según el tipo de variable. El p-valor ha sido calculado según el tipo de variable con el test Mann-Whitney o con el test chi-cuadrado aplicando posteriormente la corrección de Bonferroni.

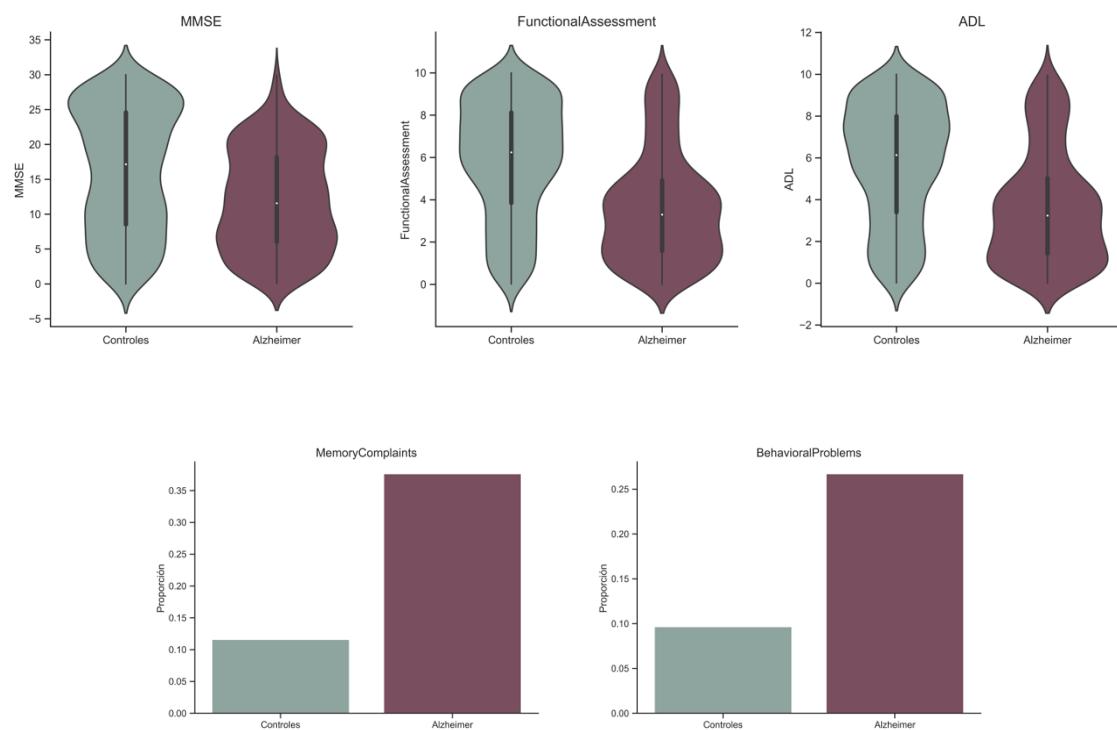


Figura 10. Distribución de las variables estadísticamente significativas en función de los grupos de diagnóstico del *dataset* factores de riesgo.

4.1.2. Dataset biomarcadores

Este *dataset* dispone de 9486 variables y 609 registros, incluyendo pacientes con Alzheimer, Parkinson, Cáncer de mama, controles de mayor edad y controles jóvenes. Las variables incluidas son los identificadores, la edad, el sexo, el MMSE, el diagnóstico y los niveles en suero de 9480 proteínas. Las características demográficas y la distribución de los pacientes por grupo de diagnóstico se muestran en la **tabla 5**.

Tabla 5. Características demográficas de los sujetos incluidos en el *dataset* biomarcadores.

Variable	General	Controles mayores	Controles jóvenes	Alzheimer	Parkinson	Cáncer de mama	p-valor
Age (años)	64,9 ± 20,5	57,8 ± 7,6	24,7 ± 3,7	78,5 ± 8,7	74,3 ± 9,0	46,9 ± 5,8	*
Gender (Hombre/Mujer)	56/44	100/0	70/30	42/58	58/42	0/100	*
MMSE	16,5 ± 5,4	-	-	16,5 ± 5,4	-	-	
Diagnosis		57	20	13	5	5	

Los resultados se muestran como la media ± sd, o como el porcentaje para cada clase, según el tipo de variable. El p-valor ha sido calculado según el tipo de variable con el test Kruskal-Wallis (para comparaciones múltiples) o con el test chi-cuadrado.

Para la construcción de los modelos se van a emplear únicamente las variables que hacen referencia a los niveles de las diferentes proteínas, y no se van a incluir las variables de edad, género y MMSE, ya que además esta última variable únicamente dispone de valores para el grupo de diagnóstico con Alzheimer, por lo que tampoco se podía realizar la imputación de los valores faltantes. De esta forma los modelos pueden centrarse únicamente en la información que aportan las proteínas en el suero de los pacientes para tratar de buscar algún patrón útil que permita realizar un diagnóstico más preciso y eficaz que con los métodos actuales.

Puesto que se dispone de información de 9480 proteínas para optimizar la construcción de los modelos se va a realizar una selección de variables, con la finalidad de reducir la dimensionalidad y obtener las variables más relevantes. En primer lugar, se eliminan variables que no aportan información, es decir, que tienen varianza 0 o próxima a 0 (< 0,01). Este filtro no elimina ninguna variable. El segundo paso es eliminar las variables con información redundante, y para ello se calcula el coeficiente de correlación de Spearman y se elimina una variable de cada par que presenten una correlación alta (|> 0,8|) (**Figura 11**). En este paso se eliminan 1504 variables, de forma que quedarían 7976. No obstante, continúa siendo un número muy alto de variables para la construcción de los modelos, por ello se ha realizado una selección de variables mediante dos métodos diferentes. Uno de esos métodos es mediante el filtro del p-valor, para ello en primer lugar, se ha analizado la distribución de las variables y se ha comprobado que en su mayoría no siguen una distribución normal, en consecuencia, para realizar el test estadístico y obtener cuáles son las proteínas con niveles significativamente diferentes entre el grupo de Alzheimer y el resto de grupos de diagnóstico, se ha empleado el test no paramétrico de Mann-Whitney con la posterior corrección de Bonferroni. Una vez realizado el test se obtienen 220 variables con un p-valor menor a 0,05. Sin embargo, como en algunos casos

se pueden obtener diferencias significativas, aunque los cambios entre grupos sean tenues, se ha aplicado otro filtro teniendo en cuenta el grado de diferencias entre grupos, \log_2 fold-change $> |0,5|$. El resultado de aplicar estos dos filtros obtiene 98 variables (**Figura 12**), cuya lista se recoge en el Anexo.

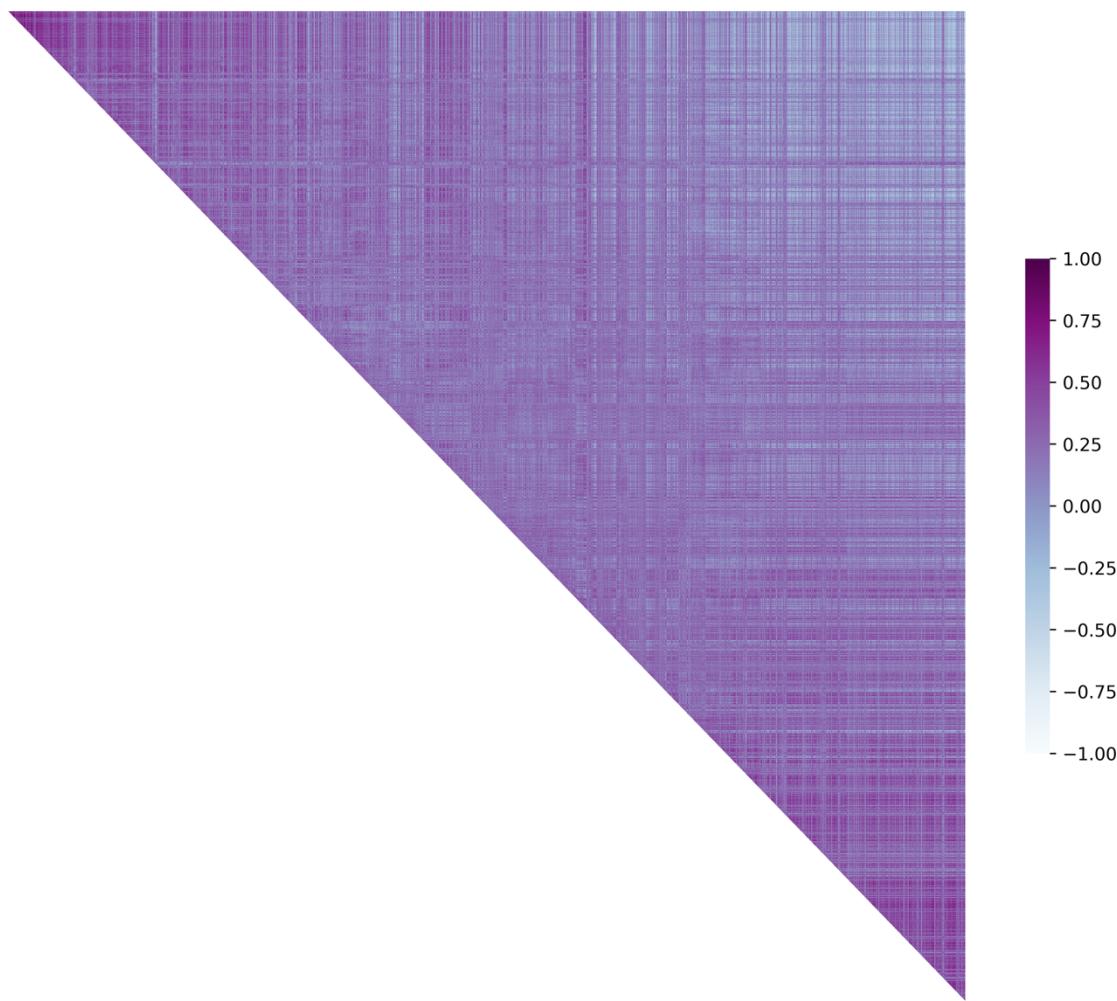


Figura 11. Matriz de correlaciones (Spearman) de las proteínas incluidas en el *dataset* biomarcadores.

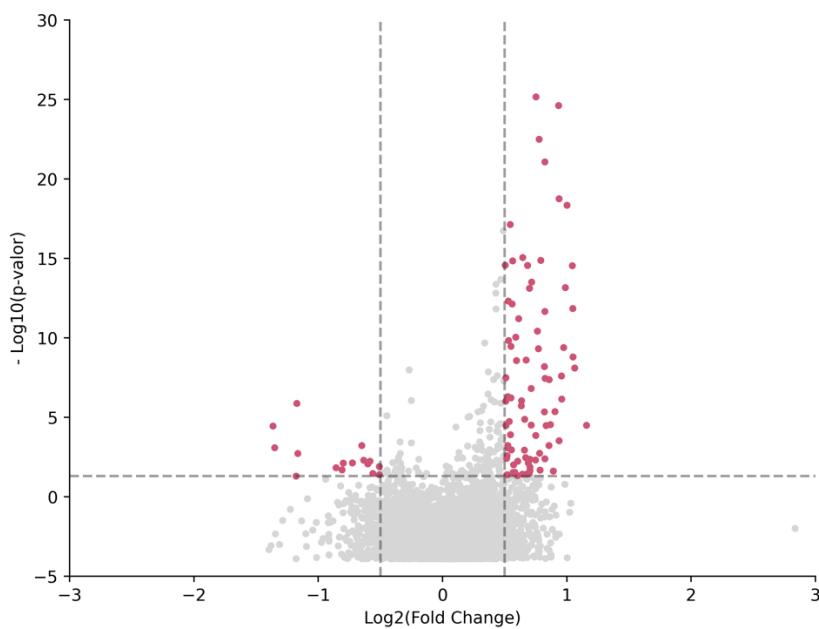


Figura 12. Volcano plot de las proteínas analizadas del dataset biomarcadores.

El otro método que se ha empleado para la selección de variables es el uso de la técnica reliefF y posteriormente la selección por wrapper. Con la técnica reliefF se asigna la puntuación de importancia a cada variable, y se seleccionan las 500 con mayor puntuación como entrada para la selección por wrapper. Por último, tras aplicar la eliminación recursiva de las características mediante en modelo RandomForest se determina que el número óptimo de variables para la tarea de clasificación de los pacientes por el diagnóstico es 23 (la lista con estas variables se recoge en el Anexo).

Por otro lado, con este *dataset* también se ha realizado una tarea de regresión para predecir el progreso del Alzheimer basándose en la predicción de los valores del MMSE de los pacientes. Para ello, se han seleccionado únicamente los pacientes con Alzheimer, que son los únicos de los que se dispone del valor de MMSE (350 pacientes). En la **figura 13** se puede observar la distribución de esta variable. Para la construcción de los modelos de regresión, de nuevo, se ha realizado la selección de variables. En este caso únicamente se ha empleado el método basado en el uso de reliefF y wrapper. Puesto que se ha cambiado la variable objetivo se debe hacer los cálculos de importancia de las variables y aplicar el modelo de eliminación recursiva de las características de nuevo. Del mismo modo que para la tarea anterior de clasificación, en primer lugar, se seleccionan las 500 variables con una mayor puntuación de importancia según la técnica reliefF, y posteriormente se aplica la eliminación recursiva sobre esas variables empleando el modelo Random Forest. En este caso se determina que el número óptimo de variables es 29 (la lista con estas variables se recoge en el Anexo).

