



**POLITÉCNICA**



UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y  
DISEÑO INDUSTRIAL

Grado en Ingeniería Electrónica y Automática Industrial

**TRABAJO FIN DE GRADO**

IDENTIFICACIÓN DE DIFERENTES  
TIPOS DE DEMENCIA MEDIANTE EL  
USO DE MODELOS DE PROGRESIÓN DE  
ENFERMEDAD Y MARCADORES DE  
PATOLOGÍA AMILOIDE, TAU Y  
NEURODEGENERACIÓN

Jorge Sepúlveda Cárcamo

*Tutor:* Carlos Platero Dueñas

*Departamento:* Ingeniería  
Eléctrica, Electrónica,

Automática y Física Aplicada



**POLITÉCNICA**



UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y  
DISEÑO INDUSTRIAL

Grado en Ingeniería Electrónica y Automática Industrial

**TRABAJO FIN DE GRADO**

IDENTIFICACIÓN DE DIFERENTES  
TIPOS DE DEMENCIA MEDIANTE EL  
USO DE MODELOS DE PROGRESIÓN DE  
ENFERMEDAD Y PERFILES  $AT(N)$

Firma Autor

*Firma Cotutor (si lo hay)*

*Firma Tutor*

Copyright ©2025. Jorge Sepúlveda Cárcamo

Esta obra está licenciada bajo la licencia Creative Commons

Atribución-NoComercial-SinDerivadas 3.0 Unported (CC BY-NC-ND 3.0). Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-nd/3.0/deed.es> o envíe una carta a Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, EE.UU.

Todas las opiniones aquí expresadas son del autor, y no reflejan necesariamente las opiniones de la Universidad Politécnica de Madrid.



**Título:** Identificación de diferentes tipos de demencia mediante el uso de modelos de progresión de enfermedad y perfiles AT(N)

**Autor:** Jorge Sepúlveda Cárcamo

**Tutor:** Carlos Platero Dueña

## EL TRIBUNAL

Presidente:

Vocal:

Secretario:

Realizado el acto de defensa y lectura del Trabajo Fin de Grado el día ..... de ..... de ... en ....., en la Escuela Técnica Superior de Ingeniería y Diseño Industrial de la Universidad Politécnica de Madrid, acuerda otorgarle la CALIFICACIÓN de:

VOCAL

SECRETARIO

PRESIDENTE



# Agradecimientos

Quiero agradecer en primer lugar a mi familia su apoyo, a lo largo de mi vida. A mi padre, que intentaba buscarme las cosquillas mientras estudia y escribía este documento, sacándome una sonrisa en el proceso. A mi madre, que siempre ha insistido en comprar ella cualquier cosa que estuviera mínimamente relacionada con los estudios, para que nunca me faltase nada. Y a mi hermana, a quién espero poder ayudar a revisar su propio trabajo de fin de grado el año que viene, aunque no tenga nada que ver con mis estudios.

Quiero agradecer también a mis amigos, tanto a los que ya tenía como los que he hecho en la universidad, los buenos ratos que he pasado con ellos, y que me han ayudado a olvidarme de la presión de los estudios durante mis cuatro años y medio de carrera.

Por último, quiero agradecer a mi tutor, Carlos Platero, por su apoyo como tutor durante la elaboración y redacción de este TFG. Por su constante implicación, indicando siempre donde podía mejorar. Por su paciencia cuando he tenido que posponer la presentación debido al suspenso de una asignatura. Este TFG no tendría el mismo resultado sin su apoyo.



# Resumen

- **Contexto:** La enfermedad de Alzheimer (*Alzheimer's Disease*, AD) es una enfermedad neurodegenerativa. El diagnóstico temprano es clave a la hora de mantener la calidad de vida de quienes la padecen. Debido a esto, es necesario estudiar el avance de la enfermedad en su fase prodrómica, así como el impacto de otras copatologías en el avance de la enfermedad.
- **Objetivos:** Identificar grupos diferentes de demencia mediante modelos de progresión de enfermedad (*Disease Progression Models*, DPM) y perfiles AT(N) de los pacientes. Valorar la idoneidad de Leaspy para construir modelos que permitan evaluar si hay diferencias entre los grupos.
- **Métodos:** Para el estudio se cuentan con 551 sMCI (MCI estable), 425 pMCI (MCI progresivo) y pacientes con 358 demencia. Todos ellos obtenidos de la base de datos de ADNI. Se realizará un estudio transversal en la primera visita y un estudio longitudinal mediante modelos DPM. Se evaluará la capacidad de leaspy de modelar la enfermedad, su robustez, su capacidad de definir la progresión de los marcadores y su capacidad de elaborar predicciones sobre el futuro del paciente.
- **Resultados:** En el estudio transversal se observa patología amiloide (A+) presente en el 89,3% de los pacientes de demencia. Además, entre el 2,8% y el 7,9% de los pacientes totales pertenecen al perfil A-T+. El mejor modelo elaborado por leaspy que combina MMSE y CDRSB, tiene una tasa de detección de pMCI del 78,04%. El modelo demuestra una buena robustez. Define correctamente las trayectorias de los marcadores, y sus cortes con el inicio de la enfermedad. En sus predicciones, puede detectar entre un 30% y un 60% de las conversiones. Al predecir la fecha exacta de la predicción tiende a fallar (correlación  $< 0,5$ ).
- **Conclusión:** Existe un grupo reducido de pacientes con demencia que pueden no padecer AD. Leaspy demuestra ser similar a otras herramientas similares como RPDPM, ofreciendo resultados ligeramente mejores en la detección de la demencia, pero peores en la de los pMCI y estimación del tiempo de conversión en la predicción. Además, el mejor modelo elaborado en leaspy requiere de menos marcadores, facilitando su implementación en ámbitos clínicos y demuestra una mayor robustez. Se demuestra también que se puede elaborar una historia natural de los pacientes mediante el uso de Leaspy. Leaspy se podría utilizar para estudiar si hay diferencias entre los grupos identificados previamente.

**Palabras clave:** Alzheimer, prodrómica, perfiles AT(N), Leaspy, DPM, Demencia, MCI.

# Abstract

- **Context:** Alzheimer’s disease (AD) is a neurodegenerative disorder. Early diagnosis is key to maintaining the quality of life of those affected. Because of this, it is necessary to study the progression of the disease in its prodromic phase, from mild cognitive impairment (MCI) to dementia, as well as the impact of other comorbidities on disease progression.
- **Objectives:** To identify different dementia groups using disease progression models (DPM) and patients’ AT(N) profiles. To evaluate the suitability of Leaspy for building models that can assess whether there are differences between these groups.
- **Methods:** The study includes 551 sMCI (stable MCI), 425 pMCI (progressive MCI), and 358 dementia patients, all obtained from the ADNI database. A cross-sectional study will be conducted at the first visit and a longitudinal study using DPM models. The ability of Leaspy to model the disease, its robustness, its capacity to define biomarker trajectories, and its ability to make predictions about the patient’s future will be evaluated.
- **Results:** The cross-sectional study shows amyloid pathology (A+) present in 89.3% of dementia patients. Additionally, between 2.8% and 7.9% of all patients belong to the A-T+ profile. The best model created by Leaspy, combining MMSE and CDRSB, has a pMCI detection rate of 78.04%. The model demonstrates good robustness. It correctly defines biomarker trajectories, and its cutoffs for disease onset approximately match those used in clinical diagnosis. In its predictions, it can detect between 30% and 60% of conversions. When predicting the exact conversion date, it tends to fail (correlation < 0,5).
- **Conclusion:** There is a small group of dementia patients who may not suffer from AD. Regarding the use of Leaspy as a tool for modeling and evaluation, Leaspy proves similar to other tools like RPDPM, offering slightly better results in dementia detection, but worse in pMCI detection and conversion time estimation in prediction. Moreover, the best model developed in Leaspy requires fewer biomarkers, facilitating its implementation in clinical settings, and shows greater robustness. It is also demonstrated that a natural history of patients can be built using Leaspy. In summary, Leaspy could be used to study whether there are differences between the previously identified groups.

**Keywords:** Alzheimer, prodromal, AT(N) profiles, Leaspy, DPM.



# Índice general

|   |           |
|---|-----------|
| Agradecimientos   | IX        |
| Resumen   | XI        |
| Abstract  | XIII      |
| Índice  | XVII      |
| <b>1. Introducción</b>  | <b>1</b>  |
| 1.1. Motivación del proyecto  | 1         |
| 1.1.1. Impacto socio-económico  | 2         |
| 1.1.2. Impacto Económico  | 2         |
| 1.2. Objetivos  | 3         |
| 1.3. Estructura del Documento   | 4         |
| <b>2. Estado del arte</b>   | <b>5</b>  |
| 2.1. Enfermedad de Alzheimer  | 5         |
| 2.1.1. Diagnóstico de AD mediante el uso de biomarcadores. Perfiles AT(N) | 8         |
| 2.2. Modelos de progresión de AD  | 9         |
| 2.2.1. Modelos Lineales de Efecto Mixto                                   | 10        |
| 2.2.2. Modelos de Progresión de Enfermedad                                | 12        |
| 2.2.3. Construcción de los modelos - Leaspy                               | 15        |
| 2.2.4. RPDPM  | 20        |
| 2.3. Aportación al Trabajo  | 21        |
| 2.3.1. Perfiles AT(N)   | 21        |
| 2.3.2. Modelos de progresión de enfermedad                                | 21        |
| <b>3. Materiales</b>  | <b>23</b> |
| 3.1. Python   | 23        |
| 3.1.1. Leaspy   | 23        |
| 3.2. Matlab   | 23        |
| 3.2.1. RPDPM  | 23        |
| 3.3. ADNImerge  | 23        |
| 3.4. Magerit  | 24        |
| 3.5. Comunicación con Curie y Magerit                                     | 25        |
| 3.5.1. PuTTY  | 25        |
| 3.5.2. Filezilla  | 25        |
| 3.5.3. Xming  | 25        |

|   |           |
|---|-----------|
| 3.5.4. Tmux . . . . .   | 25        |
| 3.6. Marcadores . . . . .   | 26        |
| 3.6.1. Marcadores neuropsicológicos (NM) . . . . .                          | 26        |
| 3.6.2. Marcadores de líquido cefalorraquídeo (CSF) . . . . .                | 27        |
| <b>4. Métodos</b>   | <b>29</b> |
| 4.1. Estudio Transversal . . . . .  | 29        |
| 4.1.1. Estudio Transversal de cada individuo en su primera visita . . . . . | 29        |
| 4.1.2. Perfiles AT(N) . . . . .   | 30        |
| 4.1.3. Tiempo de Conversión y Tiempo de Censura . . . . .                   | 30        |
| 4.2. Modelos de Progresión de Enfermedad: Fase Prodrómica . . . . .         | 31        |
| 4.2.1. Paralelización de Tareas . . . . .                                   | 31        |
| 4.2.2. Leaspy . . . . .   | 32        |
| 4.2.3. Construcción del modelo . . . . .                                    | 33        |
| 4.2.4. Evaluación de los Modelos . . . . .                                  | 34        |
| 4.3. Comprobación de robustez ante datos faltantes (NaN) . . . . .          | 34        |
| 4.4. Historia Natural . . . . .   | 35        |
| 4.5. Pronostico . . . . .   | 36        |
| <b>5. Resultados</b>  | <b>37</b> |
| 5.1. Estudio Transversal . . . . .  | 37        |
| 5.1.1. Estudio transversal de cada individuo en su primera visita . . . . . | 37        |
| 5.1.2. Perfiles AT(N) . . . . .   | 37        |
| 5.1.3. Tiempo de conversión y Tiempo de Censura . . . . .                   | 37        |
| 5.2. Modelos Población General . . . . .                                    | 40        |
| 5.3. Comprobación de robustez ante datos faltantes (NaN) . . . . .          | 41        |
| 5.4. Historia Natural . . . . .   | 42        |
| 5.5. Pronostico . . . . .   | 45        |
| <b>6. Discusión</b>   | <b>49</b> |
| 6.1. Estudio Transversal . . . . .  | 49        |
| 6.2. Modelos Población General . . . . .                                    | 49        |
| 6.3. Comprobación de robustez ante datos faltantes (NaN) . . . . .          | 50        |
| 6.4. Historia Natural . . . . .   | 50        |
| 6.5. Pronostico . . . . .   | 51        |
| <b>7. Conclusiones</b>  | <b>53</b> |
| 7.1. Conclusión . . . . .   | 53        |
| 7.2. Desarrollos futuros . . . . .  | 53        |
| <b>A. Anexo: Geometría de Riemann [24]</b>                                  | <b>55</b> |
| A.1. Variedades . . . . .   | 55        |
| A.2. Métricas y Variedades Riemannianas . . . . .                           | 55        |
| A.3. Geodiscos . . . . .  | 56        |
| A.4. Mapeado Exponencial . . . . .  | 56        |
| A.5. Transporte Paralelo . . . . .  | 56        |

|   |           |
|---|-----------|
| <b>B. Anexo: Contraste de Hipótesis. P-valor. ANOVA. Chi-cuadrado</b> |           |
| [43]  | <b>57</b> |
| B.1. Contraste de Hipótesis y P-valor . . . . .                       | 57        |
| B.2. Análisis de Varianza (ANOVA) . . . . .                           | 58        |
| B.3. Chi Cuadrado . . . . .   | 59        |
| <b>C. Anexo: Clasificadores KDE-Bayes</b>                             | <b>61</b> |
| C.1. Clasificadores de Bayes . . . . .                                | 61        |
| C.2. Estimación de Densidad Kernel (KDE) . . . . .                    | 62        |
| <b>D. Anexo: Estructura de Carpetas</b>                               | <b>63</b> |
| <b>E. Anexo: Informe de Turnitin</b>                                  | <b>69</b> |
| <b>Bibliografía</b>   | <b>71</b> |



# Índice de figuras

|  |    |
|--|----|
| 1.1. Tasa relativa de población con Alzheimer en centros de mayores [18] . . . . .   | 2  |
| 2.1. Ejemplos de Posibles Modelos. [24] . . . . .  | 15 |
| 3.1. Logo de Leaspy[25] . . . . .  | 23 |
| 3.2. Logo de ADNI[2] . . . . .   | 24 |
| 3.3. Logo del CeSViMa[4] . . . . .   | 24 |
| 3.4. Logo de Filezilla[23] . . . . .   | 25 |
| 3.5. Logo de Xming[14] . . . . .   | 25 |
| 3.6. Logo de tmux[28] . . . . .  | 26 |
| 5.1. Comparación de tiempos de censura y conversión en diferentes poblaciones . . . . .  | 40 |
| 5.2. Trayectorias de los marcadores - Lespy - train+test . . . . .   | 42 |
| 5.3. Trayectorias de los marcadores - RPDPM - train+test . . . . .   | 43 |
| 5.4. Reserva cognitiva de individuos sMCI, pMCI y Demencia - train+test - Leaspy . . . . .   | 43 |
| 5.5. Reserva cognitiva de individuos sMCI, pMCI y Demencia - train+test - RPDPM . . . . .  | 44 |
| 5.6. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - train . . . . .                               | 45 |
| 5.7. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - test . . . . .                                | 45 |
| 5.8. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - Todos los pacientes . . . . .                 | 46 |
| 5.9. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Train . . . . .                                     | 46 |
| 5.10. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Test . . . . .                                     | 47 |
| 5.11. Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Medidas de Calidad - Todos los pacientes . . . . . | 47 |
| 5.12. Pronostico, experimento de detección de conversión - Medidas de Calidad - Todos los pacientes . . . . .                                | 48 |
| E.1. Informe de similitud de Turnitin[42] . . . . .  | 69 |



# Índice de tablas

|      |   |    |
|------|---|----|
| 5.3. | Perfiles AT de los pacientes con muestras CSF, clasificados en sM-CI, pMCI y demencia. Se valoran dos criterios diferentes para A+: $pTau/A\beta_{42} > 0,028$ y $A\beta_{42} < 880$ . En los dos casos se sigue el criterio $pTau > 27pg/ml$ para valorar T+. [37] | 37 |
| 5.1. | Estudio Transversal - Todos los sujetos   | 38 |
| 5.2. | Estudio Transversal - Solo sujetos con CSF  | 39 |
| 5.4. | Comparación de las mejores combinaciones de marcadores  | 41 |
| 5.5. | Comparación del MAE de las mejores combinaciones de marcadores. Resultados de los datos de test   | 41 |
| 5.6. | Comparación de la robustez a datos NaN añadidos de la mejor combinación de marcadores para los modelos realizados en leaspy y un modelo de referencia en RPDPM  | 42 |
| 5.7. | Comparación del MAE Para diferentes valores de NaN. Mejor modelo de Leaspy y mejor modelo de RPDPM. Resultados de los datos de test   | 42 |
| 5.8. | Comparación de las mejores combinaciones de marcadores  | 43 |
| 5.9. | Evaluación del corte al inicio de la enfermedad para cada marcador. Comparación entre Leaspy y la medida real   | 44 |



# Capítulo 1

## Introducción

### 1.1. Motivación del proyecto

La enfermedad de Alzheimer (*Alzheimer's Disease*, AD) es una de las enfermedades neurodegenerativas más conocidas, así como una de las más comunes. Se estima que un 30 % de las personas mayores de 85 años padecen AD [13]. Además, según datos del Instituto Nacional de Estadística (INE), la esperanza de vida al nacimiento de la población española fue de aproximadamente 82 años en 2021 [16] y la población mayor de 70 años supera los 7,4 millones de personas [17]. Con todo esto, queda patente que la AD se encuentra muy presente en la sociedad actual y en la vida de una parte importante de la población.

Uno de los pasos más importantes al enfrentar el reto que suponen las enfermedades neurodegenerativas es la detección temprana de la enfermedad. En el caso del Alzheimer, se define una fase prodrómica llamada deterioro cognitivo leve (*Mild Cognitive Impairment*, MCI) como paso previo a la demencia [13, 7]. La correcta identificación del MCI, así como la predicción de su evolución a demencia podría abrir nuevas vías en la investigación de terapias y tratamientos para retrasar la demencia, mejorando enormemente la calidad de vida de los pacientes.

Además, actualmente el diagnóstico de AD se realiza en base a marcadores neuropsicológicos (*Neuropsychological Markers*, NM) que estudian los síntomas asociados a la enfermedad, y no la enfermedad en si misma. Esto lleva a que solo se diagnostique como posible o probable en pacientes vivos con demencia amnésica y, en ocasiones, sin diferenciar adecuadamente la causa tras la demencia. Afortunadamente, el Instituto Nacional sobre el Envejecimiento y la Asociación de Alzheimer (NIA-AA) de EEUU propone diferentes perfiles basados en biomarcadores que permiten estudiar la presencia de AD y todas las causas tras la demencia [21, 20]. Teniendo esto en cuenta, se observa la importancia de elaborar modelos y perfiles para cada paciente que puedan identificar adecuadamente el origen del MCI o la demencia y permitan seleccionar el tratamientos efectivos para su caso clínico concreto.

Por estas razones, a lo largo del siguiente trabajo se realizará un estudio longitudinal de los individuos facilitados por *Alzheimer's Disease Neuroimaging Initiative* (ADNI) con la ayuda de herramientas como leaspy [24]. Además, se elaborarán los perfiles de amiloide, tau y neurodegeneración (AT(N)) como propone la NIA [21, 20].

Juntando estos modelos y perfiles se buscará clasificar los pacientes por su tipo de demencia y estudiar como es la progresión de la enfermedad para cada caso, lo que podría abrir nuevas vías en la personalización de los tratamientos par cada paciente.

### 1.1.1. Impacto socio-económico

#### Impacto Social

La enfermedad de Alzheimer exige una atención continua al paciente una vez alcanza la demencia. Es a menudo la familia la que debe supervisar al paciente, suministrarle la medicación o incluso tomar decisiones legales, todo esto con un conocimiento imperfecto e incompleto sobre la enfermedad [13]. Además, los síntomas habituales de la demencia, no solo los físicos, sino también los psicológicos pueden impactar en los cuidadores y familiares. Estos a menudo ven como su ser querido va perdiendo la memoria, encontrándose más cansado y desorientado, o comportándose de forma más apática o agresiva con ellos; afectando de esta manera al estado anímico y bienestar psicológico de quienes rodean al paciente [13, 10]. Además, se trata de una enfermedad relativamente común, representando alrededor del 14 % de la población residente en centros de mayores [18], y siendo la cuarta causa de muerte más común [19].

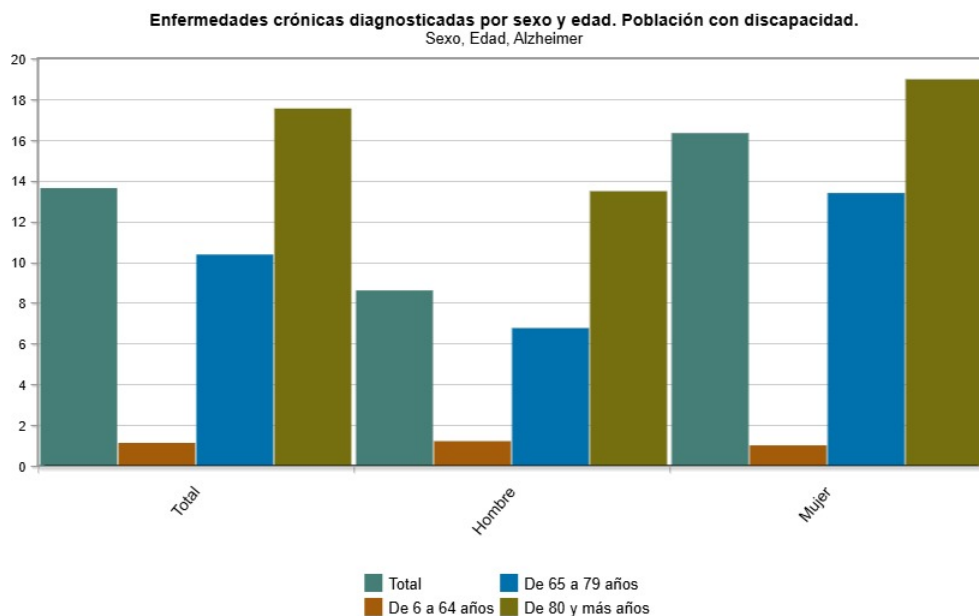


Figura 1.1: Tasa relativa de población con Alzheimer en centros de mayores [18]

### 1.1.2. Impacto Económico

En cuanto a los costes económicos asociados a la demencia, en 2015 se estimaban en 645 mil millones de dólares estadounidenses, mayores que los asociados al cáncer o a las enfermedades de corazón [34, 38]. La mayor parte de los costes del tratamiento recaen en la sanidad pública de los estados, aunque también suponen un gran gasto para familias y cuidadores [38].