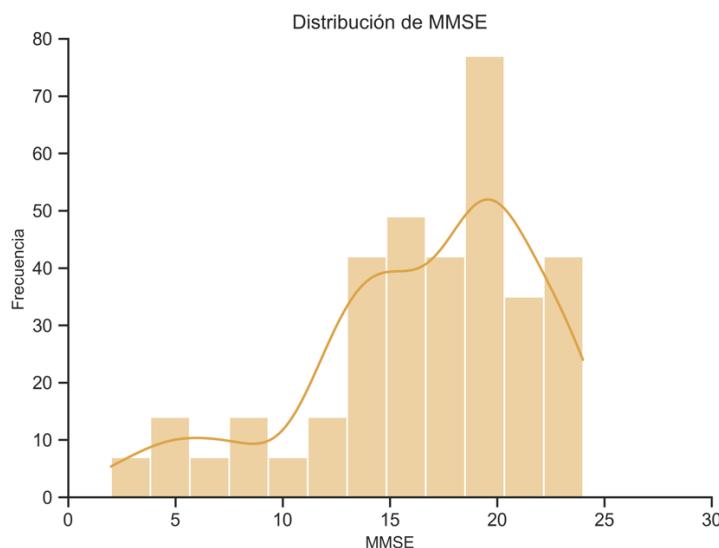


Figura 13. Distribución de la variable MMSE del *dataset* biomarcadores para los pacientes con Alzheimer.

4.2. Resultados obtenidos para el *dataset* factores de riesgo

Una vez construidos y optimizados los diferentes modelos de clasificación se han obtenido los modelos más eficientes para cada uno de los algoritmos empleados (**Tabla 6**). Como se puede observar en la tabla de resultados del *dataset* factores de riesgo, el modelo que ha conseguido mejores resultados en cuanto a precisión global del modelo, *recall* o sensibilidad para la clase de Alzheimer con respecto a los individuos sanos, y el AUC es CatBoost (**Figura 14**). Sin embargo, Random Forest consiguió una precisión ligeramente superior para la clase de Alzheimer (**Figura 14**). A pesar de ello, dado que en este contexto clínico es prioritario minimizar el número de falsos negativos (pacientes de Alzheimer no detectados), se ha dado mayor peso al valor de *recall*.

Por lo tanto, el modelo seleccionado para este *dataset* es CatBoost optimizado con los parámetros ‘interactions’: 50 y ‘learning_rate’: 0,01. Este resultado podría ser esperable, ya que este es el algoritmo que trabaja mejor con variables categóricas, que son predominantes en este conjunto de datos, y también debido a que se trata de un método de boosting por gradiente que además incluye otras implementaciones, y que, por tanto, mejora a otros algoritmos también analizados como Decision Trees, Random Forest y XGBoost.

Tabla 6. Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el *dataset* factores de riesgo.

Modelo	Precisión global	Precisión AD	Recall AD	AUC ROC AD
Decision Tree	93,123	90,052	90,526	92,533
Random Forest	94,424	94,944	88,947	93,181
SVM	82,900	76,923	73,684	80,808
KNN Model	76,952	76,190	50,526	70,953
Logistic Regresion	83,086	77,654	73,158	80,832
XGBoost	94,796	94,022	91,053	93,946
CatBoost	95,539	94,624	92,632	94,879

AD: Alzheimer; AUC: Área bajo la curva; ROC: *Receiver Operating Characteristic*.

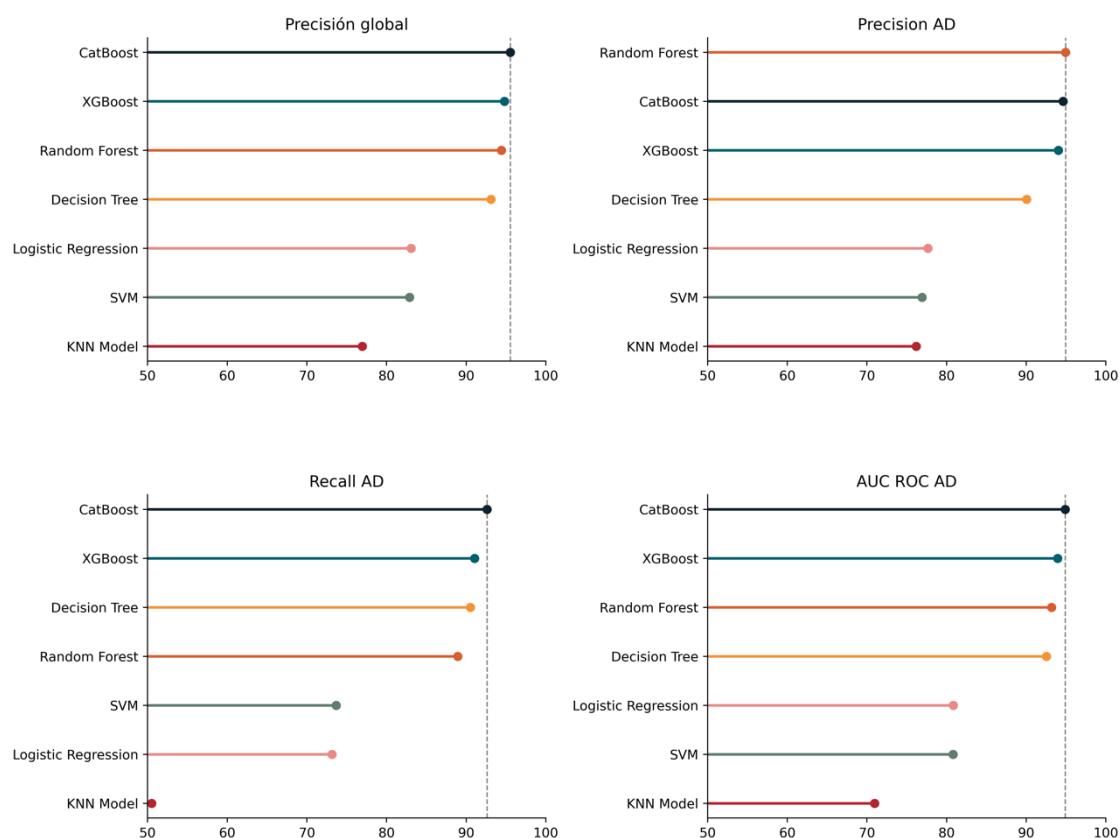


Figura 14. Representación del ranking de los modelos para el *dataset* factores de riesgo para las diferentes métricas.

4.3. Resultados obtenidos para el *dataset* biomarcadores

4.3.1. Predicción del Alzheimer

Con el *dataset* biomarcadores, mediante el que se dispone de información sobre los niveles en suero de diferentes proteínas, se ha realizado una tarea de clasificación con la finalidad de poder identificar a los pacientes con Alzheimer, con respecto al resto de grupos incluidos en el *dataset* como son controles de mayor edad y más jóvenes, pacientes con Parkinson y pacientes con cáncer de mama. Puesto que se disponía de un número muy alto de proteínas para la construcción de los modelos, se han empleado dos métodos de selección de características con la finalidad de entrenar dos grupos de modelos con particiones de datos diferentes. Por un lado, se ha analizado para cada método de selección de características cuales son los modelos que obtienen mejores resultados, y posteriormente, se busca analizar que método de selección de características ha conseguido que los modelos entrenados obtengan mejores resultados.

4.3.1.1. Resultados obtenidos con el filtro *p*-*valor*

Como se ha mencionado anteriormente con este filtro se seleccionan 98 variables para llevar a cabo el entrenamiento y optimización de los modelos, y como se puede observar existe cierta correlación entre las variables seleccionadas (**Figura 15**). Una vez se han seleccionado los parámetros óptimos para cada uno de los modelos se han obtenido sus resultados (**Tabla 7**). Como se puede ver en la tabla de resultados, el modelo que consigue una mayor precisión para la clase con Alzheimer es XGBoost ('max_depth': 3, 'n_estimators': 200, 'learning_rate': 0,1), y el que consigue un valor mayor de *recall* también para la clase de Alzheimer es SVM ('C': 10, 'gamma': 0,1, 'kernel': 'rbf'). No obstante, teniendo en cuenta los valores del AUC, que sería una métrica que tiene en cuenta tanto la precisión como la sensibilidad del modelo, se observa que los mejores resultados los obtienen los modelos Random Forest ('max_depth': None, 'n_estimators': 100) y CatBoost ('interactions': 200, 'learning_rate': 1) (**Figura 16**). En cuanto a la comparación de estos 4 modelos, se puede descartar XGBoost ya que, aunque presenta una mayor precisión, como se ha comentado antes, en este caso se busca priorizar valores altos de *recall*. Por otro lado, aunque el modelo que obtiene el mayor valor para *recall* es SVM, el resto de métricas no presentan valores tan buenos como si lo hacen los modelos Random Forest y CatBoost, sobre todo teniendo en cuenta el AUC. Por último, en cuanto a la comparación de los modelos Random Forest y CatBoost, estos presentan los mismos valores, considerándose valores altos y que dan lugar a buenos modelos para las métricas de precisión, *recall* y AUC de la clase con Alzheimer, y únicamente presentan diferencias en la precisión global del modelo haciendo que Random Forest sea ligeramente superior.

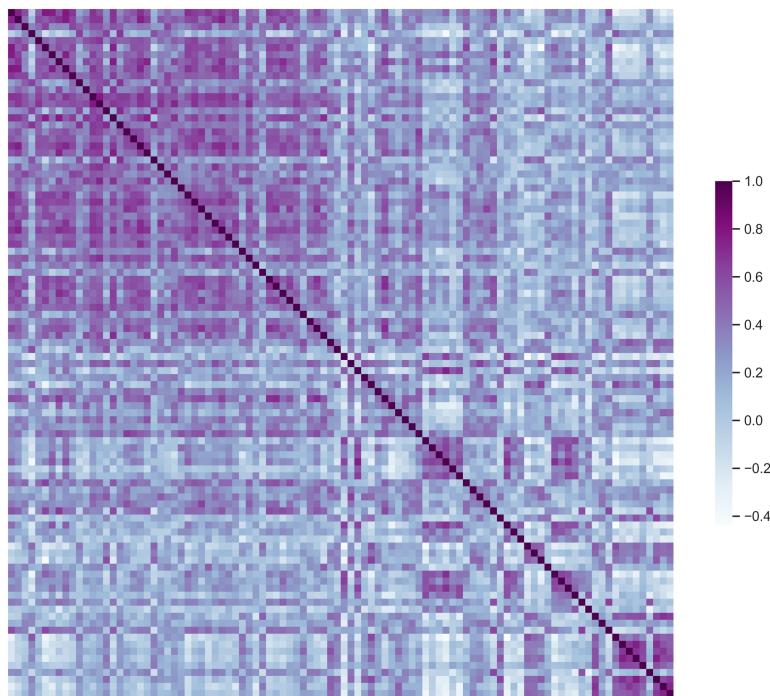


Figura 15. Correlación entre las 98 variables del *dataset* biomarcadores seleccionadas con el filtro p-valor.

Tabla 7. Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el *dataset* biomarcadores tras aplicar la selección de características por el filtro del p-valor.

Modelo	Precisión global	Precisión AD	Recall AD	AUC ROC AD
<i>Decision Tree</i>	76,471	88,608	79,545	82,850
<i>Random Forest</i>	89,542	93,182	93,182	91,976
SVM	75,817	74,138	97,727	75,787
KNN	67,320	94,340	56,818	76,101
<i>Logistic Regresion</i>	75,163	84,444	86,364	82,413
XGBoost	86,275	95,181	89,773	91,809
CatBoost	87,582	93,182	93,182	91,976

AD: Alzheimer; AUC: Área bajo la curva; ROC: *Receiver Operating Characteristic*.

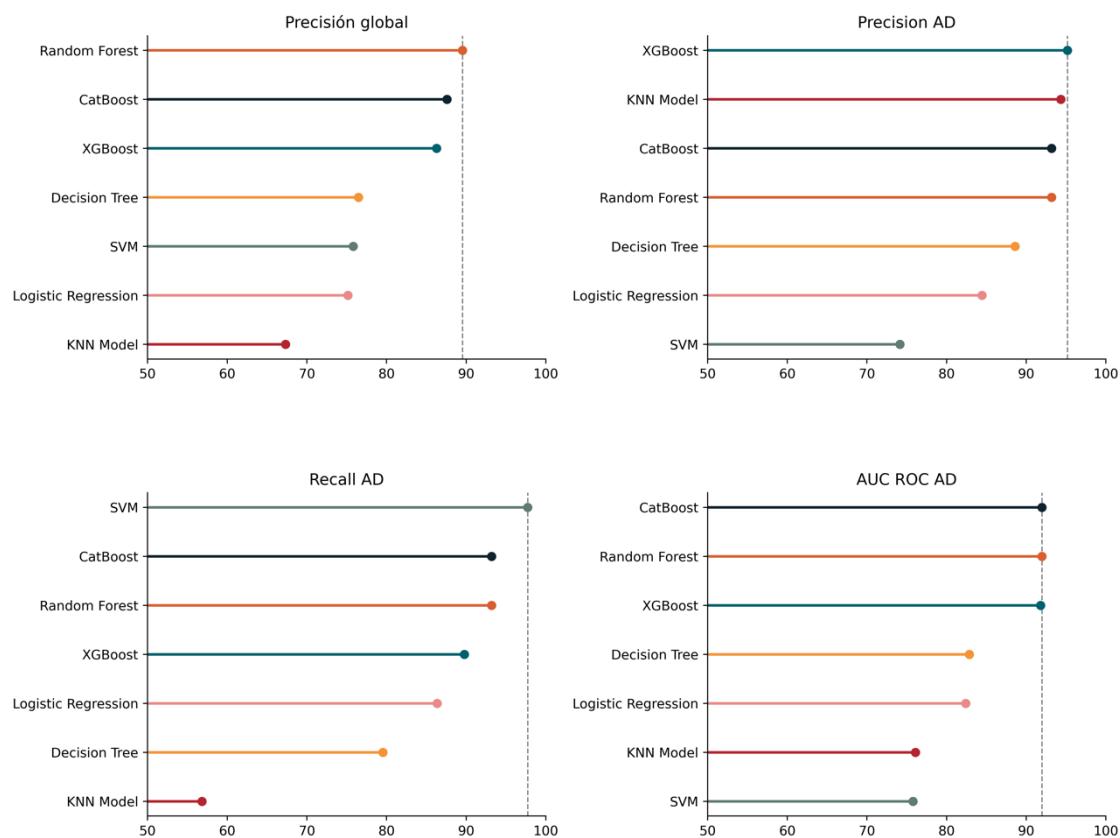


Figura 16. Representación del ranking de los modelos para el *dataset* biomarcadores con el filtro p-valor para las diferentes métricas.