Si bien estimar los costes totales de la demencia no es sencillo, unicamente centrándose en los costes en sanidad, se observa que en países como Países Bajos o Alemania ronda entre el 3% y el 6% del gasto total, incluyendo cuidados a largo plazo [34]. Una vez más, si bien es difícil cuantificar el resto de los gastos, se estima que los gastos médicos directos se corresponden con el 20% de los gastos totales de demencia, mientras que los cuidados a largo plazo representan el 40%. El resto se corresponde con el impacto en las familias y comunidades. [5].

Además, la importancia e impacto de esta enfermedad unicamente continuará creciendo dado al creciente envejecimiento de la población. En España, la esperanza de vida al nacimiento de la población fue de aproximadamente 82 años en 2021 [16] y la población mayor de 70 años supera ya los 7,4 millones de personas [17], algo especialmente relevante para una enfermedad cuyo principal factor de riesgo es la edad [13].

Por último, centrandó la atención únicamente en la población Española que sufre la enfermedad, se estima que su incidencia es de un 1.07% para la población entre 65 y 69 años. Subiendo conforme aumenta la edad a un 3,4% entre 70 y 74; un 6,9% entre 75 y 79; un 12,1% entre 80 y 84; un 20,1% entre 85 y 89; y un 39,2% a mayores de 90 [31]. Además, las cifras del pasado año 2024 de causas de mortalidad atribuyen específicamente al Alzheimer aproximadamente 13.000 defunciones del total de alrededor de 430.000.[1]

## 1.2. Objetivos

Por las razones detalladas anteriormente, se buscará identificar grupos diferentes de demencia a partir de los datos de la muestra de la cohorte de ADNI. Se elaboraran para ello modelos DPM de los sujetos que se encuentren en la fase prodrómica empleando leaspy. Se elaboraran también sus perfiles AT(N) y se valorarán junto al modelo para realizar la clasificación. Para ello se seguirán los siguientes pasos:

- **Estudio transversal:** Se estudiarán los datos de los pacientes de la cohorte de ADNI que se utilizarán para elaborar los modelos en su primera visita. Se valorarán los pacientes MCI y Demencia con más de una visita tras descartar las visitas cognitivamente normales (CN).
- **Elaboración de perfiles AT(N) en línea base:** Se estudiarán los perfiles AT(N) de los pacientes en su primera visita, lo que permitirá realizar una primera valoración sobre el impacto de los marcadores en líquido cefalorraquídeo (CSF) en la aparición de demencia.
- **Elaboración de DPM con Leaspy:** Se utilizará leaspy para construir varios DPM de la población general y personalizarlo para cada individuo, como se propone en la tesis de Koval [24]. Se combinará con un clasificador bayesiano que permita identificar mediante los modelos que individuos pertenecen a MCI o a demencia.

### 1.3. Estructura del Documento

- **Introducción:** Este capítulo contiene los objetivos del proyecto, así como el contexto en el que se desarrolla.
- **Estado del arte:** Este capítulo contiene un resumen de los artículos, tesis y teorías que sirven como bases para el desarrollo del proyecto.
- **Materiales:** Este capítulo contiene información sobre los materiales y herramientas de software empleadas para la realización del proyecto.
- **Métodos:** Detalla la metodología y procedimientos seguidos durante la realización del proyecto.
- **Resultados:** Expone los resultados del estudio y centra la atención en los hallazgos más relevantes.
- **Discusión:** Análisis de los resultados del proyecto
- **Apéndices:** Esta sección contiene información adicional para apoyar lo desarrollado en este documento.

## Capítulo 2

# Estado del arte

### 2.1. Enfermedad de Alzheimer

La enfermedad de Alzheimer es la enfermedad neurodegenerativa más común. La enfermedad lentamente destruye la memoria y aptitudes del pensamiento hasta volver al paciente incapaz de realizar hasta las tareas más simples. Afecta al 8% de la población superior a 65, aunque la incidencia aumenta mucho con la edad. Hay registrado entre un 1% y un 2% de casos de Alzheimer en la población de entre 65 y 70 años, y su incidencia se dobla cada 5 años. En edades superiores a 85 años la incidencia es del 30% aproximadamente. La supervivencia media de un paciente es de entre 8 y 12 años, pero se han registrado casos de más de 20 años. La supervivencia depende principalmente de los cuidados y terapias suministradas al paciente por sus cuidadores. [13]

El principal factor de riesgo de esta enfermedad es la edad. Además, a edades más avanzada hay registrados más casos de Alzheimer entre la población de genero femenino. Niveles bajos de escolarización también aumentan el riesgo de padecer la enfermedad. La hipótesis más aceptada es que la actividad intelectual aumenta la reserva cognitiva. También son especialmente importantes los factores de riesgo vascular. Hipertensión arterial o colesterol alto en las edades medias aumentan el riesgo de desarrollar Alzheimer o demencia. De manera similar, tensión o colesterol bajos en edades avanzadas también aumentan el riesgo. Un mal control de la diabetes tipo II, llevando a hipoglucemia también aumenta el riesgo. La anemia y alteraciones en hormonas estimulantes del tiroides y hormonas sexuales aumentan también el riesgo de Alzheimer. Un caso especial es el de la depresión. La depresión aumenta en dos veces el riesgo de sufrir la enfermedad. Sin embargo, la depresión también puede ser un síntoma precoz de la demencia. Por último, también se consideran factores de riesgo ser fumador, la obesidad y el infrapeso. Se cree que condiciones ambientales y traumatismos también pueden influir. Para acabar, el consumo de café reduce el riesgo ligeramente. Hay que mencionar también que una parte muy pequeña de los casos tiene un componente genético. [13]

En un paciente con Alzheimer se observa una disminución de volumen. La atrofia se concentra en el lóbulo temporal, el hipocampo y la amígdala. La corteza cerebral se vuelve más delgada, y presenta circunvoluciones atróficas. La degeneración la sufren tanto la sustancia gris como la blanca, ya que esta última contiene los axones de

las neuronas dañadas. Con el desarrollo de la enfermedad se va reduciendo el número de sinapsis. Los procesos de destrucción neuronal y de sinapsis son progresivos. Además, se cree que son los responsables de los síntomas clínicos. En el cerebro del enfermo de Alzheimer se pueden encontrar depósitos de  $A\beta$  amiloide y de proteína tau. Los depósitos de  $A\beta$  amiloide aparecen como placas extracelulares. También se encuentran depósitos a nivel vascular, aumentando la fragilidad de los vasos y el riesgo de hemorragia cerebral. La proteína tau se encuentra en neuritas distróficas alrededor de la amiloide, y en neuronas con degeneración neurofibrilar. Alrededor de las placas de  $A\beta$  amiloide se observa neuroinflamación y daños por oxidación. Se observa también procesos de neurogénesis en áreas cercanas a las placas. Los sistemas de neurotransmisión también se ven afectados por la enfermedad. [13]

Antes del desarrollo de la demencia, el paciente pasa por un estado habitualmente llamado deterioro cognitivo leve. Entre el 10 % y el 15 % de los pacientes con deterioro cognitivo leve desarrollan demencia al cabo de un año. Se considera que hay deterioro cognitivo leve cuando se cumplen varias condiciones.

1. El paciente no debe tener demencia
2. El paciente se queja perdidas de memoria
3. El paciente se queja de un declive en la función cognitiva en comparación a un año antes
4. Se observa un declive en memoria u otras funciones cognitivas
5. El deterioro no tiene consecuencias en su vida diaria

La principal queja entre pacientes con deterioro cognitivo leve son perdidas de memoria. También se puede observar problemas al encontrar las palabras, desorientación y problemas de planificación. Se puede observar también depresión, irritabilidad, ansiedad, agresión o apatía. Mediante técnicas de neuroimagen, test neuropsicológicos y biomarcadores podemos identificar los pacientes con mayor riesgo de conversión a demencia. Es importante realizar un seguimientos para identificar la posible progresión a Alzheimer. [13]

Durante el desarrollo de la demencia se observan varios síntomas. El paciente pierde progresivamente su capacidad de interacción con el medio. Puede sufrir síntomas conductuales, inmovilidad, trastornos de lenguaje, trastornos de alimentación, incontinencia e infecciones. Inicialmente, se producen perdidas de memoria esporádica. A este se le añaden déficits semánticos, alteraciones en las funciones ejecutivas, modificaciones en los niveles de atención, modificaciones del lenguaje, apraxia visoconstructiva y agnosia. La apraxia visoconstructiva es la perdida de la capacidad para llevar a cabo tareas de ensamblaje y copia de dibujos. La agnosia es la incapacidad para recordar información que llegue desde los sentidos. Además de esto, el paciente sufre de perdidas de memoria, que dificultan su vida diaria. En cuanto a los trastornos del lenguaje alteraciones en la memoria semántica causan afasia anómica fluente, es decir, incapacidad de denominar las cosas por su nombre. El paciente acaba sufriendo desorientación, primero temporal, después espacial, y por último personal. Los hábitos de trabajo y la capacidad de organización se ven alterados. Pierde la capacidad de realizar varias tareas simultáneamente, el juicio

crítico, el razonamiento, el pensamiento abstracto y la iniciativa. Se encuentra también acalculia y apraxia ideomotora (capacidad de realizar el movimiento adecuado ante una orden). [13]

Acompañando a los síntomas físicos se encuentran síntomas psicológicos y conductuales. El más frecuente es la apatía. La apatía es un estado de desinterés y falta de motivación ante estímulos externos. También se puede observar agitación, agresividad e irritabilidad. Frecuentemente sufren de alteraciones de percepción, destacando delirios de robos e infidelidad, que generan situaciones de agresividad. Pueden sufrir alucinaciones visuales, insomnio y síndrome crepuscular. Además, tienden a la deambulación errática. Se pueden usar los síntomas conductuales como marcador del progreso de la enfermedad. [13]

El grado de alteración funcional depende de la combinación entre síntomas físicos y conductuales. En estados más leves, se altera la capacidad de realizar tareas complejas. Posteriormente la capacidad de realizar actividades instrumentales se ve afectada. Por último, los pacientes pierden la capacidad de autocuidado. Con el objetivo de poder adaptar las terapias al estado del paciente, existen diversas escalas que miden el deterioro. Se destaca la General Deterioration Scale de Reisberg (GDS). [13]

El proceso de diagnóstico actual sigue siendo clínico, sin embargo, los doctores pueden apoyarse en biomarcadores y técnicas de neuroimagen. Los biomarcadores principales son los niveles de  $A\beta$  amiloide y tau en líquido cefalorraquídeo. Para ello se emplea la punción lumbar. En cuanto a la neuroimagen, destacan la resonancia magnética nuclear (RMN), tomografía por emisión de fotón único (SPECT), la tomografía por emisión de positrones con fluorodeoxiglucosa (PET-FDG) y magnetoencefalografías (MEG). La RMN se puede emplear para cuantificar la atrofia, aunque no permite diferenciar las demencias. La SPECT permite conocer el riesgo bascular en diferentes áreas cerebrales. La PET permite encontrar hipometabolismo de glucosa, así como cuantificar el tamaño de depósitos de  $A\beta$  amiloide y tau. Por último la magnetoencefalografía permite valorar el estado del cerebro en diferentes fases de la enfermedad. [13]