4.3.1.2. Resultados obtenidos con *ReliefF* y *wrapper*

Tras aplicar el filtro *reliefF* y *wrapper* se han seleccionado 23 variables para el entrenamiento de los modelos. Como se puede observar algunas de estas variables presentan cierta correlación entre ellas (**Figura 17**). Con este filtro, el modelo que ha obtenido un valor mayor para la precisión de la clase con Alzheimer es KNN ('n_neighbours': 3), mientras que el que mejor *recall* y AUC consigue es XGBoost ('max_depth': 5, 'n_estimators': 200, 'learning_rate': 0,1). Por lo tanto, en este caso se considera que el mejor modelo es XGBoost (**Tabla 8** y **Figura 18**).

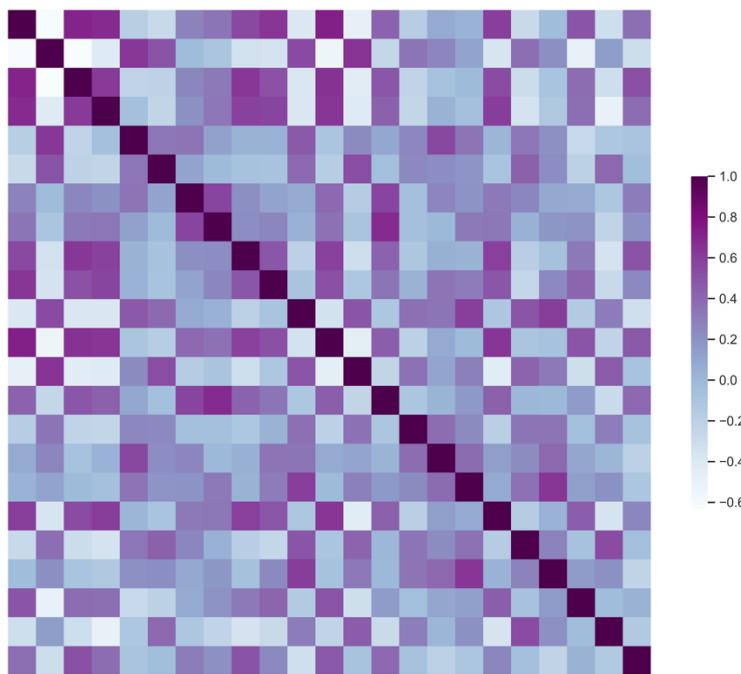


Figura 17. Correlación entre las 23 variables del *dataset* biomarcadores seleccionadas con el filtro reliefF y wrapper.

Tabla 8. Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de clasificación con el *dataset* biomarcadores tras aplicar la selección de características por el filtro reliefF y wrapper.

Modelo	Precisión global	Precisión AD	Recall AD	AUC ROC AD
<i>Decision Tree</i>	73,856	84,524	80,682	80,341
<i>Random Forest</i>	86,275	94,318	94,318	93,313
SVM	78,431	91,358	84,091	86,661
KNN	75,163	97,059	75,000	85,962
<i>Logistic Regression</i>	71,895	88,732	71,591	79,642
XGBoost	91,503	96,552	95,455	95,420
CatBoost	86,275	93,023	90,909	90,839

AD: Alzheimer; AUC: Área bajo la curva; ROC: *Receiver Operating Characteristic*.

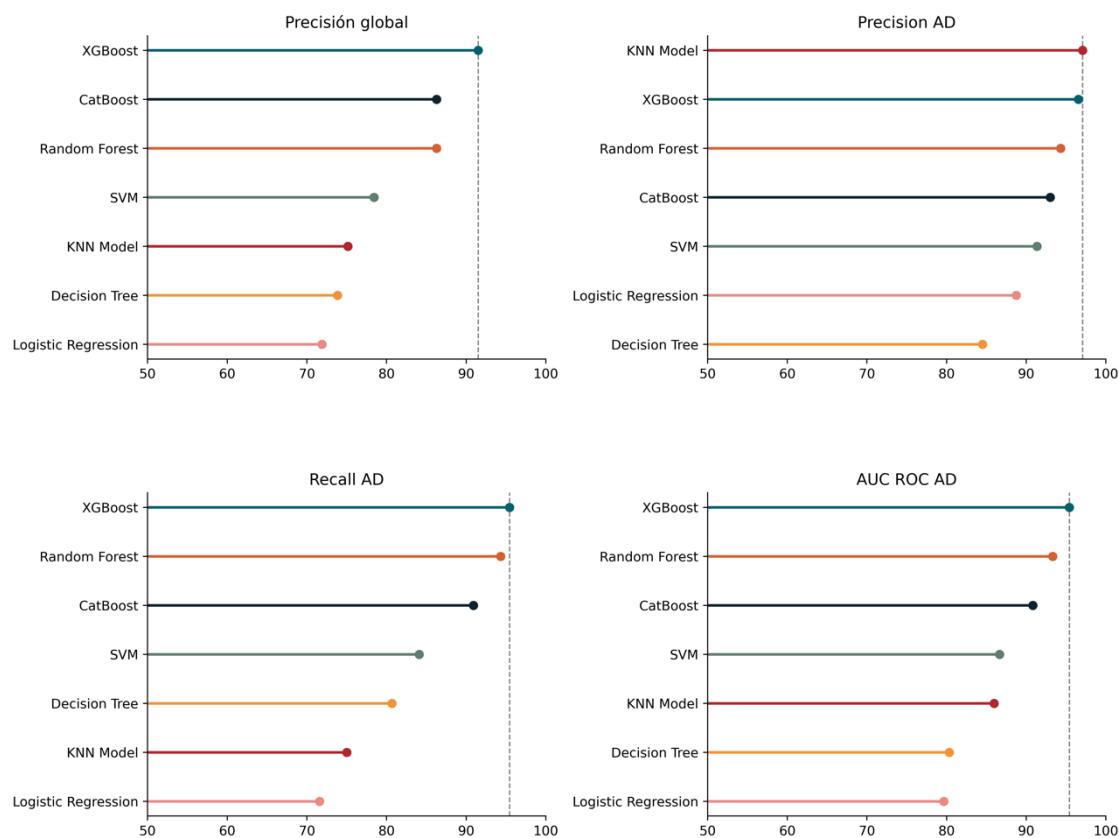


Figura 18. Representación del ranking de los modelos para el *dataset* biomarcadores con el filtro reliefF y wrapper para las diferentes métricas.

4.3.2. Predicción de la evolución del Alzheimer (MMSE)

El último bloque de modelos que se han desarrollado está orientado a la predicción del avance del Alzheimer, prediciendo mediante modelos de regresión el valor del MMSE, que representa una escala por la que se valora el deterioro cognitivo de los pacientes. La selección de variables para esta tarea ha obtenido 29 variables, y como se puede observar alguna de ellas presentan cierta correlación (**Figura 19**).

Para esta tarea de predicción el modelo que obtiene mejores resultados, evaluando las diferencias entre los valores reales y los valores predichos por los modelos, es CatBoost (**Tabla 9** y **Figura 20**). Este modelo es el que obtiene valores inferiores para MAE, MSE y RMSE, y un R^2 mayor lo que hace que este modelo explique una mayor parte de la varianza de los datos. Cabe destacar que en este tipo de problemas de predicción es importante tener en cuenta la escala de la variable que se quiere predecir para saber si se han obtenido buenos modelos. En este caso el RMSE relativo es de 15%, por lo que no

se puede considerar que es un modelo de un rendimiento muy bueno ($<10\%$), pero sí que sería un modelo medio.

Analizando en detalle las predicciones realizadas por el modelo CatBoost, si se compara los valores reales con respecto a los valores predichos se puede observar que para los valores menores del MMSE el modelo comete más errores (**Figura 21**). En la gráfica de residuos no se observa un patrón claro, lo que puede sugerir la ausencia de sesgos sistemáticos en la predicción (**Figura 21**). Sin embargo, en el *Q-Q plot* se puede observar que los residuos no siguen una distribución normal, si bien los valores centrales se alinean adecuadamente a la distribución teórica, los extremos, especialmente el inferior, se desvían de esta distribución (**Figura 21**). Esta observación se ha confirmado mediante el test de normalidad Shapiro-Wilk, que ha obtenido un p-valor $< 0,05$, indicando una desviación significativa respecto a la distribución normal. Estos resultados apuntan a que el modelo aprende mejor en las zonas con mayor densidad de datos (valores de MMSE entre 15 y 20), y presenta un mayor margen de error en los extremos (cerca de 0 o 30), generando residuos más grandes y asimétricos. Esta falta de homogeneidad en la distribución de los datos de entrenamiento (**Figura 13**) puede estar limitando la capacidad del modelo para generalizar correctamente en esas regiones, impactando negativamente en la distribución de los residuos.

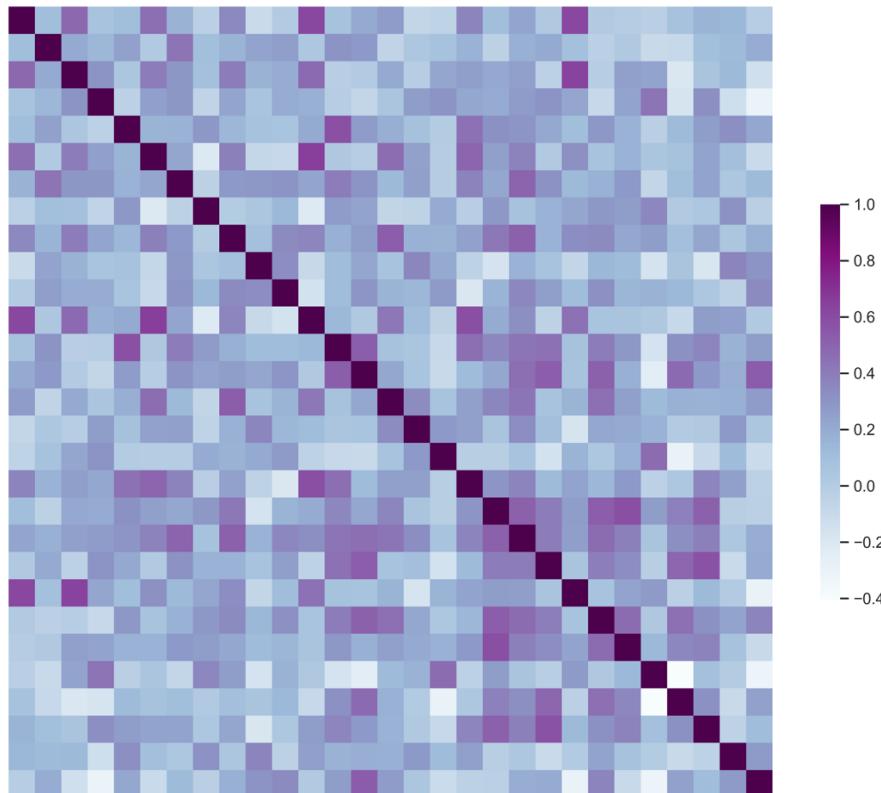


Figura 19. Correlación entre las 29 variables del *dataset* biomarcadores seleccionadas para la predicción del MMSE.

Tabla 9. Resultados obtenidos tras la optimización de los diferentes modelos para la tarea de regresión y predicción del MMSE en pacientes con Alzheimer con el *dataset* biomarcadores.

Modelo	MSE	RMSE	MAE	R ²	RMSE relativo (%)
Decision Tree	20,273	4,503	3,500	0,278	0,205
Random Forest	11,727	3,425	2,409	0,582	0,156
SVR	12,011	3,466	2,375	0,572	0,158
KNN	13,284	3,645	2,625	0,527	0,166
Ridge Regression	16,307	4,038	3,102	0,419	0,184
XGBoost	13,443	3,666	2,648	0,521	0,167
CatBoost	11,216	3,349	2,284	0,600	0,152

MSE: Error cuadrático medio; RMSE: raíz cuadrada del error cuadrático medio; MAE: Error absoluto medio.

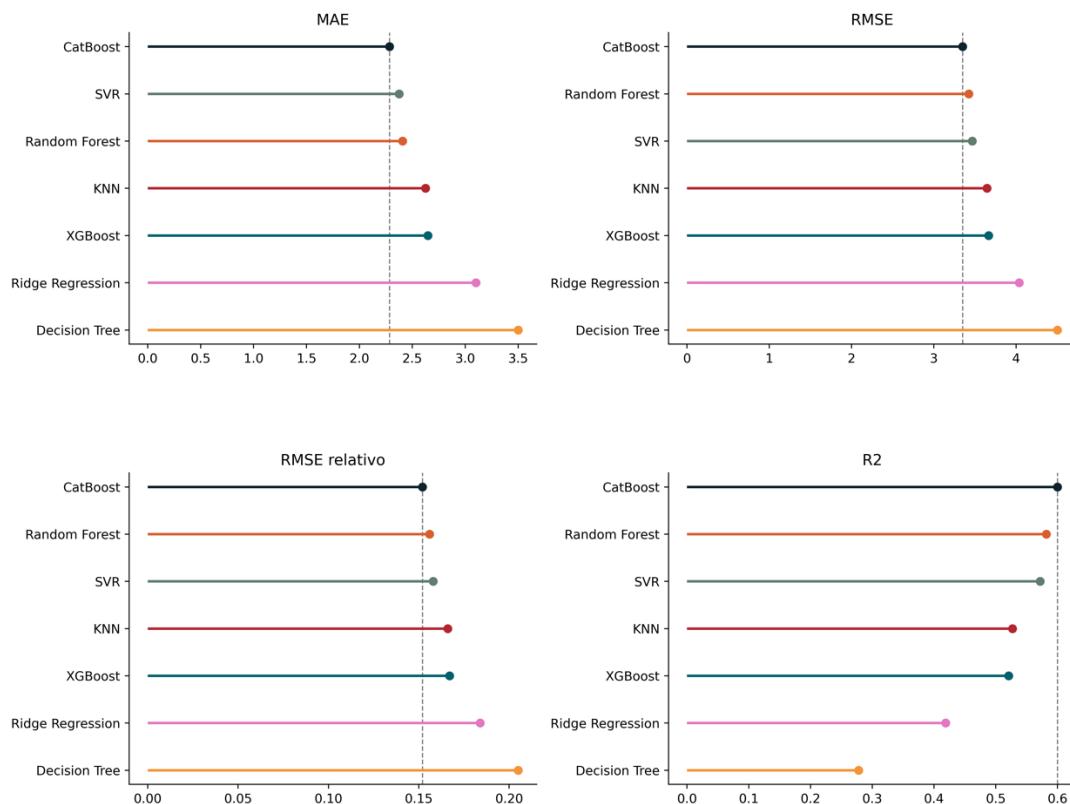


Figura 20. Representación del ranking de los modelos para el *dataset* biomarcadores para la predicción del MMSE para las diferentes métricas.

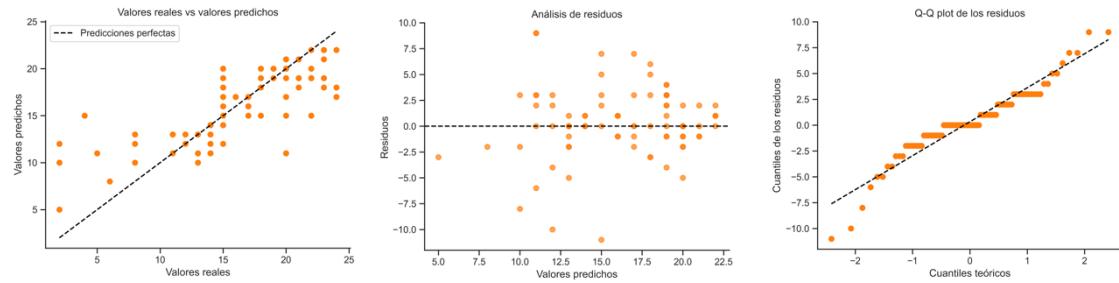


Figura 21. Análisis de los residuos de los valores predichos de MMSE con el modelo CatBoost con respecto a los valores reales.

4.4. Comparación de los modelos de clasificación

Una vez seleccionados los modelos que mejor funcionan para cada conjunto de datos, se puede realizar otra comparación para determinar con qué *dataset* y con qué método de selección de características se consiguen mejores resultados para poder identificar a los pacientes con Alzheimer. El modelo que obtuvo el mayor valor de *recall* fue SVM para el conjunto de datos del *dataset* biomarcadores con la selección de características por el p-valor (**Figura 18**). Sin embargo, este modelo no fue el seleccionado para ese conjunto de datos ya que el resto de métricas no eran favorables y se consideró mejor modelo el Random Forest. Por lo tanto, teniendo en cuenta el resto de modelos los mejores valores de *recall* fueron de los modelos Random Forest y XGBoost para el conjunto de datos del *dataset* de biomarcadores con la selección de características por *reliefF* y *wrapper*. Lo mismo se observa para la métrica AUC (**Figura 18**), aunque en este caso las diferencias son menores con los resultados obtenidos con el *dataset* factores de riesgo, sobre todo con el modelo CatBoost.

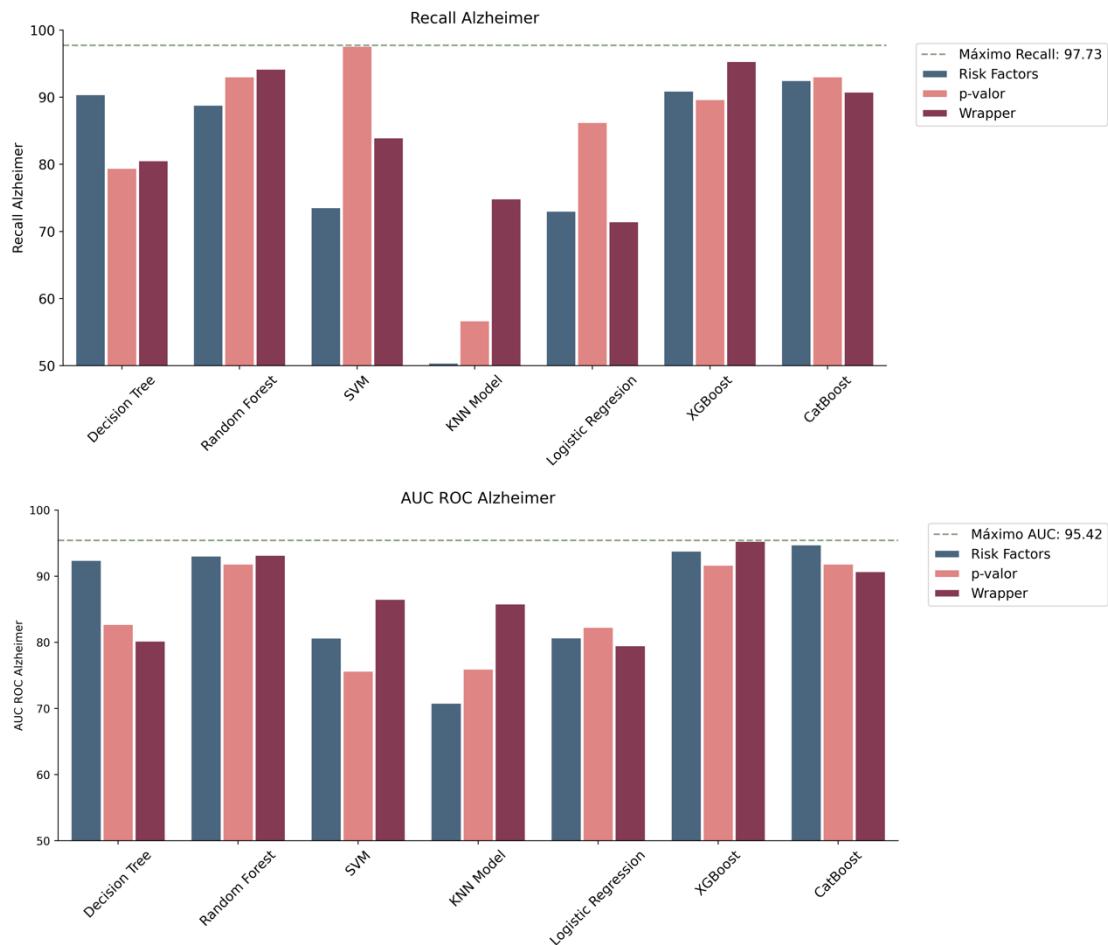


Figura 22. Comparación del *recall* y AUC de los modelos para los diferentes *datasets* y los dos filtros de selección de variables aplicados.

4.5. Interpretabilidad de los modelos

El último paso en el análisis de los resultados es determinar cuáles han sido las variables que participan en mayor medida en la toma de decisiones de los modelos.

4.5.1. *Dataset* factores de riesgo

Para el *dataset* de factores de riesgo el modelo seleccionado fue CatBoost, y puesto que en este caso únicamente se dispone de 32 variables se ha analizado la importancia de cada una de ellas en la toma de decisiones de este modelo (**Figura 19**). Las variables más relevantes para el modelo fueron la evaluación funcional, la puntuación de la calidad de las actividades de la vida cotidiana, los problemas de memoria, el MMSE, y los

problemas conductuales. Estas variables coinciden con las que en el análisis exploratorio de los datos se había observado que presentaban diferencias significativas entre el grupo de pacientes con Alzheimer y sujetos sanos. Seguidas de estas, pero ya con menos relevancia en el modelo se puede encontrar la tensión sistólica o los niveles de colesterol.

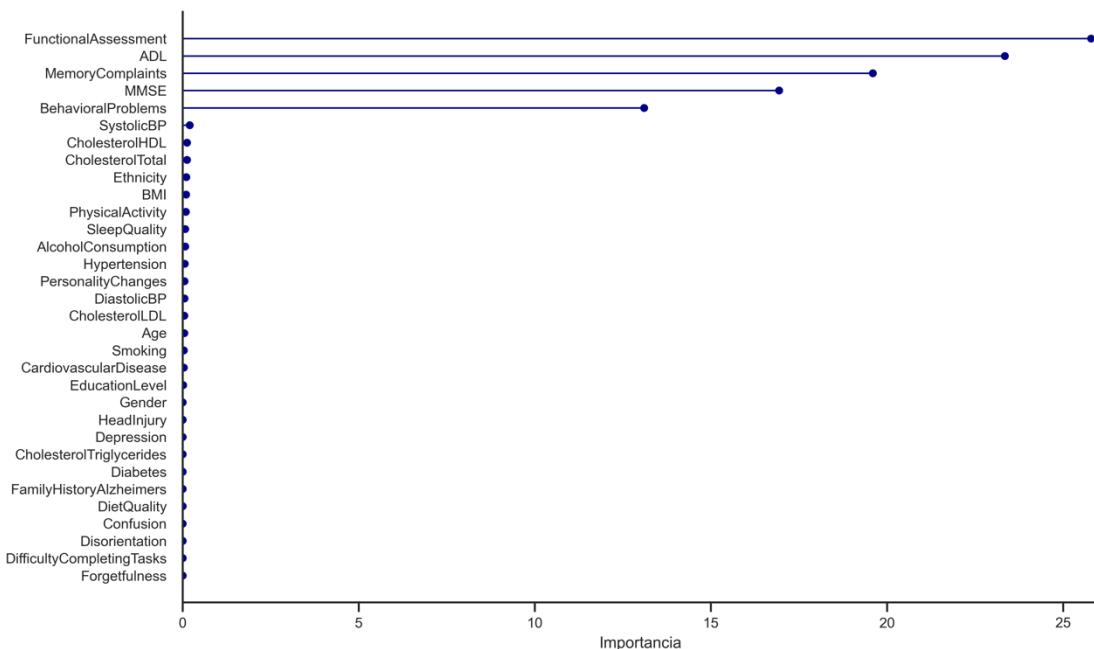


Figura 23. Ranking de la relevancia en el modelo CatBoost de las variables del *dataset* factores de riesgo para la predicción del Alzheimer.

4.5.2. *Dataset* biomarcadores

4.5.2.1. Filtro *p*-valor

El modelo que mejores resultados ha obtenido para el *dataset* biomarcadores con la selección de variables realizada con el filtro del p-valor es Random Forest. En la **Figura 24**, se muestran por orden de importancia las 20 proteínas más relevantes para este modelo. Las proteínas más importantes para la clasificación de los pacientes son PPP1R8 (subunidad 8 del inhibidor de la fosfatasa 1), ADARB1 (adenosina quinasa B1), NUDT16L1 (proteína de unión a nucleótidos difosfato), SKIP (inositol fosfatasa del músculo esquelético y riñón), RPL11 (proteína L11 ribosomal) y RALGPS2 (proteína Ral GEF con dominio PH y el motivo de unión 2 a SH3) (**Figura 24**).