En cuanto a los tratamientos, incluyen intervención farmacológica y terapéutica. Los objetivos de la intervención farmacológica son paliar la degeneración neuronal, potenciar la actividad de las neuronas no afectadas, contrarrestar el déficit de neurotransmisores y neutralizar o aliviar los síntomas cognitivos y no cognitivos. Para paliar la degeneración neuronal se puede buscar disminuir la producción de  $\beta$  amiloide y tau, bajar la neuroinflamación, reducir el estrés oxidativo o aumentar el número de sinapsis. Con estos objetivos, el tratamiento emplea dos grupos de fármacos. Los inhibidores de acetilcolinesterasa (IACE) mejoran los niveles del neurotransmisor colinérgico en las sinapsis. Se suministra en cuatro fases. En primer lugar es necesario asegurar que el paciente este en un grado de deterioro leve o moderado. Después se aumenta la dosis hasta que el paciente alcanza la dosis máxima que puede tolerar. Tras esto se debe hacer un seguimiento del efecto del fármaco en el paciente. Por último, se retira cuando deja de ser eficaz. El otro fármaco habitual son los antagonistas de los receptores N-Metil-D-Aspartato tipo memantina, que buscan bloquear los daños causados por la disregulación del glutamato. Otra vía de investigación

farmacológica abierta son los dirigidos a la limpieza de  $A\beta$  amiloide y tau, aunque a día de hoy es solo una línea de investigación. [13]

En cuanto a las terapias, buscan tanto el beneficio de los pacientes como de sus cuidadores. En primer lugar, hay que mencionar los programas de orientación a la realidad. Se administran en sesiones grupales. En ellos se le aporta al paciente información sobre su entorno. Mejora aspectos cognitivos, conductuales y funcionales. Otra terapia relevante es la terapia de reminiscencia. La terapia de reminiscencia busca emplear recuerdos del paciente, junto con fotos o vídeos suyos, para mejorar su conducta, afectividad y rendimiento cognitivo. En tercer lugar, encontramos la terapia validación, que busca ayudar al paciente a aceptar y conocer la situación en la que se encuentra. También es habitual la estimulación cognitiva. En la estimulación cognitiva se trabajan funciones concretas, como la memoria o la atención. Para ello se busca el recuerdo de algún hecho o actividades de la vida diaria. La última terapia que se menciona es la estimulación magnética transcraneal repetitiva. En ella se provocan corrientes eléctricas en partes concretas del cerebro para aumentar o disminuir la excitabilidad de las neuronas. No hay evidencias de que sea realmente efectivo. Por último, además de terapias para el paciente es importante mejorar la formación de los cuidadores encargados del paciente. [13]

### 2.1.1. Diagnóstico de AD mediante el uso de biomarcadores. Perfiles AT(N)

Sabiendo esto, para poder estudiar adecuadamente la enfermedad conviene recordar que muchas veces, el termino AD se usa para referirse tanto a los cambios producidos por el Alzheimer como para cambios neuropatológicos con síntomas compatibles con Alzheimer. Dado que esta doble definición de la AD puede llevar a confusión en ámbitos de investigación, falsos positivos en ámbitos médicos y opaca la presencia de la fase preclínica, el NIA ya propone en su informe de 2018 la redefinición de AD de forma independiente de sus síntomas clínicos [21].

Desvincular la enfermedad de su sintomatología lleva a la necesidad de establecer nuevos marcadores separados de los neuropsicológicos para su diagnóstico. Debido a esto, los marcadores CSF y la neuroimagen se vuelve clave para el diagnóstico de la enfermedad. En su artículo de 2018 el NIA propone emplear marcadores de la presencia de placas de proteína  $A\beta$ -amiloide (A), acumulación de proteína  $\tau$  (T) y marcadores de neurodegeneración (N). Valorando los tres marcadores a la vez se podrían crear perfiles AT(N) para cada paciente, que podrían ser expandidos para abarcar nuevos marcadores (ATX(N), siendo X cualquier grupo nuevo de marcadores relevantes). Para la valoración de los marcadores se establecen puntos de corte a partir de los cuales los valores son anormales. Por debajo del punto de corte el paciente tiene valores normales, representados con un " - ". Por encima se considera que padece de una patología, representado con un " + ". [21]. Como se verá más adelante, el empleo de estos perfiles para agrupar a los diferentes individuos permitirá identificar factores de riesgo de demencia y aportará información relevante para separar los casos de AD de las demencias debidas a otras patologías.

Posteriormente, en 2024, con el fin de aumentar la información aportada por los modelos, el NIA propone nuevos grupos de marcadores. En primer lugar, expande los marcadores ya propuestos en 2018 en función de si son o no específicos a la AD y de su orden de aparición. Más importante, expande los marcadores propuestos añadiendo los grupos de neuroinflamación (I), vascular (V) y de  $\alpha$ -sinucleína (S). Son especialmente relevantes los marcadores V y S ya que son indicadores de otros tipos de demencia diferentes, ayudando a descartar AD y aportando nuevas explicaciones al origen de la demencia [20]. Estos grupos de marcadores no se utilizarán en los perfiles de este estudio, pero es importante considerarlos para investigaciones futuras, ya que facilitarían el proceso de clasificar las diferentes demencias. Con esta finalidad es especialmente interesante los marcadores de riesgo vasculares. Son fáciles de obtener y están considerados uno de los principales factores de riesgo de demencia [13], al mismo tiempo que ayudan a identificar el origen de la demencia como vascular.

Tras proponer la clasificación en perfiles, el NIA pasa a valorar como se puede utilizar para identificar AD. Tomaron como base para el diagnóstico los marcadores más específicos de AD, el A y T. Además, dado que se observa que el marcador A solo es anómalo en pacientes con AD, concluyen que cualquier individuo con valores anormales de A (A+) ha sufrido "cambios patológicos por Alzheimer". Por otra parte, sabiendo que A es desencadenante de cambios futuros, pero no de los síntomas, se considera necesario T+ para diagnosticar al paciente con AD [21].

Tras esto, pasaron a definir la fase preclínica de la enfermedad. Para ello, definieron la enfermedad como un proceso continuo, el "Continuo de Alzheimer". Se considera que un paciente se encuentra en el continuo de Alzheimer si se observa A+ [21]. En 2018 se definió la evolución normal de este continuo como A+T-(N-), A+T+(N-) y A+T+(N+). En caso de A+T-(N+) se sigue considerando que el paciente está en el continuo de Alzheimer [21]. Además, en el artículo de 2024 se detalla que habitualmente este caso, además de ser raro, se debe a problemas en la sensibilidad de las pruebas o al asignar los puntos de corte. El caso A+T-N+ se puede considerar por tanto equivalente al A+T+(N+) a efectos prácticos [20]. Por otro lado, esta clasificación permite también identificar que la neurodegeneración en casos de A-T-N+ y A-T+N+ se debe a otras copatologías [21, 20].

## 2.2. Modelos de progresión de AD

A la hora de estudiar fenómenos como las enfermedades neurodegenerativas es importante poder estudiar toda su evolución en el tiempo, lo que implica trabajar con datos longitudinales. Un estudio longitudinal es aquel en el que las variables se miden de forma repetida en el tiempo para la misma muestra de la población [8]. El objetivo de estos estudios es observar los cambios en las medidas para cada individuo a lo largo del tiempo, así como para el conjunto de la población y obtener conclusiones sobre los factores que afectan a la medida [8]. Realizar el seguimiento con los mismos individuos en todas las medidas permite obtener valores más precisos, y sin el efecto de la variabilidad entre dos individuos diferentes por razones biológicas, clínicas o experimentales. Además, también permite extrapolar mejor esta información a toda la población, mediante la identificación de las fuentes de variabilidad y

sus efectos [8].

Por la naturaleza de los datos, son comunes grandes desviaciones entre sujetos, carencias de información y tiempos no uniformes entre los datos [8]. Debido a esto, los métodos estadísticos más comunes no son lo suficientemente potentes para analizar adecuadamente la información de los estudios, ni para realizar predicciones en base a ellos [8]. Para solucionar esto, lo más habitual es recurrir a modelos paramétricos continuos que ajustan la dinámica de las variables a funciones habitualmente basadas en modelos de efecto mixto [8, 24, 30]. Así, la herramienta que se emplea para trabajar con estos datos en el contexto, no solo de AD, también de otras enfermedades neurodegenerativas son los **modelos de progresión de enfermedad (DPM)**, basados en el paradigma de los **modelos lineales de efecto mixto (LMEM)** y los **modelos no lineales de efecto mixto (NLMEM)** [24, 37].

### 2.2.1. Modelos Lineales de Efecto Mixto

Los datos longitudinales tienen aspectos a modelar: la trayectoria media de las medidas y la correlación entre ellas. Los dos son interdependientes entre ellos [8]. Los modelos LME se basan en la regresión lineal para describir las medidas y su trayectoria. En estos modelos, la medida media se describe como combinación lineal de variables independientes. La trayectoria, por su parte, viene definida por el tiempo y variables dependientes del tiempo [8]. Gracias a esto, el modelo se puede aplicar a individuos sin medidas de tiempo comunes entre ellos. Una estrategia común para decidir el modelo de la trayectoria es asumirla lineal. Si se quiere una trayectoria que se ajuste mejor a los datos, se aconseja un análisis gráfico de los datos para buscar un modelo más complejo [8].

La ecuación general de los LME es la siguiente [8]:

$$Y_i = X_i\beta + Z_ib_i + e_i \quad (2.1)$$

Donde:

- $Y_i$  es el vector  $n_i \times 1$  de medidas para el sujeto  $i$ , siendo  $n_i$  el número de medidas.
- $X_i$  es la matriz de tamaño  $n_i \times p$  de efectos fijo del individuo. Puede incluir factores como la edad, género o educación.
- $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$  es el vector  $p \times 1$  de coeficientes de regresión de efectos fijos.
- $Z_i$  es la matriz de tamaño  $n_i \times q$ , con  $q \leq p$  de efectos aleatorios. Un ejemplo sería el tiempo de la muestra.
- $b_i = (b_{i1}, b_{i2}, \dots, b_{iq})^T$  es el vector  $q \times 1$  de efectos aleatorios.
- $e_i = (e_{i1}, e_{i2}, \dots, e_{in_i})^T$  es el vector  $n_i \times 1$  de medidas de error.

Todos los efectos aleatorios deben ser variantes con el tiempo [8].

Habitualmente se asumen las siguientes distribuciones [8]:

$$b_i \sim N(0, D),$$

$$e_i \sim N(0, \sigma^2 I_{n_i})$$

Donde  $N(0, \Sigma)$  es una Gaussiana multivariable de media 0 y matriz de covarianza  $\Sigma$ ; y  $I_{n_i}$  es la matriz de identidad  $n_i \times n_i$  [8].

Estos modelos también ofrecen una distinción entre la media condicional y marginal de  $Y_i$ . La media condicional de  $Y_i$  viene dada por la siguiente ecuación [8]:

$$E(Y_i|b_i) = X_i\beta + Z_i b_i \quad (2.2)$$

Por otro lado, la media marginal viene dada por la ecuación [8]:

$$E(Y_i) = X_i\beta \quad (2.3)$$

De esta manera se concluye que el vector  $\beta$  contiene características comunes a todos los individuos del grupo, mientras que  $b_i$  contiene información propia de cada individuo.