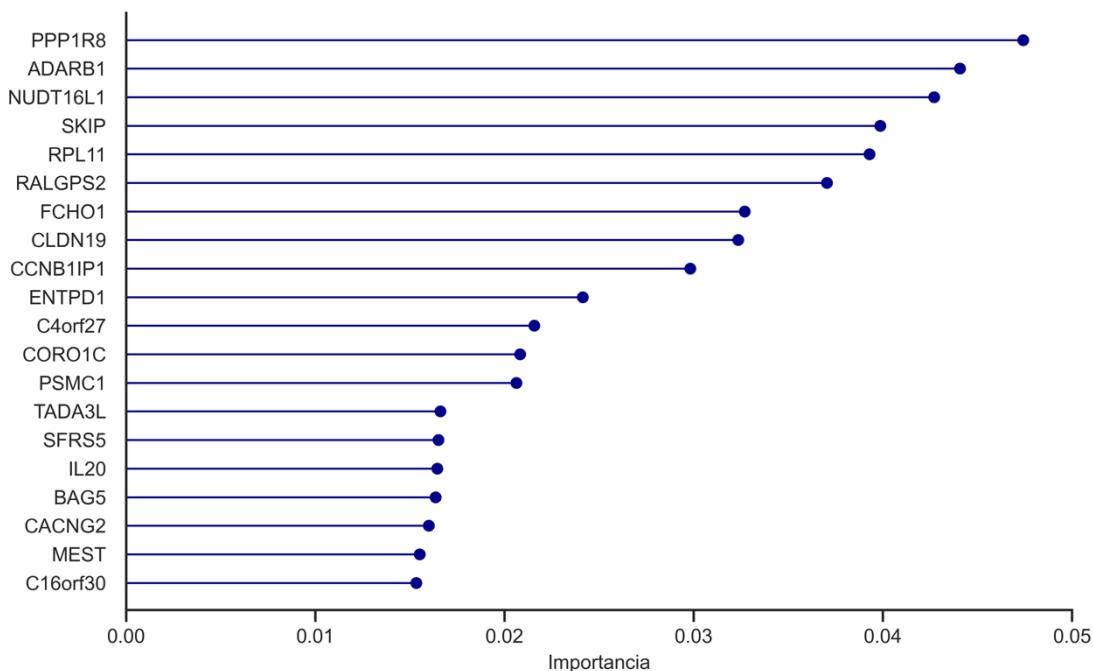


Figura 24. Ranking de la relevancia de las variables del *dataset* biomarcadores seleccionadas mediante el filtro p-valor para la predicción del Alzheimer con el modelo Random Forest.

4.5.2.2. Filtro ReliefF y wrapper

El modelo que mejores resultados ha obtenido para el *dataset* biomarcadores con la selección de variables realizada con el filtro reliefF y wrapper es XGBoost. En la **Figura 25**, se muestran por orden de importancia las 23 proteínas que se han empleado para el desarrollo del modelo. Las proteínas más relevantes son RAB10 (oncogen miembro de la familia RAS), ABAT (aminotransferasa 4-aminobutirato), FLJ32658, ENTPD1, EFCBP1 (proteína 1 de unión a la mano EF de calcio), y GTF2I (factor de transcripción general II-I) (**Figura 25**).

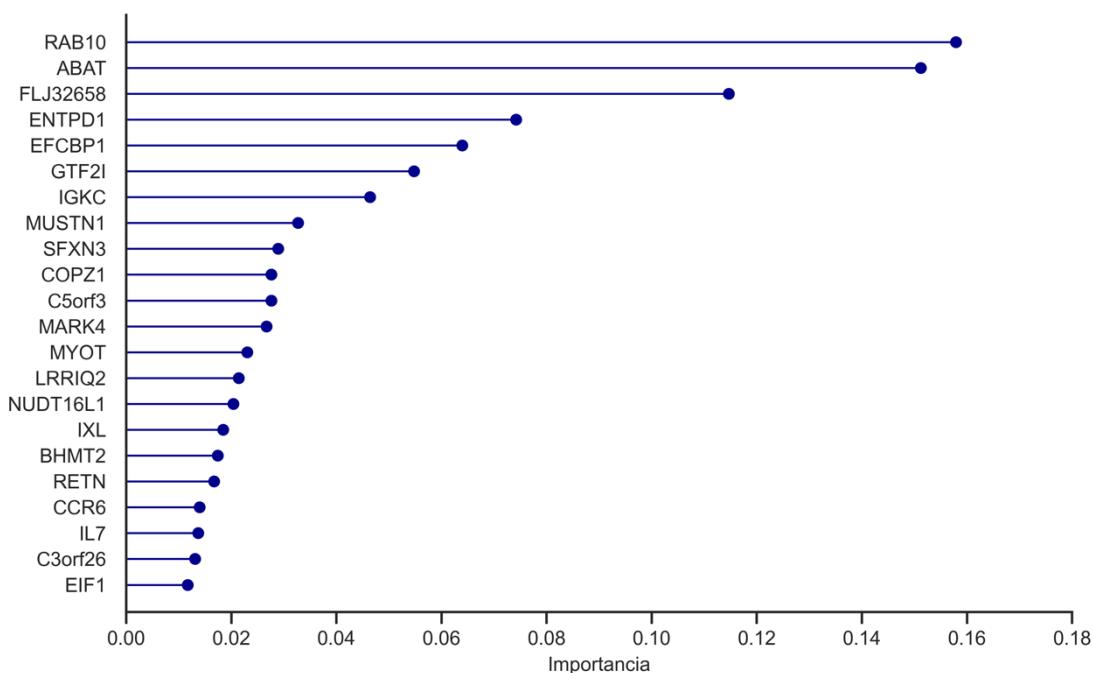


Figura 25. Ranking de la relevancia de las variables del *dataset* biomarcadores seleccionadas mediante el filtro reliefF y wrapper para la predicción del Alzheimer con el modelo XGBoost.

4.5.2.3. Predicción MMSE

Para el entrenamiento y optimización de los modelos de regresión para la predicción de los valores del MMSE, se emplearon 29 variables. Tras el entrenamiento y optimización de los modelos se ha determinado que el modelo que mejor funcionaba era CatBoost, y se ha analizado el orden de importancia de esas variables en la obtención de las predicciones del modelo (**Figura 26**). En este caso, la proteína de mayor relevancia es EIF2AK2, que es una quinasa activada por RNA bicatenario inducida por interferón, seguidas de proteínas como ECH1 (isomerasa mitocondrial), PRELP (prolargin), NUDT2 (hidrolasa nudix 2), y GATA3 (factor de transcripción específico de células T transactivador GATA-3).

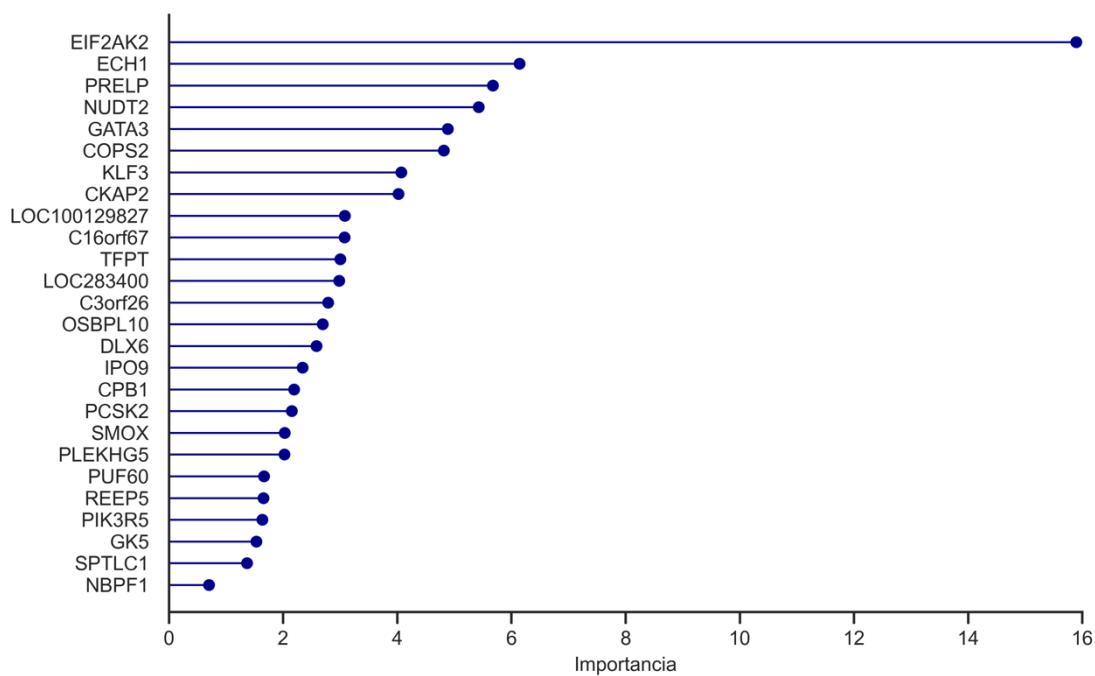


Figura 26. Ranking de la relevancia en el modelo CatBoost de las variables del *dataset* biomarcadores para la predicción del valor del MMSE.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Para el desarrollo de este Trabajo Fin de Máster se han empleado dos *datasets*, uno que incluye diferentes factores de riesgo asociados con el Alzheimer y el otro con una serie de biomarcadores analizados mediante un estudio de proteómica con muestras de suero de pacientes. El objetivo principal de estudio es poder identificar de manera precisa los pacientes que padecen Alzheimer. Con el *dataset* factores de riesgo se han desarrollado 7 modelos (Decision Tree, Random Forest, KNN, SVM, Logistic Regression, XGBoost y CatBoost). Con el *dataset* biomarcadores, se ha realizado el balanceado de clases y se han empleado dos métodos diferentes para la selección de características (filtro por el p-valor y mediane reliefF y wrapper). Una vez seleccionadas las variables para cada conjunto se han desarrollado también los mismos 7 modelos que con el *dataset* factores de riesgo. Por otro lado, con el *dataset* biomarcadores también se han empleado diferentes modelos (Decision Tree, Random Forest, KNN, SVR, Ridge Regression, XGBoost y CatBoost) para tratar de predecir el pronóstico y avance de la enfermedad en los pacientes que ya están diagnosticados, prediciendo el valor del MMSE. En este último caso la selección de variables para entrenar los modelos se ha realizado mediante reliefF y wrapper.

Como se ha comentado, en el caso del *dataset* biomarcadores se ha empleado la reducción o selección de características, que permite tanto evitar el sobreajuste y obtener

modelos más generalizables como obtener un número de variables más manejable que se pueda emplear a nivel clínico. Reduciendo el número de variables se puede buscar una combinación de biomarcadores que se puedan emplear a nivel clínico para el diagnóstico de la enfermedad, reduciendo el coste y consiguiendo métodos más precisos. En el caso del filtro del p-valor se han seleccionado las proteínas que mayor diferencia presentan entre el grupo de pacientes con Alzheimer, y el resto de grupos. Sin embargo, con este filtro se pueden obtener variables que estén relacionadas entre sí y que tengan comportamientos similares entre estos grupos, por lo que puede que, aunque se estén seleccionando las variables que permiten diferenciar el grupo de Alzheimer del resto puede que no se estén seleccionando las variables más relevantes. Por ello, se ha empleado otro método complementario para la selección de variables basado en el uso de reliefF y la aplicación posterior de métodos wrapper. Estas técnicas permiten considerar interacciones entre las variables y el contexto multivariado, obteniendo una evaluación más global de la relevancia de cada variable en el rendimiento de los modelos. A diferencia del método anterior que era univariado, estos métodos tienen en cuenta la contribución de cada variable al modelo en presencia del resto de variables, permitiendo identificar variables que, aunque no sean muy diferentes entre los grupos de forma individual, pueden ser más relevantes si se combinan con otras. En este sentido, ReliefF evalúa la relevancia de cada variable para distinguir entre observaciones cercanas, pero de clases diferentes. Por otro lado, los métodos wrapper seleccionan conjuntos de variables basándose en el rendimiento de los modelos. En este caso concreto se ha empleado la eliminación recursiva de variables con validación cruzada con el modelo Random Forest, de forma que se van eliminando las variables con menos importancia en cada modelo entrenado hasta que se alcanza un número óptimo de variables. De esta forma también es el propio modelo el que determina el número óptimo de variables. En este caso se ha determinado que el número óptimo de variables para el desarrollo de los modelos de predicción del diagnóstico del Alzheimer es 23 (Anexo). Esto permite obtener modelos más robustos y clínicamente útiles, evitando la redundancia y optimizando el conjunto de variables seleccionadas.

Analizando los resultados para cada caso, se ha obtenido que el modelo que mejores resultados consigue para el *dataset* factores de riesgo es CatBoost, con un *recall* de 92,6% y un AUC de 94,9%. Para el *dataset* biomarcadores con la selección de variables realizada con el filtro del p-valor el modelo que mejores resultados obtiene es Random Forest, con un *recall* de 93,2% y un AUC de 92,0%. Para este mismo *dataset* pero con las variables seleccionadas con reliefF y wrapper el mejor modelo es XGBoost, con un *recall* de 95,5% y un AUC de 95,4%. Según los valores de referencia para determinar la calidad de los modelos desarrollados se puede considerar que los tres modelos obtenidos serían muy buenos modelos para predecir el diagnóstico del Alzheimer [69,70,76], resaltando los valores obtenidos para del *dataset* biomarcadores con reliefF y wrapper. Además, si se compara estos resultados con los de la **Tabla 1** del estado del arte se puede confirmar

que estos modelos mejoran a los obtenidos previamente, a excepción del estudio de Karaglani et al., [47] con un AUC de 97,5% con datos de transcriptómica (mRNA), y el estudio de Chiu et al., [45] con un AUC de 95-98%, que emplean los modelos SVM y *information gain*, respectivamente.

A partir de los datos obtenidos, se ha observado que algunos algoritmos como KNN, SVM y Logistic Regression no han alcanzado buenos rendimientos en comparación con los modelos más avanzados como Random Forest, XGBoost o CatBoost. Esto puede deberse a la naturaleza de los datos empleados en el estudio, ya que puede haber relaciones no lineales y complejas entre las variables que afectan negativamente a algoritmos como Logistic Regression o KNN. Por otro lado, SVM, aunque es más sensible por poder manejar relaciones no lineales mediante el uso de kernels, la presencia de valores atípicos y de ruido o la distribución no homogénea de las clases pueden haber dificultado el entrenamiento del modelo o incluso producir un sobreajuste a los datos de entrenamiento.

En cuanto a las tareas de regresión que se han realizado en el estudio, con la finalidad de predecir el pronóstico y avance de la enfermedad, mediante la puntuación del MMSE, el modelo que mejores resultados ha obtenido es CatBoost, con un RMSE de 3,35 lo que implica un RMSE relativo del 15%. Teniendo en cuenta los valores de referencia se puede considerar que el modelo desarrollado es de calidad media, ya que podría obtener buenos resultados, pero necesita mejoras para alcanzar una mayor precisión [69,70,76]. Uno de los motivos por los que no se alcanzan resultados mejores puede ser debido a que no hay una distribución equitativa de los sujetos en los diferentes rangos del MMSE. Debido a esto valores de MMSE por debajo de 14 están poco representados y por tanto los modelos no pueden adaptarse correctamente a este rango de valores, contribuyendo a los errores del modelo.

Por otro lado, mediante la plataforma The Database for Annotation, Visualization, and Integrated Discovery (DAVID) de National Institutes of Health (NIH) (<https://davidbioinformatics.nih.gov/home.jsp>) se puede analizar la función de las proteínas seleccionadas con los diferentes filtros y también las que mayor relevancia tienen en los modelos obtenidos, con la finalidad de descubrir cuales son las rutas o funciones que podrían estar alteradas en la enfermedad, y que además también pueden llevar al descubrimiento de nuevas dianas terapéuticas. Mediante las anotaciones funcionales de esta plataforma se puede observar que las vías principalmente representadas con las proteínas obtenidas con el filtro del p-valor son rutas implicadas con el metabolismo del RNA (PPP1R8, ADARB1, RPL11) y la fosforilación de proteínas. Con el filtro de reliefF y wrapper las vías principalmente implicadas están relacionadas con el sistema inmunitario (RAB10, CCR6, IGKC, RETN), y también con la fosforilación de proteínas (ABAT, GTF2I). Para la predicción del avance de la enfermedad las proteínas seleccionadas están relacionadas con procesos del sistema inmune, apoptosis, unión y daños en el ADN y fosforilación de proteínas. Y concretamente, las que mayor relevancia han presentado en el modelo, EIF2AK2 y GATA3 están relacionadas con el

sistema inmune, y ECH1 y NUDT2, están involucradas en la función mitocondrial. Todos estos hallazgos, en cuanto a la función de las proteínas filtradas y que mayor relevancia presentan en los modelos, concuerdan con la bibliografía publicada sobre las vías que se conocen que están implicadas en la enfermedad. Sin embargo, este estudio podría ayudar a determinar cuáles son las proteínas concretas que están alterando estas vías y que dan lugar a la patogénesis de la enfermedad.

En conclusión, los resultados obtenidos en este estudio demuestran que con la metodología empleada es posible identificar con una precisión alta y un número reducido de biomarcadores a los pacientes con Alzheimer. Esto podría resultar de gran interés a nivel clínico ya que se podría diagnosticar de manera más efectiva y precisa esta enfermedad, reduciendo la cantidad de pruebas diagnósticas invasivas y su coste ya que en este caso se emplea suero de los pacientes que es un método poco invasivo y ampliamente empleado en los estudios rutinarios. Además, se han identificado proteínas que pueden ser cruciales en la enfermedad lo que podría ayudar en el avance de la investigación básica del Alzheimer. No obstante, para que estos resultados puedan trasladarse a la práctica clínica y que el modelo tenga una aplicabilidad real, se debería realizar una evaluación del rendimiento del modelo en nuevas cohortes, que permita determinar su robustez ante variaciones técnicas y biológicas. Con la finalidad de validar los biomarcadores seleccionados para la construcción de los modelos y el diagnóstico, se podrían realizar estudios prospectivos en los que calcular la sensibilidad y especificidad del uso de estos biomarcadores en el diagnóstico y su comparación con los métodos de diagnóstico actuales. Todos estos pasos son esenciales para confirmar la utilidad diagnóstica del modelo y de los biomarcadores seleccionados, su reproducibilidad en condiciones reales y su potencial para mejorar la detección temprana del Alzheimer en contextos clínicos hospitalarios.

5.2. Limitaciones del estudio

Este estudio también presenta una serie de limitaciones. Por un lado, en el *dataset* factores de riesgo las variables que se recogen y que son las más relevantes para la decisión del modelo seleccionado son los criterios que se emplean actualmente para el diagnóstico de la enfermedad, por lo que en este caso no ha permitido obtener conocimiento nuevo sino confirmar el buen funcionamiento de los criterios actuales. Por otro lado, en cuanto al *dataset* biomarcadores, este no incluye un número muy alto de registros y aunque incluía diferentes clases de interés, que permitirían diferenciar a los pacientes con Alzheimer tanto de controles sanos como de pacientes con otras enfermedades más relacionadas, como el Parkinson, o diferentes, como el cáncer de mama, estas estaban muy desbalanceadas y para el entrenamiento del modelo se han empleado nuevas muestras generadas sintéticamente, por lo que puede que el modelo generado no se adapte bien a

los datos reales. Además, otra de las limitaciones que se ha comentado anteriormente sería que no se dispone de muestras suficientes para el rango inferior de los valores del MMSE, para que el modelo consiga realizar predicciones más precisas.

5.3. Trabajo futuro

Una vez realizado el estudio se plantean una serie de propuestas para futuros estudios que podrían contribuir a enriquecer los resultados obtenidos, así como inferir nuevos conocimientos:

- Ampliar el número de registros para el *dataset* biomarcadores, con la finalidad de obtener modelos más robustos y generalizables.
- Ampliar el número de registros para alcanzar el balanceado de clases con casos reales y no generados sintéticamente.
- También sería interesante incluir el valor del MMSE para todas las clases y no únicamente para la clase con Alzheimer, permitiendo así obtener valores en todos los rangos y que los modelos puedan alcanzar una mayor precisión.
- Otro aspecto importante podría ser introducir nuevas clases de pacientes que tengan diagnosticado demencia pero que no sea Alzheimer, para la obtención de modelos que sean capaces de discriminar entre diferentes tipos de demencia, lo que es crucial para determinar el tratamiento óptimo para los pacientes.
- Aunque los modelos empleados han obtenido buenos resultados, también podría probarse técnicas más sofisticadas como el *Deep Learning* y las redes neuronales.
- Por último, se podría realizar una clasificación de los pacientes con Alzheimer que permita determinar diferentes tipos de la enfermedad, lo que podría ayudar a predecir el pronóstico de la enfermedad y a establecer tratamientos más personalizados.

Capítulo 6

Glosario

ACC	Precisión
AD	Alzheimer
ANN	<i>Artifical neural network</i>
AUC	Área bajo la curva ROC
bvFTD	<i>Behavioural variant frontotemporal dementia</i>
C	Control
CatBoost	<i>Categorical Boosting</i>
DD	Diagnóstico con demencia
DL	<i>Deep learning</i>
DT	<i>Decision Tree</i>
FN	Falsos negativos
FP	Falsos positivos
GDPR	Reglamento General de Protección de Datos
KNN	<i>K-nearest neighbors</i>
LD	Viviendo con demencia
LDA	<i>Linear Discriminant Analysis</i>
LR	<i>Logistic Regression</i>
LSTM	<i>Long Short Term Memory</i>
MAE	Error absoluto medio

MCI	<i>Mild cognitive impairment</i>
MMSE	<i>Mini-Mental State Examination</i>
MNCD	<i>Major neurocognitive disorder</i>
MSE	Error cuadrático medio
NB	Naïve Bayes
RF	<i>Random Forest</i>
RMSE	Raíz del error cuadrático medio
ROC	<i>Receiver Operating Characteristic</i>
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
SVM	<i>Support vector machine</i>
VMD	<i>Very-mild dementia</i>
VN	Verdaderos negativos
VP	Verdaderos positivos
XGBoost	<i>Extreme Gradient Boosting</i>

Bibliografía

1. *Dementia in Europe Yearbook 2019: Estimating the Prevalence of Dementia in Europe*, 2020.
2. Lane, C.A.; Hardy, J.; Schott, J.M. Alzheimer's Disease. *Eur J Neurol* 2018, *25*, 59–70.
3. Estimation of the Global Prevalence of Dementia in 2019 and Forecasted Prevalence in 2050: An Analysis for the Global Burden of Disease Study 2019 *Lancet Public Health* 2022, *7*, e105–e125.
4. 2023 Alzheimer's Disease Facts and Figures *Alzheimers Dement* 2023, *19*, 1598–1695.
5. Lai, Y.; Lin, P.; Lin, F.; Chen, M.; Lin, C.; Lin, X.; Wu, L.; Zheng, M.; Chen, J. Identification of Immune Microenvironment Subtypes and Signature Genes for Alzheimer's Disease Diagnosis and Risk Prediction Based on Explainable Machine Learning *Front Immunol* 2022, *13*, 1046410.
6. Knopman, D.S.; Amieva, H.; Petersen, R.C.; Chételat, G.; Holtzman, D.M.; Hyman, B.T.; Nixon, R.A.; Jones, D.T. Alzheimer Disease *Nat Rev Dis Primers* 2021, *7*, 33.
7. Singh-Manoux, A.; Dugavot, A.; Fournier, A.; Abell, J.; Ebmeier, K.; Kivimäki, M.; Sabia, S. Trajectories of Depressive Symptoms before Diagnosis of Dementia: A 28-Year Follow-Up Study *JAMA Psychiatry* 2017, *74*, 712–718.
8. Gottesman, R.F.; Albert, M.S.; Alonso, A.; Coker, L.H.; Coresh, J.; Davis, S.M.; Deal, J.A.; McKhann, G.M.; Mosley, T.H.; Sharrett, A.R. *et al.* Associations between Midlife Vascular Risk Factors and 25-Year Incident Dementia in the Atherosclerosis Risk in Communities (ARIC) Cohort *JAMA Neurol* 2017, *74*, 1246–1254.
9. Thambisetty, M.; An, Y.; Tanaka, T. Alzheimer's Disease Risk Genes and the Age-at-Onset Phenotype *Neurobiol Aging* 2013, *34*, 2696.e1–5.
10. Tzioras, M.; McGeachan, R.I.; Durrant, C.S.; Spires-Jones, T.L. Synaptic Degeneration in Alzheimer Disease *Nat Rev Neurol* 2023, *19*, 19–38.
11. Hampel, H.; Hardy, J.; Blennow, K.; Chen, C.; Perry, G.; Kim, S.H.; Villemagne, V.L.; Aisen, P.; Vendruscolo, M.; Iwatsubo, T. *et al.* The Amyloid-B Pathway in Alzheimer's Disease *Mol Psychiatry* 2021, *26*, 5481–5503.

12. Jucker, M.; Walker, L.C. Alzheimer's Disease: From Immunotherapy to Immunoprevention *Cell* 2023, *186*, 4260–4270.
13. Jucker, M.; Walker, L.C. Propagation and Spread of Pathogenic Protein Assemblies in Neurodegenerative Diseases *Nat Neurosci* 2018, *21*, 1341–1349.
14. Masters, C.L.; Bateman, R.; Blennow, K.; Rowe, C.C.; Sperling, R.A.; Cummings, J.L. Alzheimer's Disease *Nat Rev Dis Primers* 2015, *1*, 15056.
15. Arnold, S.E.; Hyman, B.T.; Flory, J.; Damasio, A.R.; Van Hoesen, G.W. The Topographical and Neuroanatomical Distribution of Neurofibrillary Tangles and Neuritic Plaques in the Cerebral Cortex of Patients with Alzheimer's Disease *Cereb Cortex* 1991, *1*, 103–116.
16. D'Onofrio, G.; Sancarlo, D.; Panza, F.; Copetti, M.; Cascavilla, L.; Paris, F.; Seripa, D.; Matera, M.G.; Solfrizzi, V.; Pellegrini, F. *et al.* Neuropsychiatric Symptoms and Functional Status in Alzheimer's Disease and Vascular Dementia Patients *Curr Alzheimer Res* 2012, *9*, 759–771.
17. Cano, A.; Turowski, P.; Ettcheto, M.; Duskey, J.T.; Tosi, G.; Sánchez-López, E.; García, M.L.; Camins, A.; Souto, E.B.; Ruiz, A. *et al.* Nanomedicine-Based Technologies and Novel Biomarkers for the Diagnosis and Treatment of Alzheimer's Disease: From Current to Future Challenges *J Nanobiotechnology* 2021, *19*, 122.
18. Blennow, K.; Hampel, H.; Weiner, M.; Zetterberg, H. Cerebrospinal Fluid and Plasma Biomarkers in Alzheimer Disease *Nat Rev Neurol* 2010, *6*, 131–144.
19. Jack, C.R.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J. *et al.* NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease *Alzheimers Dement* 2018, *14*, 535–562.
20. McKhann, G.M.; Knopman, D.S.; Chertkow, H.; Hyman, B.T.; Jack, C.R.; Kawas, C.H.; Klunk, W.E.; Koroshetz, W.J.; Manly, J.J.; Mayeux, R. *et al.* The Diagnosis of Dementia due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease *Alzheimers Dement* 2011, *7*, 263–269.
21. Atri, A. The Alzheimer's Disease Clinical Spectrum: Diagnosis and Management *Med Clin North Am* 2019, *103*, 263–293.
22. Davda, N.; Corkill, R. Biomarkers in the Diagnosis and Prognosis of Alzheimer's Disease *J Neurol* 2020, *267*, 2475–2477.
23. Chang, C.; Lin, C.; Lane, H. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease *Int J Mol Sci* 2021, *22*, 2761.
24. Ausó, E.; Gómez-Vicente, V.; Esquia, G. Biomarkers for Alzheimer's Disease Early Diagnosis *J Pers Med* 2020, *10*, 114.
25. Canevelli, M.; Bacigalupo, I.; Gervasi, G.; Lacorte, E.; Massari, M.; Mayer, F.; Vanacore, N.; Cesari, M. Methodological Issues in the Clinical Validation of