Se puede hacer la misma distinción para covarianza condicional y marginal [8]:

$$Cov(Y_i|b_i) = Cov(e_i) = \sigma^2 I_{n_i} \quad (2.4)$$

$$Cov(Y_i) = Cov(Z_i b_i) = Z_i D Z_i^T + \sigma^2 I_{n_i} \quad (2.5)$$

Se observa por tanto que se puede modelar la correlación entre variables.

Conocido todo esto, ahora es necesario poder estimar cada uno de los parámetros del modelo. Estos son los coeficientes del vector  $\beta$ ,  $\sigma$  y  $D$ . Para estimaciones conocidas de  $\hat{\sigma}$  y  $\hat{D}$  existe una solución cerrada para la estimación más probable (ML) de  $\beta$  [8]:

$$\hat{\beta} = \left( \sum_{i=1}^m X_i^T \hat{\Sigma}_i^{-1} X_i \right)^{-1} \left( \sum_{i=1}^m X_i^T \hat{\Sigma}_i^{-1} y_i \right) \quad (2.6)$$

Para  $\hat{\Sigma}_i = Z_i \hat{D} Z_i^T + \hat{\sigma}^2 I_{n_i}$  y siendo  $y_i$  los componentes del vector  $Y_i$ .

Por otra parte, necesitamos la estimación de  $\hat{\sigma}$  y  $\hat{D}$ , que se puede obtener maximizando la función de probabilidad restringida [8]:

$$l_{ReML} = \frac{1}{2} \sum_{i=1}^m \log |\Sigma_i^{-1}| - \frac{1}{2} \sum_{i=1}^m (y_i - X_i \hat{\beta})^T \Sigma_i^{-1} (y_i - X_i \hat{\beta}) - \frac{1}{2} \log \left| \sum_{i=1}^m X_i^T \Sigma_i^{-1} X_i \right| \quad (2.7)$$

Donde  $\Sigma_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$ . Se trata de un problema de optimización sin solución cerrada, por lo que se deben recurrir a métodos numéricos para resolverlo.

Conocidos los parámetros del modelo, es necesario seleccionar los efectos aleatorios para fijar la estructura de la covarianza. Para ello, el método más común es el test de ratio de probabilidad. En este método se empieza por un modelo simple con el error como único efecto aleatorio. Tras esto se van añadiendo al modelo variables dependientes del tiempo que representan los efectos aleatorios. Cada vez que se añade una variable se comprueba si produce cambios significativos en la función de probabilidad restringida, y si no es el caso, se elimina. Para comprobar si son relevantes se puede usar chi-cuadrado [8].

Se realizará por tanto la validación de hipótesis, empleando la matriz  $\hat{\beta}$  y su matriz de covarianza asintótica estimada:

$$\hat{C}ov_{asintotica}(\hat{\beta}) = \left( \sum_{i=1}^m X_i^T \hat{\Sigma}_i^{-1} X_i \right)^{-1} \quad (2.8)$$

En general, usando una matriz L de contraste, hay dos posibles hipótesis, la hipótesis nula,  $H_0 : L\beta = 0$  y la hipótesis no nula,  $H_A : L\beta \neq 0$ . En condiciones de hipótesis nula, la función F de distribución será:

$$F = \frac{(L\hat{\beta})^T (L\hat{C}ov_{asintotica}(\hat{\beta})L^T)^{-1} L\hat{\beta}}{rank(L)} \quad (2.9)$$

Determinar los grados de libertad de la función es un proceso complicado. Existen varias aproximaciones diferentes para ello, como la que se emplea en el informe, basada en Satterthwaite [8]:

$$F = \frac{\kappa(L\hat{\beta})^T (L\hat{C}ov_{KR}(\hat{\beta})L^T)^{-1} L\hat{\beta}}{rank(L)} \quad (2.10)$$

Donde  $\hat{C}ov_{KR}$  es la estimación, corregido el sesgo por muestra pequeña, de la matriz de covarianza de  $\hat{\beta}$ . Esto permite tener en cuenta la covarianza de la función de máxima probabilidad restringida en la estimación de los grados de libertad [8].

### 2.2.2. Modelos de Progresión de Enfermedad

Los LMEM y NLMEM, aunque son un buen punto de partida, no son del todo completos a la hora de evaluar enfermedades. Para solucionar este problema se pasa a emplear DPMs, que añaden al modelo la edad de la enfermedad en el individuo [37] pasando la ecuación a ser la siguiente [37]:

$$y_{ijk} = g_k(t_{ijk}^c + \delta_i) + x'_{t_{ijk}^c} \beta_k + \alpha_{0_{ik}} + \alpha_{1_{ik}} t_{ijk}^c + e_{ijk} \quad (2.11)$$

Donde:

- $g_k$  es una función continua, monótona y diferenciable.
- $\delta_i$  es un tiempo desconocido y específico para cada sujeto, a estimar. Representa la edad de la enfermedad en el individuo. Sigue una distribución normal de media cero y varianza  $\sigma_\delta^2$

- $t_{ijk}^c$  es el tiempo de observación a corto plazo del individuo. Indica años centrados en relación a la evolución de las visitas:  $t_{ijk}^c = t_{ij} - (t_{i_1} + t_{i_{end}})/2$  Donde  $t_{ij}$  es el tiempo de la visita y  $t_{i_1}$  y  $t_{i_{end}}$  son la primera y última visita registrada para el paciente  $i$ .
- $x'_{t_{ijk}^c}$  es el vector de efectos fijos, igual que en un LME.
- $\beta_k$  es el coeficiente de efectos fijos, una vez más basado en un LME.
- $\alpha_{0_{ik}}$  y  $\alpha_{1_{ik}}$  son la pendiente e intersección específicas para el resultado e individuo.
- $e_{ijk}$  es el vector de medidas de error. Sigue una Gaussiana de media cero y varianza  $\sigma_k$

Al elaborar DPMs es importante tener en cuenta que se pueden elaborar para los diferentes individuos de una población y para la población general. Utilizando geometría de Riemann, considerando que cada observación pertenece a una variedad de Riemann y que existe un geodisco que representa la trayectoria espacio-temporal media del grupo [24]. Se considera que las trayectorias individuales son variaciones espacio-temporales de esta trayectoria y derivan de ella gracias a:

- Una variación espacial definida por un desplazamiento espacial  $w_i \in T_{p_0}M$ . Caracteriza la dirección con la que la trayectoria media del grupo se desplaza con la información de cada individuo. Se puede aplicar transporte paralelo al vector  $w_i$ . El resultado es  $P_{\gamma_0, t_0, t}(w_i)$ . Sabiendo esto, las trayectorias individuales vienen definidas por su mapa exponencial  $\eta^{w_i}(t) := \text{Exp}_{\gamma_0(t)}(P_{\gamma_0, t_0, t}(w_i))$ . [24]
- Una variación temporal definida por un factor de aceleración  $\alpha_i \in \mathbb{R}$  y un desplazamiento temporal  $\tau_i \in \mathbb{R}$ . Una observación individual en un tiempo  $t_{ij}$  corresponde a la edad  $\psi_i(t_{ij})$ . El factor de aceleración indica la velocidad de la dinámica. Las más rápidas tienen un valor de  $\alpha_i > 1$  respecto a  $\alpha_i < 1$  para las más lentas. El desplazamiento temporal indica un inicio de la progresión más temprano ( $\tau_i < 1$ ) frente a una más tardía ( $\tau_i > 1$ ). [24]

Dado que el espacio de características es normalmente de alta dimensión, se puede representar  $w_i$  en una matriz  $w_i = As_i$  [24].  $A \in \mathbb{R}^{N \times N_s}$  es la matriz de mezclado y  $s_i = (s_{ij})_{1 \leq j \leq N_s}$  son las fuentes [24]. Con todo lo anterior, una medida individual del  $i$ -ésimo individuo en un tiempo  $t_{ij}$  es:

$$y_{ij} = \eta^{w_i}(\psi_i(t_{ij})) + \epsilon_{ij}, \quad (2.12)$$

siendo  $\epsilon_{ij}$  el ruido no captado por el modelo [24].

Además, en el Espacio de Riemann podemos describir con facilidad las trayectorias espacio-temporales de los individuos. Por un lado la progresión de las enfermedades a largo plazo se corresponde al geodisco  $\gamma_0$  parametrizado por  $p_0, t_0$  y  $v_0$ . Por otro lado, la función  $t \mapsto \eta^{w_i}(\psi_i(t_{ij}))$  que corresponde a la descripción geométrica de la trayectoria de un individuo en la variedad de Riemann. Deriva de  $\gamma_0$  gracias a  $\alpha_0, t_0$  y  $w_i$ . [24]

La separación entre las trayectorias individuales y las de la población (derivando las trayectorias individuales de las del conjunto) se puede ver como un modelo de efecto mixto. El geodisco se corresponde con los efectos fijos mientras que los parámetros  $\alpha_i, \tau_i$  y  $w_i$  son los efectos aleatorios. [24]

Al elaborar DPMs es habitual elaborar varios modelos para los mismos marcadores, empleando técnicas como el *bootstrapping* [37], que consiste en el remuestreo y reinferencia de la población para medir la calidad de la muestra obtenida.

Además, es necesario establecer un punto de corte para los marcadores a partir del cual se considera que se entra en demencia. Para ello, se pueden definir varias puntuaciones basadas en la edad de la enfermedad (*Disease Progression Score*, DPS) y el progreso a demencia [37]. Definiendo el año inicial de la demencia como  $t_{onset}$ , cabría esperar que las trayectorias de los sMCI se encuentren a su izquierda [37]. Por otro lado, lo pMCI cruzarán el  $t_{onset}$  moviéndose hacia la derecha. Se definirán por tanto varias puntuaciones que permitan evaluar el estado de los grupos [37]. En primer lugar, se evalúa el número de sMCI cuyas últimas visitas ocurrieron antes de  $t_{onset}$  en relación al número de sMCI [37]:

$$SCORE_1 = \frac{\#\{i|(i \in sMCI) \cap ()(t_{i_1} + \delta_i) < t_{onset})\}}{\#sMCI} \quad (2.13)$$

Se emplea la proporción de pMCI que convirtió a demencia y no tenía demencia de base [37]:

$$SCORE_2 = \frac{\#\{i|(i \in pMCI) \cap ()(t_{i_1} + \delta_i) < t_{onset})\}}{\#pMCI} \quad (2.14)$$

Los pacientes cuya última visita era mayor que  $t_{onset}$  [37]:

$$SCORE_3 = \frac{\#\{i|(i \in pMCI) \cap ()(t_{i_1} + \delta_i) > t_{onset})\}}{\#pMCI} \quad (2.15)$$

Por último, los pacientes con demencia al principio del estudio [37]:

$$SCORE_4 = \frac{\#\{i|(i \in Dementia) \cap ()(t_{i_1} + \delta_i) > t_{onset})\}}{\#Dementia} \quad (2.16)$$

El valor óptimo de  $t_{onset}$  será el que maximice todas las puntuaciones y les de valores similares [37].