- Biomarkers for Alzheimer's Disease: The Paradigmatic Example of CSF Front Aging Neurosci 2019, *11*, 282.
26. Rossini, P.M.; Di Iorio, R.; Vecchio, F.; Anfossi, M.; Babiloni, C.; Bozzali, M.; Bruni, A.C.; Cappa, S.F.; Escudero, J.; Fraga, F.J. *et al.* Early Diagnosis of Alzheimer's Disease: The Role of Biomarkers Including Advanced EEG Signal Analysis. Report from the IFCN-Sponsored Panel of Experts Clin Neurophysiol 2020, *131*, 1287–1310.
27. Khoury, R.; Ghossoub, E. Diagnostic Biomarkers of Alzheimer's Disease: A State-of-the-Art Review Biomarkers in Neuropsychiatry 2019, *1*, 100005.
28. Blennow, K.; Dubois, B.; Fagan, A.M.; Lewczuk, P.; de Leon, M.J.; Hampel, H. Clinical Utility of Cerebrospinal Fluid Biomarkers in the Diagnosis of Early Alzheimer's Disease Alzheimers Dement 2015, *11*, 58–69.
29. Maimin Oded; Rokach Lior. Data Mining and Knowledge Discovery Handbook2005.
30. Dwyer, D.B.; Falkai, P.; Koutsouleris, N. Machine Learning Approaches for Clinical Psychology and Psychiatry Annu Rev Clin Psychol 2018, *14*, 91–118.
31. Myszczynska, M.A.; Ojamies, P.N.; Lacoste, A.M.B.; Neil, D.; Saffari, A.; Mead, R.; Hautbergue, G.M.; Holbrook, J.D.; Ferraiuolo, L. Applications of Machine Learning to Diagnosis and Treatment of Neurodegenerative Diseases Nat Rev Neurol 2020, *16*, 440–456.
32. Javeed, A.; Dallora, A.L.; Berglund, J.S.; Ali, A.; Ali, L.; Anderberg, P. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions J Med Syst 2023, *47*, 17.
33. Hane, C.A.; Nori, V.S.; Crown, W.H.; Sanghavi, D.M.; Bleicher, P. Predicting Onset of Dementia using Clinical Notes and Machine Learning: Case-Control Study JMIR Med Inform 2020, *8*, e17819.
34. Ryu, S.; Shin, D.; Chung, K. Prediction Model of Dementia Risk Based on XGBoost using Derived Variable Extraction and Hyper Parameter Optimization IEEE Access 2020, *8*, 177708–177720.
35. Stamate, D.; Alghamdi, W.; Ogg, J.; Hoile, R.; Murtagh, F. A Machine Learning Framework for Predicting Dementia and Mild Cognitive Impairment. In ; pp. 671–678.
36. Bansal, D.; Chhikara, R.; Khanna, K.; Gupta, P. Comparative Analysis of various Machine Learning Algorithms for Detecting Dementia Procedia Computer Science 2018, *132*, 1497–1502.
37. Cleret de Langavant, L.; Bayen, E.; Yaffe, K. Unsupervised Machine Learning to Identify High Likelihood of Dementia in Population-Based Surveys: Development and Validation Study J Med Internet Res 2018, *20*, e10493.
38. Stamate, D.; Kim, M.; Proitsi, P.; Westwood, S.; Baird, A.; Nevado-Holgado, A.; Hye, A.; Bos, I.; Vos, S.J.B.; Vandenberghe, R. *et al.* A Metabolite-Based Machine Learning Approach to Diagnose Alzheimer-Type Dementia in Blood:

- Results from the European Medical Information Framework for Alzheimer Disease Biomarker Discovery Cohort Alzheimers Dement (N Y) 2019, *5*, 933–938.
39. Garcia-Gutierrez, F.; Delgado-Alvarez, A.; Delgado-Alonso, C.; Díaz-Álvarez, J.; Pytel, V.; Valles-Salgado, M.; Gil, M.J.; Hernández-Lorenzo, L.; Matías-Guiu, J.; Ayala, J.L. *et al.* Diagnosis of Alzheimer's Disease and Behavioural Variant Frontotemporal Dementia with Machine Learning-Aided Neuropsychological Assessment using Feature Engineering and Genetic Algorithms Int J Geriatr Psychiatry 2021, *37*.
40. Balea-Fernandez, F.J.; Martinez-Vega, B.; Ortega, S.; Fabelo, H.; Leon, R.; Callico, G.M.; Bibao-Sieyro, C. Analysis of Risk Factors in Dementia through Machine Learning J Alzheimers Dis 2021, *79*, 845–861.
41. Fouladvand, S.; Mielke, M.M.; Vassilaki, M.; Sauver, J.S.; Petersen, R.C.; Sohn, S. Deep Learning Prediction of Mild Cognitive Impairment using Electronic Health Records Proceedings (IEEE Int Conf Bioinformatics Biomed) 2019, *2019*, 799–806.
42. Salem, F.A.; Chaaya, M.; Ghannam, H.; Al Feel, R.E.; El Asmar, K. Regression Based Machine Learning Model for Dementia Diagnosis in a Community Setting Alzheimer's & Dementia 2021, *17*, e053839.
43. Qazi, N.; Raza, K. Effect of Feature Selection, SMOTE and Under Sampling on Class Imbalance Classification. In ; pp. 145–150.
44. Gurevich, P.; Stuke, H.; Kastrup, A.; Stuke, H.; Hildebrandt, H. Neuropsychological Testing and Machine Learning Distinguish Alzheimer's Disease from Other Causes for Cognitive Impairment Front Aging Neurosci 2017, *9*, 114.
45. Chiu, P.; Tang, H.; Wei, C.; Zhang, C.; Hung, G.; Zhou, W. NMD-12: A New Machine-Learning Derived Screening Instrument to Detect Mild Cognitive Impairment and Dementia PLoS One 2019, *14*, e0213430.
46. Nori, V.S.; Hane, C.A.; Martin, D.C.; Kravetz, A.D.; Sanghavi, D.M. Identifying Incident Dementia by Applying Machine Learning to a very Large Administrative Claims Dataset PLoS One 2019, *14*, e0203246.
47. Karaglani, M.; Gourlia, K.; Tsamardinos, I.; Chatzaki, E. Accurate Blood-Based Diagnostic Biosignatures for Alzheimer's Disease Via Automated Machine Learning J Clin Med 2020, *9*, 3016.
48. Ryzhikova, E.; Ralbovsky, N.M.; Sikirzhynski, V.; Kazakov, O.; Halamkova, L.; Quinn, J.; Zimmerman, E.A.; Lednev, I.K. Raman Spectroscopy and Machine Learning for Biomedical Applications: Alzheimer's Disease Diagnosis Based on the Analysis of Cerebrospinal Fluid Spectrochim Acta A Mol Biomol Spectrosc 2021, *248*, 119188.
49. Shahzad, A.; Dadlani, A.; Lee, H.; Kim, K. Automated Prescreening of Mild Cognitive Impairment using Shank-Mounted Inertial Sensors Based Gait Biomarkers IEEE Access 2022, *10*, 15835–15844.

50. Jin, H.; Chien, S.; Meijer, E.; Khobragade, P.; Lee, J. Learning from Clinical Consensus Diagnosis in India to Facilitate Automatic Classification of Dementia: Machine Learning Study JMIR Ment Health 2021, *8*, e27113.
51. Facal, D.; Valladares-Rodríguez, S.; Lojo-Seoane, C.; Pereiro, A.X.; Anido-Rifon, L.; Juncos-Rabadán, O. Machine Learning Approaches to Studying the Role of Cognitive Reserve in Conversion from Mild Cognitive Impairment to Dementia Int J Geriatr Psychiatry 2019, *34*, 941–949.
52. Hsiu, H.; Lin, S.; Weng, W.; Hung, C.; Chang, C.; Lee, C.; Chen, C. Discrimination of the Cognitive Function of Community Subjects using the Arterial Pulse Spectrum and Machine-Learning Analysis Sensors 2022, *22*, 806.
53. Patil, T.R.; Sherekar, M.S.S. Performance Analysis of Naive Bayes and J 48 Classification Algorithm for Data Classification 2013.
54. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993 Mach Learn 1994, *16*, 235–240.
55. Casas Roma, J.; Minguillón Alfonso, J. Preparación De Los Datos. Editorial UOC.
56. Montoliu Colás, R. Preprocesado De Datos. Editorial UOC.
57. Minguillón Alfonso, J.; Caihuelas Quiles, R. Proceso De Minería De Datos. Editorial UOC.
58. Gironés Roig, J. Gestión De Características. Editorial UOC.
59. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection The Journal of Machine Learning Research 2003, *3*, 1157–1182.
60. Google developers. Datasets: Imbalanced Datasets. <Https://Developers.Google.Com/Machine-Learning/Crash-Course/Overfitting/Imbalanced-Datasets>. 2025.
61. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions Prog Artif Intell 2016, *5*, 221–232.
62. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique jair 2002, *16*, 321–357.
63. Hira, Z.M.; Gillies, D.F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data Adv Bioinformatics 2015, *2015*, 198363.
64. Bernadó Mansilla, E.; Vinuesa, T.S. Contrastos De Hipótesis. Editorial UOC.
65. Bland, J.M.; Altman, D.G. Multiple Significance Tests: The Bonferroni Method. BMJ 1995, *310*, 170.
66. Mullan, K.A.; Bramberger, L.M.; Munday, P.R.; Goncalves, G.; Revote, J.; Mifsud, N.A.; Illing, P.T.; Anderson, A.; Kwan, P.; Purcell, A.W. *et al.* ggVolcanoR: A Shiny App for Customizable Visualization of Differential Expression Datasets Comput Struct Biotechnol J 2021, *19*, 5735–5740.
67. Scikit-learn. RFECV. Https://Scikit-Learn/Stable/Modules/Generated/Sklearn.Feature_selection.RFECV.Html.
68. Montoliu Colás, R. Modelos Supervisados. Editorial UOC.

69. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *An Introduction to Statistical Learning with Applications in Python*, 2023.
70. Cichosz, P. *Data Mining Algorithms*; Wiley, 2015.
71. Hackeling, G. *Mastering Machine Learning with Scikit-Learn*.; Packt Publishing, 2017.
72. López, O.A.M.; López, A.M.; Crossa, D.J. Support Vector Machines and Support Vector Regression. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet]*.; Anonymous .; Springer, 2022.
73. van Wieringen, W.N. *Lecture Notes on Ridge Regression*, 2023.
74. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System2016, 785–794.
75. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features 2019.
76. Montoliu Colás, R. Evaluación De Los Modelos. Editorial UOC.

Anexo

Lista de variables en el *dataset* factores de riesgo

1. Age (edad): edad de los pacientes (60-90 años).
2. Gender (Sexo): Sexo de los pacientes, 0 = Hombre, 1 = Mujer.
3. Ethnicity (origen étnico): Origen étnico de los pacientes, 0 = Caucásico, 1 = Afroamericano, 2 = Asiatico, 3 = Otro.
4. EducationLevel (Nivel de educación): Nivel educativo de los pacientes, 0 = Sin educación reglada, 1 = Graduado escolar, 2 = Grado Universitario, 3 = Educación superior.
5. BMI: índice de masa corporal de los pacientes (15-40).
6. Smoking (Tabaquismo): Pacientes fumadores, 0 = No, 1 = Sí.
7. AlcoholConsumption (Consumo de alcohol): Consumo semanal de alcohol (0-20).
8. PhysicalActivity (Actividad física): Horas semanales de actividad física (0-10)
9. DietQuality (Calidad de la diesta): Puntuación de la calidad de la dieta (0-10).
10. SleepQuality (Calidad del sueño): Puntuación de la calidad del sueño (4-10).
11. FamilyHistoryAlzheimers (Historial familiar de Alzheimer): Historia familiar del Alzheimer, 0 = No, 1 = Sí.

12. CardiovascularDisease (Enfermedades cardiovasculares): Presencia de enfermedades cardiovasculares, 0 = No, 1 = Sí.
13. Diabetes: Presencia de diabetes, 0 = No, 1 = Sí.
14. Depression (Depresión): Presencia de depresión, 0 = No, 1 = Sí.
15. HeadInjury (lesiones en la cabeza): Historial de lesiones en la cabeza, 0 = No, 1 = Sí.
16. Hypertension (Hipertensión): Presecia de hipertensión, 0 = No, 1 = Sí.
17. SystolicBP (Presión arterial sistólica): Tensión arterial sistólica (90-180 mm Hg).
18. DiastolicBP (Presió arterial distólica): Tensión arterial distólica (60-120 mm Hg).
19. CholesterolTotal (Colesterol total): Niveles de colesterol total (150-300 mg/dL).
20. CholesterolLDL (Colesterol LDL): Niveles de colesterol LDL (50-200 mg/dL).
21. CholesterolHDL (Colesterol HDL): Niveles de colesterol HDL (20-100 mg/dL).
22. CholesterolTriglycerides (Triglicéridos): Niveles de triglicéridos (50-400 mg/dL).
23. MMSE: Puntuación del mini examen del estado mental (0 – 30). Niveles bajos determinan un deterioro cognitivo. 27-30 = Estado cognitivo normal, 24-26 = Deterioro cognitivo leve, 19-23 = Deterioro moderado (demencia incipiente, Alzheimer en etapas tempranas), 14-18: Deterioro cognitivo moderado-grave (demencia establecida), 0-13: Deterioro cognitivo grave (demencia avanzada, Alzheimer severo).
24. FunctionalAssessment (Evaluación funcional): Puntuación de la evaluación funcional (0-10). Puntuaciones bajas indican un mayor deterioro.
25. MemoryComplaints (Problemas de memoria): Presencia de problemas de memoria, 0 = No, 1 = Sí.
26. BehavioralProblems (Problemas conductuales): Presencia de problemas conductuales, 0 = No, 1 = Sí.
27. ADL: Puntuaciones de las actividades de la vida cotidiana (0-10). Puntuaciones más bajas indican un mayor deterioro.
28. Confusion: Presencia de confusiones, 0 = No, 1 = Sí.