Estos modelos muestran una buena capacidad para clasificar entre pacientes sMCI y pMCI [37]. Además, se observa que es más fácil predecir la conversión a demencia cuando se produce menos de 4 años desde el diagnóstico MCI [37]. De la misma manera, los sMCI de más de 4 años son más fáciles de clasificar como no conversión a demencia [37]. Se pueden emplear los marcadores de riesgo para discriminar también dos subgrupos: pMCI-rápido, con la mayor probabilidad de conversión a demencia (80%) y sMCI-lento, con la menor (14%) [37]. El biomarcador más significativo al realizar esta clasificación es la A-beta, pero también se tienen en cuenta diagnósticos tardíos de MCI (*Late Mild Cognitive Impairment*, LMCI) y el estado de APOE4 (factor de riesgo genético) [37].

Una vez más, es importante recordar que los modelos definen unicamente la trayectoria de los marcadores. Para obtener una clasificación en función del estado de la enfermedad se recurre a estimadores de densidad kernel (*Kernel Density Estimator*, KDE) con clasificadores Bayesianos [30].

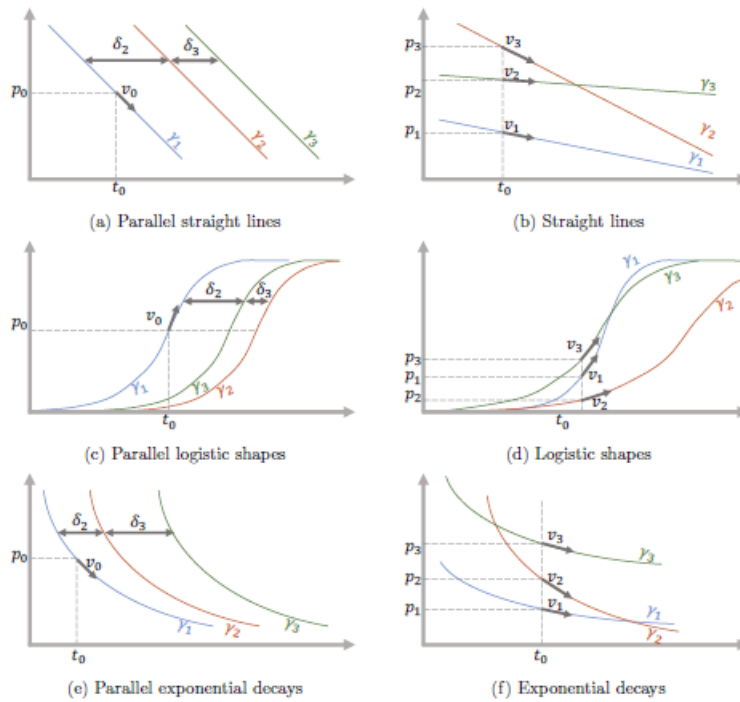
### 2.2.3. Construcción de los modelos - Leaspy

Sabiendo en que consisten los DPMS, es necesario saber las diferentes modelos concretos que se pueden utilizar y cómo elaborar un modelo genérico.

En primer lugar, es necesario saber que para asegurar la identificabilidad del modelo en presencia de variabilidad espacio-temporal, el desplazamiento espacial  $w_i$  debe ser ortogonal a la velocidad  $v_0$  [24]. Si los dos vectores no fuesen ortogonales, la proyección de  $w_i$  sobre  $\gamma_i$  sería no nula e interferiría con la progresión temporal controlada por  $\alpha_i$  y  $\tau_i$  [24]. En conclusión, se debe asegurar  $g^M(w_i, v_0) = 0$  siendo  $g^M$  las métricas de la variedad  $M$ . [24]

En cuanto a los diferentes modelos que se pueden utilizar, en la tesis de Koval se proponen como ejemplo los siguientes modelos más comunes:

Figura 2.1: Ejemplos de Posibles Modelos. [24]



- La primera posibilidad es considerar **líneas paralelas rectas**. El  $k$ -ésimo geodisco alcanza el valor  $p_0$  con una velocidad  $v_0$  en un tiempo  $t_0 + \delta_k$ . La diferencia entre los geodiscos es unicamente el desfase  $\delta_k$ . Para poder identificar mejor cada modelo,  $\delta_1 = 0$ .
- **Líneas Rectas**. Similar al anterior, pero el  $k$ -ésimo geodisco alcanzará el valor  $p_k$  con valor  $v_k$  en un tiempo  $t_0$ .

- **Curvas logísticas paralelas.** Se consideran curvas con asíntota de 0 en  $-\infty$  y 1 en  $+\infty$ . Como en el primer ejemplo, las curvas se encuentran desfasadas un valor  $\delta_k$
- **Curvas logísticas.** Como el modelo anterior, pero se consideran todas las coordenadas independientes.
- **Decaimientos Exponenciales Paralelos.** Se consideran curvas exponenciales con asíntotas  $+\infty$  en  $-\infty$  y de 0 en  $+\infty$ . Una vez más, los demás geodiscos son el primero trasladado  $\delta_k$ .
- **Decaimientos Exponenciales.** Como el modelo anterior, pero se consideran todas las coordenadas independientes.

Sobre estos modelos se pueden realizar variaciones. Una forma habitual es mediante la selección de los parámetros del modelo. Por ejemplo, en vez de observar el punto y velocidad que se alcanzan a distintos tiempos, se puede observar la velocidad y el tiempo que se tarda en alcanzar distintos puntos [24]. Es importante recordar que no todas las parametrizaciones son igual de fáciles de estimar. Otra posible variación para trabajar con datos unidimensionales es puede aplicar los modelos anteriores forzando la condición  $w_i = (w_{i1}) = 0$  [24]. Por último, el número de fuentes  $N_s$  define el subespacio de direcciones posibles para el desplazamiento espacial  $w_i$  [24]. Para seleccionarlo se tiene en cuenta que un número muy pequeño no logra abarcar todo  $\mathbb{R}^N$  mientras que uno demasiado grande satura la cantidad de direcciones posibles y dispara el coste computacional [24]. El mejor valor es el valor mínimo que sature el ruido [24].

Teniendo todo esto en cuenta, es necesario realizar la estimación de los modelos. Para ello, se parte del siguiente modelo genérico [24]:

$$y_i = f(\theta, z_i) + \epsilon_i \quad (2.17)$$

Siendo:

- $y_i$  la variable de salida
- $z_i$  las variables aleatorias asociadas al sujeto  $i$ . También llamadas variables latentes
- $f$  la función del modelo
- $\theta$  los parámetros del modelo

En datos longitudinales, con los que se trabajaran,  $y_i$  es un conjunto de observaciones en distintos momentos temporales,  $y_i = y_{i,j}$  [24]. El objetivo es encontrar los parámetros del modelo que mejor describan las observaciones.

La probabilidad depende de las variables latentes  $z_i$  que son por definición desconocidas. Esto complica considerablemente calcularla de forma directa, por lo que se recurre a algoritmos iterativos. Uno de estos es el Algoritmo de Máxima Expectación [24]:

---

**Algorithm 1** Algoritmo de Máxima Expectación [24]
 

---

```

 $\theta \leftarrow \theta_0$ 
 $k \leftarrow 0$ 

while Convergence of  $\theta$  do
   $k \leftarrow k + 1$ 

  Calculo de Expectación:  $Q(\theta, \theta^k) = \mathbb{E}_{p(z|y; \theta^k)}[\log p(y, z; \theta)]$ 

  Maximización:  $\theta^{k+1} = \operatorname{argmax}_{\theta} Q(\theta, \theta^k)$ 
end while
return  $\theta^{(k)}$ 

```

---

El algoritmo alterna entre el paso de Expectación, que define una función que calcula la expectativa sobre las variables latentes dado un valor actual del parámetro. Y el paso de maximización, que toma el parámetro que maximice la función anterior.

Para modelos complejos no lineales, la integral sobre las variables latentes se vuelve intratable. Se reemplaza por tanto el paso de expectación con uno de aproximación estocástica. Al utilizar una única muestra por iteración, ahorra en coste computacional en comparación con otras aproximaciones como la de Monte Carlo [24].

---

**Algorithm 2** Algoritmo de Maximización de Expectación con Aproximación Estocástica [24]
 

---

```

 $\theta \leftarrow \theta_0$ 
 $Q_0 \leftarrow 0$ 
 $(\epsilon_k)_k \geq 0$  tal que  $\sum_{k \geq 0} \epsilon_k = +\infty$  y  $\sum_{k \geq 0} \epsilon_k^2 < +\infty$ 
 $k \leftarrow 0$ 

while Convergence of  $\theta$  do
   $k \leftarrow k + 1$ 

  Paso Estocástico:  $z^{(k)} \sim p(z|y; \theta^{(k)})$ 
  Paso de Aproximación:  $Q_k(\theta) = Q_{k-1}(\theta) + \epsilon_k(\log p(y, z^{(k)}; \theta) - Q_{(k-1)})$ 

  Maximización:  $\theta^{k+1} = \operatorname{argmax}_{\theta} Q_k(\theta)$ 
end while
return  $\theta^{(k)}$ 

```

---

Otra dificultad que se puede encontrar es que  $p(z|y; \theta^{(k)})$  puede ser desconocido. Del teorema de Bayes se deriva [24]:

$$p(z|y; \theta^{(k)}) = \frac{p(y|z; \theta^{(k)})p(z; \theta^{(k)})}{p(y; \theta^{(k)})} = \frac{p(y|z; \theta^{(k)})p(z; \theta^{(k)})}{\int p(y|z; \theta^{(k)})p(z; \theta^{(k)})dz} \quad (2.18)$$

Donde  $p(y|z; \theta^{(k)})$  es el propio modelo,  $p(z; \theta^{(k)})$  es el prior conocido de las variables ocultas, y el denominador es una constante. Se puede utilizar por tanto el algoritmo de Metropolis-Hastings para aproximar  $z^{(k+1)}$  [24]. A partir de aquí se pueden calcular los parámetros con la Aproximación Estocástica de la Cadena de Markov de Monte Carlo para la Maximización de Expectación [24].

**Algorithm 3** Algoritmo de Metropolis-Hastings [24]

---

$z^{(k)}$   
 Elegir una ley de proposición:  $q_k(\cdot|z^{(k)})$   
 Tomar  $z^c \sim q_k(\cdot|z^{(k)})$   
 Actualizar  $z^{(k+1)} = z^c$  con probabilidad  $\tau = \min\left(\frac{p(z^c|y;\theta^{(k)})q_k(z^{(k)}|z^c)}{p(z^{(k)}|y;\theta^{(k)})q_k(z^c|z^{(k)})}, 1\right)$   
 $(z^{(k+1)} = z^{(k)} \text{ en otro caso})$   
**return**  $z^{(k+1)}$

---

**Algorithm 4** Aproximación Estocástica de la Cadena de Markov de Monte Carlo para la Maximización de Expectación [24]

---

$\theta \leftarrow \theta_0$   
 $\tilde{S}_0 \leftarrow 0$   
 $(\epsilon_k)_k \geq 0$  tal que  $\sum_{k \geq 0} \epsilon_k = +\infty$  y  $\sum_{k \geq 0} \epsilon_k^2 < +\infty$   
 $k \leftarrow 0$   
**while** Convergence of  $\theta$  **do**  
      $k \leftarrow k + 1$   
     **Paso de Simulación:** dado  $(y, \theta^{(k)})$ , muestrear  $z^{(k)}$  de  $z^{(k-1)}$  por el algoritmo de Metropolis-Hastings  
     **Aproximación Estocástica:**  $\tilde{S}_k \leftarrow \tilde{S}_k + \epsilon_k(S(y, z^{(k)}) - \tilde{S}_{k-1})$   
     **Maximización:**  $\theta^k = \operatorname{argmax}_{\theta}(-\log C(\theta) + \langle \tilde{S}_k, \Phi(\theta) \rangle)$   
**end while**  
**return**  $\theta^{(k)}$

---

Para variables multivariantes complejas es difícil muestrear  $z^{(k)}$  mediante el método de la Cadena de Markov. Para solucionarlo, se combina el algoritmo de Metropolis-Hastings con el muestreador de Gibbs[24].

**Algorithm 5** Muestreador de Gibbs [24]

---

Dado un conjunto de variables ocultas  $z^{(k)} = (z_l^{(k)} | 1 \leq l \leq L)$  y parámetros  $\theta^{(k)}$   
 Muestrear:  $z_1^{(k+1)}$  de  $q(z_1|y, z_1^{(k)}, \dots, z_L^{(k)}, \theta^{(k+1)})$   
 ...  
 Muestrear:  $z_l^{(k+1)}$  de  $q(z_l|y, z_1^{(k+1)}, \dots, z_{l-1}^{(k+1)}, z_{l+1}^{(k)}, \dots, z_L^{(k)}, \theta^{(k+1)})$   
 ...  
 Muestrear:  $z_L^{(k+1)}$  de  $q(z_L|y, z_1^{(k+1)}, \dots, z_{L-1}^{(k+1)}, z_L^{(k)}, \theta^{(k+1)})$   
**return**  $\theta^{(k+1)}$

---

Tras detallar los algoritmos que podemos utilizar para estimar los parámetros, es necesario profundizar en los pasos necesarios para estimar los modelos. Como ya se ha descrito anteriormente, el modelo esta caracterizado por sus parámetros geométricos y por variables ocultas individuales. Los parámetros geométricos dependen de la instanciación del modelo. Las variables ocultas por su parte son las variaciones de cada individuo sobre la media del grupo [24].

Sabiendo esto, para la estimación del DPM se puede dividir en los siguientes procesos [24]:

- **Calibración:** Dado  $y$ , estimar los parámetros  $\hat{\theta}$  que mejor se ajustan a la trayectoria media del grupo [24].
- **Personalización:** Dados  $\hat{\theta}$ , estimar las variables ocultas individuales  $z_i^*$  que mejor derivan la trayectoria promedio del grupo para reconstruir las mediciones individuales [24].
- **Reconstrucción, imputación de valores faltantes y predicción de futuro:** Dados  $\hat{\theta}$  y  $z_i^*$ , estimar valores de  $\tilde{y}_{ij} = f(\theta, z_i^*, t_{ij})$  tanto para valores de tiempo observados, como para valores intermedios y futuros [24].
- **Simulación:** Dados  $\hat{\theta}$  y  $z_i$ , añadir una nueva variable oculta  $z_{i'}$  que determine un individuo nuevo [24].

El proceso de calibración busca encontrar los parámetros  $\hat{\theta}$  que mejor describe el modelo, habitualmente, que maximiza la probabilidad  $p(y; \theta)$  [24]. El modelo, así como la variabilidad de la población quedan caracterizados por un valor óptimo  $\hat{\theta} \in \operatorname{argmax}_{\Theta} p(y; \theta)$  [24]. Dado que las variables ocultas son desconocidas, la probabilidad logarítmica es intratable.

Suponiendo que el modelo pertenece a la familia exponencial de tal manera que la probabilidad logarítmica completa sea:  $\log p(y, z; \theta) = \langle \Phi(\theta), S(y, z) \rangle + \log C(\theta)$ , se puede asegurar la convergencia del algoritmo [24].

Para realizar la personalización se considera un conjunto de datos longitudinales  $y$  y el valor óptimo de  $\hat{\theta}$  que maximiza  $p(y; \theta)$ , se toma un conjunto de  $k_i$  observaciones del mismo paciente [24]. La personalización consiste en encontrar el valor óptimo de  $z_i^*$  que maximice la probabilidad  $p(y_i, z_i; \hat{\theta}) = p(y_i | z_i; \hat{\theta}) p(z_i; \hat{\theta})$  [24]. Este proceso anterior especifica la trayectoria espacio-temporal de un individuo para cualquier  $t$  [24].

Elegir una  $t$  que se corresponda con la de una de las muestras se llama reconstrucción de los datos. Esto permite conocer el error provocado por el modelo, llamado error de reconstrucción [24].

Si se elige una  $t$  que se encuentre entre dos muestras, se corresponde con la interpolación de los datos entre las muestras [24].

Por último, elegir una  $t > t_k$  se corresponde con una predicción de futuro. Esta predicción implica suponer que la trayectoria media del grupo describe un escenario en el futuro que es aplicable a las trayectorias individuales. Por lo tanto, es desaconsejable realizar predicciones de tiempos no incluidos en la base de datos de entrenamiento [24].

Por último, el proceso de simulación permite generar una cohorte virtual que reproduzca las características de la original. Gracias a esto, se puede contar con más pacientes y visitas. Los pacientes simulados son combinaciones lineales de los reales, ya que de esta manera se puede asegurar que es muy probable que existan, aunque

no se encuentren en la muestra [24].

Conocidos como son los modelos, queda conocer los paquetes y librerías para diferentes lenguajes de programación que se pueden utilizar para elaborarlos. Los más habituales en la literatura son Leaspy [36], de Python; Grace [37, 36], de R; y RPDPM [30, 36], de Matlab.

#### 2.2.4. RPDPM

De los modelos mencionados anteriormente es especialmente relevante detenerse en RPDPM (*Robust Parametric Disease Progression Model*) [30]. En él se pueden ver muchos de los principios descritos anteriormente para construir los modelos, pero además se incluye el uso de estimadores de máxima verosimilitud para aumentar la robustez ante datos erróneos o faltantes sin necesidad de aumentar el número de parámetros [30].

Entrando más en profundidad en como es RPDPM, lo primero a mencionar es cual es la función del modelo a la que ajusta las variables, así como cuales son los parámetros a ajustar. La función que emplea es una función de Stannard modificada [30]:

$$g(s; \theta) = \left[ 1 + \frac{1}{\gamma} e^{-\frac{b}{\gamma}(s-c)} \right]^{-\gamma} \quad (2.19)$$

Siendo los parámetros  $\theta = \{b, c, \gamma\}$ .

Con esto en cuenta, se define, al igual que en la literatura anterior [8, 24], la estimación de cada biomarcador. En este caso, se considera  $y_{i,j,k}$  como la medida,  $y$ , del biomarcador  $k$ , para el sujeto  $i$  en la visita  $j$  [30]. De esta manera, la ecuación de las medidas de los biomarcadores es:

$$y_{i,j,k} = f(s_{i,j}; \theta_k) + \sigma_k \epsilon_{i,j,k} \quad (2.20)$$

Siendo  $\sigma_k$  la desviación típica de las medidas para el marcador y  $\epsilon_{i,j,k}$  ruido blanco de tipo Gaussiano. Por otra parte,  $s_{i,j}$  es la DPS del paciente  $i$  en su visita  $j$ , dado por la siguiente ecuación. [30]

$$s_{i,j} = \alpha_i t_{i,j} + \beta_i \quad (2.21)$$

Siendo  $t_{i,j}$  la edad del paciente  $i$  en la visita  $j$ ,  $\alpha_i$  su tasa de progresión de la enfermedad y  $\beta_i$  su onset de su enfermedad. [30]. Así, la función de regresión no lineal que optimiza RPDPM es: [30]

$$\{\hat{\alpha}, \hat{\beta}, \hat{\theta}\} = \min_{\alpha, \beta, \theta} \sum_{i,j,k} w_i \rho \left( \frac{y_{i,j,k} - f(\alpha_i t_{i,j} + \beta_i; \theta_k)}{\sigma_k} \right) \quad (2.22)$$

Siendo  $p(\cdot)$  la función tipo máxima verosimilitud que introduce RPDPM, y  $w_i = 1/N_i$  un factor ponderado de normalización de la función que depende del número de medidas disponibles por sujeto  $N_i$  [30]. El rango de la función es  $(0, 1)$ . [30]

En cuanto a como ajustar los parámetros, el método es muy similar al ya descrito anteriormente [24, 30]. Si es importante mencionar que la función de pérdida que dio mejores resultados fue la logística, dada por la ecuación [30]:

$$\rho(r) = \ln(\cosh(r)) \quad (2.23)$$

Para un factor de escala  $\tau = 1,205$  [30]

En cuanto a los resultados de la aplicación de RPDPM, en el artículo se comprueba que los modelos permiten discriminar con facilidad las trayectorias de CN y demencia, pero hay solapamiento entre CN y MCI, y entre MCI y demencia [30]. Además, los modelos definen el mayor crecimiento de los marcadores CSF durante la etapa MCI, y de los marcadores cognitivos durante la demencia [30].

## 2.3. Aportación al Trabajo

### 2.3.1. Perfiles AT(N)

La clasificación de los pacientes en perfiles AT(N) es una de las bases en las que se fundamenta el trabajo, así como considerar la AD como la presencia de placa amiloide [21]. Partiendo de esta idea, se podrá elaborar una clasificación en perfiles AT(N) en un primer estudio transversal, que permitirá identificar la presencia de demencias no debidas al Alzheimer en el grupo. Además, para trabajos futuros se podrían incorporar otros marcadores, como los vasculares. [20].

### 2.3.2. Modelos de progresión de enfermedad

Para la segunda mitad del trabajo, como ya se ha indicado, se elaboraran modelos longitudinales de progresión de la enfermedad. Para ello, es necesario conocer la base matemática de los LME [8] y DPMs [37, 24]. Además de como elaborar los DPM, cual es el proceso para construir modelos de la población y personalizarlos, y algunas herramientas que se pueden utilizar, como Leaspy. [24, 32, 25]. Hay que remarcar también que los DPM por si solos no ofrecen el diagnóstico, por lo que será necesario después aplicar estimación de densidad Kernel a los resultados, junto con un clasificador Bayesiano para conseguir la clasificación clínica [30]. Los clasificadores de Bayes-KDE se desarrollan más en el anexo C. Por último, en el caso concreto de RPDPM, se utilizará para su comparación con los modelos elaborados en Leaspy.



## Capítulo 3

# Materiales

### 3.1. Python

#### 3.1.1. Leaspy

Leaspy es un paquete para Python 3.8 o superior, centrado en el análisis de datos longitudinales. El propio paquete de Leaspy contiene métodos que permiten elaborar el modelo de la población general, personalizarlo a cada individuo, reconstruir datos faltantes y realizar la predicción de datos futuros[25]. Gracias a esto, el paquete simplifica enormemente la elaboración de los DPM.



Figura 3.1: Logo de Leaspy[25]

### 3.2. Matlab

#### 3.2.1. RPDPM

RPDPM son las siglas de *Robust Parametric Disease Progression Modeling*. Es un algoritmo programado en Matlab para la elaboración de DPMs y el análisis de datos longitudinales [30]. Si bien no se empleará directamente durante el trabajo, si que se compararán los resultados obtenidos en Leaspy con los resultados ofrecidos por una implementación de RPDPM [36].

### 3.3. ADNImerge

Los datos de los pacientes para el estudio pertenecerán al paquete ADNImerge, de la *Alzheimer's Disease Neuroimaging Initiative* (ADNI). ADNI es una iniciativa

de colaboración público-privada formada por centros de investigación de Estados Unidos y Canadá. ADNI reúne información sobre una cohorte de pacientes norteamericanos en diferentes fases de la progresión de la enfermedad. [2].