29. Disorientation (Desorientación): Presencia de desorientación, 0 = No, 1 = Sí.
30. PersonalityChanges (Cambios de personalidad): Presencia de cambios de personalidad, 0 = No, 1 = Sí.
31. DifficultyCompletingTasks (Dificultad para completar tareas): Presencia de dificultad para completar tareas, 0 = No, 1 = Sí.
32. Forgetfulness (Pérdidas de memoria): Presencia de pérdidas de memoria, 0 = No, 1 = Sí.

Lista de variables del dataset biomarcadores con el filtro p-valor

Variable	Abreviatura génica	Proteína
3	C16orf30	chromosome 16 open reading frame 30 (C16orf30)
4	DYNC1I1	dynein, cytoplasmic 1, intermediate chain 1 (DYNC1I1)
13	BHMT2	betaine-homocysteine methyltransferase 2 (BHMT2)
20	ENTPD1	cDNA clone MGC:12418 IMAGE:3934658, complete cds
27	CCDC99	coiled-coil domain containing 99 (CCDC99)
31	CLDN19	claudin 19 (CLDN19)
38	GBA2	glucosidase, beta (bile acid) 2 (GBA2)
40	EFCBP1	EF-hand calcium binding protein 1 (EFCBP1)
41	CACNG2	calcium channel, voltage-dependent, gamma subunit 2 (CACNG2)
45	RSRC2	arginine/serine-rich coiled-coil 2 (RSRC2)
59	BAIAP2	BAI1-associated protein 2 (BAIAP2), transcript variant 2
61	MUSTN1	musculoskeletal, embryonic nuclear protein 1 (MUSTN1)
62	LOC150568	PREDICTED: Homo sapiens hypothetical protein LOC150568 (LOC150568)
65	MEST	mesoderm specific transcript homolog (mouse) (MEST), transcript variant 2
76	C20orf198	chromosome 20 open reading frame 198 (C20orf198)

Variable	Abreviatura génica	Proteína
77	PACRG	Parkin coregulated gene protein homolog
78	FAM126B	family with sequence similarity 126, member B (FAM126B)
90	ZMAT4	Zinc finger matrin-type protein 4
92	AKTIP	AKT interacting protein (AKTIP), transcript variant 2
93	ADPRHL2	ADP-ribosylhydrolase like 2 (ADPRHL2)
94	PRMT1	Protein arginine N-methyltransferase 1
96	MCM10	cDNA clone IMAGE:3451214 (MCM10)
103	CDC42SE1	CDC42 small effector 1 (CDC42SE1), transcript variant 2
105	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa (TAF6), transcript variant 1
114	C3orf26	chromosome 3 open reading frame 26 (C3orf26)
123	C9orf37	chromosome 9 open reading frame 37 (C9orf37)
126	SCAMP3	secretory carrier membrane protein 3 (SCAMP3), transcript variant 2
131	MAP2K6, mutant	mitogen-activated protein kinase kinase 6 (MAP2K6), transcript variant 2; mutant protein: MAP2K6 S207E, T211E
136	EIF1	eukaryotic translation initiation factor 1 (EIF1)
144	FTSJ1	FtsJ homolog 1 (E. coli) (FTSJ1), transcript variant 1
153	BIRC5	baculoviral IAP repeat-containing 5 (survivin) (BIRC5)
207	ASCIZ	ATM/ATR-Substrate Chk2-Interacting Zn2+-finger protein (ASCIZ)
220	C2orf13	chromosome 2 open reading frame 13 (C2orf13)
235	RAB3IL1	Guanine nucleotide exchange factor for Rab3A
243	LARP6	La ribonucleoprotein domain family, member 6 (LARP6), transcript variant 1
249	TARDBP	TAR DNA-binding protein 43
251	LOC221711	Synaptonemal complex protein 2-like
257	CCDC15	Coiled-coil domain-containing protein 15
304	SMR3B	submaxillary gland androgen regulated protein 3 homolog B (mouse) (SMR3B)
312	USP47	ubiquitin specific peptidase 47 (USP47)
327	SPHK2	sphingosine kinase 2 (SPHK2)
335	HTATIP	HIV-1 Tat interacting protein, 60kDa (HTATIP)

Variable	Abreviatura génica	Proteína
336	PRKCZ	protein kinase C, zeta (PRKCZ)
353	SRP19	Signal recognition particle 19 kDa protein
401	TMEM139	transmembrane protein 139 (TMEM139)
420	AICDA	activation-induced cytidine deaminase (AICDA)
436	CXCL2	chemokine (C-X-C motif) ligand 2 (CXCL2)
447	ADARB1	adenosine deaminase, RNA-specific, B1 (RED1 homolog rat) (ADARB1), transcript variant 2
477	MAPKAPK5	mitogen-activated protein kinase-activated protein kinase 5 (MAPKAPK5), transcript variant 1
483	ACVR2A	Activin receptor type-2A
502	GTF2E2	general transcription factor IIIE, polypeptide 2, beta 34kDa (GTF2E2)
679	RNASEL	2-5A-dependent ribonuclease
787	HIST1H2AC	Histone H2A type 1-C
794	STK6	Serine/threonine-protein kinase 6
814	C9orf121	nucleoredoxin-like 2 (NXNL2)
895	C9orf116	chromosome 9 open reading frame 116 (C9orf116), transcript variant 2
916	EVI5L	ecotropic viral integration site 5-like (EVI5L)
944	PPIA	peptidylprolyl isomerase A (cyclophilin A) (PPIA)
947	ATP6V1D	ATPase, H ⁺ transporting, lysosomal 34kDa, V1 subunit D (ATP6V1D)
1061	ARHGAP9	Rho GTPase activating protein 9 (ARHGAP9), transcript variant 1
1098	RIOK2	RIO kinase 2 (yeast) (RIOK2)
1293	MGC5590	hypothetical protein MGC5590 (MGC5590)
1294	PSENEN	presenilin enhancer 2 homolog (C. elegans) (PSENEN)
1298	C20orf18	chromosome 20 open reading frame 18 (C20orf18), transcript variant 4
1314	PSG9	pregnancy specific beta-1-glycoprotein 9 (PSG9)
1335	TFPT	TCF3 fusion partner
1448	PLEKHG6	Pleckstrin homology domain-containing family G member 6
1544	PDPK1	3-phosphoinositide dependent protein kinase-1 (PDPK1), transcript variant 2
1548	NIP7	nuclear import 7 homolog (S. cerevisiae) (NIP7)
1579	C9orf114	Uncharacterized protein C9orf114
1613	PCGF1	polycomb group ring finger 1 (PCGF1)

Variable	Abreviatura génica	Proteína
1648	RALGPS2	Ral GEF with PH domain and SH3 binding motif 2 (RALGPS2), transcript variant 1
1656	FCHO1	FCH domain only 1 (FCHO1)
1853	ELP3	elongation protein 3 homolog (S. cerevisiae) (ELP3)
2042	S100A2	S100 calcium binding protein A2 (S100A2)
2083	C16orf13	chromosome 16 open reading frame 13 (C16orf13)
2425	FAM112A	family with sequence similarity 112, member A (FAM112A), transcript variant 1
2442	LOC344405	UPF0566 protein
2457	CIRH1A	cirrhosis, autosomal recessive 1A (cirhin) (CIRH1A)
2470	SKIP	skeletal muscle and kidney enriched inositol phosphatase (SKIP), transcript variant 1
2678	SLC39A7	solute carrier family 39 (zinc transporter), member 7 (SLC39A7)
3136	ARHGDIG	Rho GDP-dissociation inhibitor 3
3543	ALOXE3	Epidermis-type lipoxygenase 3
4023	COLEC10	Collectin-10
4574	PPP1R8	protein phosphatase 1, regulatory (inhibitor) subunit 8 (PPP1R8), transcript variant 1
4828	IRAK2	Interleukin-1 receptor-associated kinase-like 2
5075	IL20	interleukin 20 (IL20)
5313	EIF2B5	eukaryotic translation initiation factor 2B, subunit 5 epsilon, 82kDa (EIF2B5)
5638	C4orf27	chromosome 4 open reading frame 27 (C4orf27)
5739	BAG5	BCL2-associated athanogene 5 (BAG5)
6252	TADA3L	transcriptional adaptor 3 (NGG1 homolog, yeast)-like (TADA3L), transcript variant 2
6331	SFRS5	splicing factor, arginine/serine-rich 5 (SFRS5)
6342	NUDT16L1	nudix (nucleoside diphosphate linked moiety X)-type motif 16-like 1 (NUDT16L1)
6478	PSMC1	proteasome (prosome, macropain) 26S subunit, ATPase, 1 (PSMC1)
6883	UBE2S	Ubiquitin-conjugating enzyme E2 S
7438	RPL11	ribosomal protein L11 (RPL11)
8056	CCNB1IP1	cyclin B1 interacting protein 1 (CCNB1IP1), transcript variant 1
8059	CORO1C	coronin, actin binding protein, 1C (CORO1C), transcript variant 1

Lista de variables del *dataset* biomarcadores con el filtro reliefF y wrapper

Variable	Abreviatura génica	Proteína
1	LRRIQ2	leucine-rich repeats and IQ motif containing 2 (LRRIQ2)
2	C5orf3	chromosome 5 open reading frame 3 (C5orf3)
7	IL7	interleukin 7 (IL7)
13	BHMT2	betaine-homocysteine methyltransferase 2 (BHMT2)
15	COPZ1	coatomer protein complex, subunit zeta 1 (COPZ1)
20	ENTPD1	cDNA clone MGC:12418 IMAGE:3934658, complete cds
40	EFCBP1	EF-hand calcium binding protein 1 (EFCBP1)
61	MUSTN1	musculoskeletal, embryonic nuclear protein 1 (MUSTN1)
114	C3orf26	chromosome 3 open reading frame 26 (C3orf26)
136	EIF1	eukaryotic translation initiation factor 1 (EIF1)
363	GTF2I	General transcription factor II-I
3886	MARK4	MAP/microtubule affinity-regulating kinase 4
4079	MYOT	myotilin (MYOT)
4331	IGKC	immunoglobulin kappa constant (IGKC)
6332	RAB10	RAB10, member RAS oncogene family (RAB10)
6342	NUDT16L1	nudix (nucleoside diphosphate linked moiety X)-type motif 16-like 1 (NUDT16L1)
6812	IXL	mediator complex subunit 29 (MED29)
7506	CCR6	C-C chemokine receptor type 6
8714	ABAT	4-aminobutyrate aminotransferase (ABAT)
9329	SFXN3	sideroflexin 3 (SFXN3)
9442	FLJ32658	hypothetical protein FLJ32658 (FLJ32658)
9445	RETN	resistin (RETN)

Lista de variables del dataset biomarcadores para la predicción de MMSE

Variable	Abreviatura génica	Proteína
9217	EIF2AK2	Interferon-induced, double-stranded RNA-activated protein kinase
8050	ECH1	enoyl Coenzyme A hydratase 1, peroxisomal (ECH1)
8586	PRELP	proline/arginine-rich end leucine-rich repeat protein (PRELP), transcript variant 1
8172	NUDT2	nudix (nucleoside diphosphate linked moiety X)-type motif 2 (NUDT2), transcript variant 1
8838	GATA3	Trans-acting T-cell-specific transcription factor GATA-3
8738	COPS2	COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis) (COPS2)
5	KLF3	Kruppel-like factor 3 (basic) (KLF3)
7816	CKAP2	cytoskeleton associated protein 2 (CKAP2)
7268	LOC100129827	cDNA clone MGC:40177 IMAGE:5167345, complete cds
9026	C16orf67	hypothetical protein MGC3020 (MGC3020)
1335	TFPT	TCF3 fusion partner
5395	LOC283400	PREDICTED: Homo sapiens hypothetical protein LOC283400 (LOC283400)
114	C3orf26	chromosome 3 open reading frame 26 (C3orf26)
8222	OSBPL10	Oxysterol-binding protein-related protein 10
8339	DLX6	distal-less homeobox 6 (DLX6)
7227	IPO9	Importin-9
8322	CPB1	carboxypeptidase B1 (tissue) (CPB1)
7854	PCSK2	proprotein convertase subtilisin/kexin type 2 (PCSK2)
7903	SMOX	Spermine oxidase
8086	PLEKHG5	pleckstrin homology domain containing, family G (with RhoGef domain) member 5 (PLEKHG5), transcript variant 1
767	PUF60	Poly(U)-binding-splicing factor PUF60
8065	REEP5	Receptor expression-enhancing protein 5

Variable	Abreviatura génica	Proteína
6150	PIK3R5	phosphoinositide-3-kinase, regulatory subunit 5, p101 (PIK3R5)
8399	GK5	hypothetical protein MGC40579 (MGC40579)
8915	SPTLC1	serine palmitoyltransferase, long chain base subunit 1 (SPTLC1), transcript variant 2
8965	NBPF1	neuroblastoma breakpoint family, member 1 (NBPF1)