Figura 3.2: Logo de ADNI[2]

### 3.4. Magerit

Una vez se ha probado los algoritmos encargados de construir los modelos en Curie para verificar su funcionamiento, se pasará a una máquina más potente. Para ello se empleará Magerit-3 del Centro de Supercomputación y Visualización de Madrid (CeSViMa) de la Universidad Politécnica de Madrid (UPM) [4].

Magerit-3 es un clúster formado por 72 nodos Lenovo ThinkSystem SD530, cada uno de los cuales dispone de 2 procesadores Intel® Xeon® Gold 6230 de 20 cores @ 2.10 GHz (1.344 GFLOPS) y 192 GiB de RAM, y 48 nodos Lenovo ThinkSystem SD530 (cada uno de los cuales dispone de 2 procesadores Intel® Xeon® Gold 6240R de 24 cores @ 2.40 GHz (1.344 GFLOPS) y 768 GiB de RAM) [4]. Funciona en el sistema operativo CentOS [4], basado en la distribución de Linux Fedora [3].

Por esta razón, Magerit se utilizará elaborar todos los modelos de Leaspy con todas las combinaciones posibles de marcadores. La potencia de la máquina reducirá enormemente el tiempo necesario para esto, lo que permitirá elegir la mejor combinación de marcadores posible con la que seguir trabajando sin necesidad de limitarse a probar combinaciones concretas.



Figura 3.3: Logo del CeSViMa[4]

## 3.5. Comunicación con Curie y Magerit

### 3.5.1. PuTTY

Para la comunicación con *Curie* y *Magerit* se empleará *PuTTY*. PuTTY es un cliente de SSH de código abierto desarrollado para Windows. Con él se puede establecer conexiones eficientes y seguras con la máquina externa, además de guardar su dirección para más fácil acceso. [41]

### 3.5.2. Filezilla

Filezilla es un programa de código abierto usado para el envío y gestión de archivos de forma segura a través de diferentes redes [23]. Se utilizará para el envío y la descarga de archivos de *Curie* y *Magerit*.



Figura 3.4: Logo de Filezilla[23]

### 3.5.3. Xming

Xming es un servidor de sistema de ventanas X gratuito para el sistema operativo Windows. Permite que, al conectarse en remoto a ordenadores o redes de ordenadores que funcionan en UNIX, se pueda utilizar en el sistema Windows la interfaz gráfica de los programas que se ejecuten en el otro ordenador.[14] De esta manera, permitirá el uso de Spyder en *Curie* y *Magerit*.



Figura 3.5: Logo de Xming[14]

### 3.5.4. Tmux

Tmux es un multiplexor de terminales para sistemas Linux, macOS, BSD y Solaris. Utilizando tmux se puede multiplexar el terminal, creando varias sesiones dentro del mismo. Dentro de las sesiones creadas se pueden ejecutar programas que reciben un terminal propio gestionado por tmux [28]. De esta manera, cuando se quiera por ejemplo ejecutar el script para crear los modelos, se podrá crear una sesión de tmux en la que ejecutarlo y desconectar de la máquina, dejando el programa en segundo plano. Se utilizará durante las pruebas en *Curie* para poder iniciar el programa que calcula los modelos y, una vez este funcionando poder cortar la conexión con la máquina sin cerrarlo.



Figura 3.6: Logo de tmux[28]

## 3.6. Marcadores

### 3.6.1. Marcadores neuropsicológicos (NM)

En la práctica clínica, el diagnóstico de demencia se realiza mediante la valoración de sus síntomas. Para ello se realizan pruebas neuropsicológicas, con las que se obtienen de resultado marcadores neuropsicológicos que valoran el estado de las diferentes funciones cognitivas. Se busca que mediante los DPM realizados se puedan llegar a resultados aplicables en el ámbito clínico. Además para entrenar los modelos siempre se utiliza como información verdadera el diagnóstico clínico. Por ello, es importante conocer e incluir los NM en los DPMs. Los más utilizados en el diagnóstico de demencia por AD son los siguientes:

- **Rey Auditory Verbal Learning Test (RAVLT)**. El RAVLT es un test normalizado y fácil de realizar para evaluar la memoria verbal. Para ello se valora la capacidad de recordar palabras de una lista con y sin distracciones e interferencias, así como la retención tras un periodo largo de tiempo. [27]
- **Mini-Mental State Examination (MMSE)**. El Mini-Mental es un test que se utiliza para valorar las funciones cognitivas del paciente de forma general. Incluye test para evaluar la capacidad de orientación, atención, velocidad de cálculo y memoria visual. [27]
- **Alzheimer's Disease Assessment Scale (ADAS13)**. Evalúa la gravedad de los síntomas cognitivos del paciente y está considerada una de las pruebas más importantes para el diagnóstico. Engloba la memoria verbal, orientación y comprensión de información. [27]
- **Clinical Dementia Rating Sum of Boxes (CDRSB)**: Escala que da una puntuación general de la severidad y progresión de la demencia. Valora la memoria, orientación, y la capacidad del paciente de valerse por si mismo para la resolución de problemas y vida diaria. Es el que se emplea para decidir el diagnóstico clínico. [40]
- **Functional Activities Questionnaire (FAQ)**: Cuestionario que valora la capacidad del paciente de realizar tareas de la vida diaria, valorando también su complejidad. Especialmente útil par identificar el paso de MCI a demencia. [29]
- **Trail Making Test Part B Score (TRABSCOR)**: Se trata de la puntuación de la prueba B del Trail Making Test (TMT). Esta prueba concreta consiste en unir 25 círculos que contienen letras y números, alternando las letras y los números en orden de menor a mayor y alfabético. Es decir, 1-¿A-¿2-¿B... La puntuación se asigna en función del tiempo tardado, y permite estimar daño cerebral en general. [39]

- **Modified Preclinical Alzheimer Cognitive Composite Trails B (mPACC-CtrailsB)**: es una versión modificada del Preclinical Alzheimer's Cognitive Composite (PACC) que incorpora como uno de sus componentes el Trail Making Test Part B. Se emplea para monitorizar la fase preclínica. [22]
- **Modified Preclinical Alzheimer Cognitive Composite Digit Symbol (mPACCdigit)**: El Modified Preclinical Alzheimer's Cognitive Composite con Digit Symbol (mPACC-Digit) es una versión modificada del Preclinical Alzheimer Cognitive Composite (PACC) que incluye el Digit Symbol Substitution Test (DSST) como componente para evaluar velocidad de procesamiento del paciente. Una vez más, es especialmente útil en la fase preclínica [11].
- **Logical Memory Delayed Total Recall (LDELTOTAL)**: Durante esta prueba se le recitan al paciente dos historias con una cantidad fija de información. Inmediatamente después se les pide que las reciten de vuelta. Media hora después se les vuelve a preguntar por lo que recuerden. El objetivo es identificar y cuantificar cuanto han olvidado en esa media hora. [26]

### 3.6.2. Marcadores de líquido cefalorraquídeo (CSF)

Como se pudo ver en el capítulo 2 la presencia o no de AD se define por la presencia de patología amiloide, y la presencia de síntomas esta relacionada con la de patología tau [21]. Por esta razón, se buscará que al menos un grupo de los pacientes tengan medidas de los marcadores de  $A\beta$ , TAU y pTAU.



# Capítulo 4

## Métodos

### 4.1. Estudio Transversal

#### 4.1.1. Estudio Transversal de cada individuo en su primera visita

En primer lugar, se realizará un estudio transversal de todos los individuos facilitados por ADNI en su primera visita. Se valorará cuales de los pacientes de ADNI tienen información relevante para el estudio. Se buscará también una visión general de sus marcadores. Por último, para facilitar el estudio de la progresión de la enfermedad, se les clasificará en los siguientes grupos:

- **sMCI:** (*Deterioro Cognitivo Leve Estable.*) Pacientes con diagnóstico MCI en la primera y última visita.
- **pMCI:** (*Deterioro Cognitivo Leve Progresivo.*) Pacientes con diagnóstico de MCI en la primera visita y demencia en la última.
- **Demencia:** Pacientes con diagnóstico de demencia en la primera visita.

El primer paso es eliminar las visitas que no son relevantes para el estudio, ya sea porque pertenecen a diagnósticos cognitivamente normales (CN) o porque el paciente o diagnóstico no cumple alguna de las condiciones que exige Leaspy para poder estimar su modelo. Además, es necesario convertir el tipo de dato de algunos marcadores. El proceso, así como las visitas a eliminar se detallan en el siguiente algoritmo:

---

**Algorithm 6** Preprocesado de datos de ADNIMerge

---

1. Se cargan los datos facilitados por ADNIMerge en un DataFrame, df
  2. Se eliminan todas las visitas (filas del df) con NaN en los diagnósticos
  3. Se eliminan todas las visitas con NaN en la edad del paciente
  4. Se eliminan todas las visitas de pacientes con CN como diagnóstico en su primera visita
  5. Se eliminan todas las visitas de pacientes con CN en su última visita
  6. Se eliminan los pacientes con menos de dos visitas
  7. Se elimina los símbolos " > " y " < " de los strings de marcadores con valor saturado
  8. Se convierte a valores numéricos las cadenas de caracteres los marcadores
  9. Se realiza la clasificación en  $A+$  y  $A-$  y en  $T+$  y  $T-$  (El criterio a seguir se detalla en el siguiente apartado, perfiles AT(N))
- 

Una vez se asegura que todo datos restantes son relevantes para el estudio, y son interpretables por Leaspy, se procede a clasificarlos en los grupos mencionados ante-

riormente. Se calcularán los parámetros estadísticos de cada marcador, incluyendo media, máximo y mínimo. Además se valoran también por separado los pacientes con medidas de CSF para poder estudiar sus biomarcadores.

El código empleado para realizar el preprocesado, clasificación en grupos y cálculos forma parte del script accesible en `"/Demo/Tablas/Tablas.py"`. Como se ve en el anexo D

#### 4.1.2. Perfiles AT(N)

Como ya se ha mencionado anteriormente, los pacientes con patología amiloide (A+) se encuentran dentro del continuo de Alzheimer. Además, presencia de proteína tau (T+) está muy relacionada con la presencia de síntomas de demencia, ya sea por Alzheimer o por otras causas como demencias vasculares [?, ?].

Se elaborarán los perfiles AT de todos los pacientes con datos de CSF. Se considerarán T+ todos los pacientes con  $pTau > 27pg/ml$ . Para la identificación de los pacientes A+ se compararán dos criterios diferentes. El primero, será  $pTau/A\beta_{42} > 0,028$ . El otro criterio que se evaluará es  $A\beta_{42} < 880$  [?].

El código empleado para realizar los perfiles y mostrar la información en una tabla forma parte del script accesible en `"/Demo/Tablas/Tablas.py"`. Como se ve en el anexo D

#### 4.1.3. Tiempo de Conversión y Tiempo de Censura

Para el estudio de la conversión de MCI a demencia es necesario definir una forma de medir el tiempo que tarda en producirse la conversión. Para ello se define el tiempo de conversión como el tiempo que pasa entre el primer diagnóstico de MCI y el primer diagnóstico de demencia. Además, es importante poder conocer el tiempo que los pacientes llevan sin convertir en caso de que no conviertan a demencia. Para pacientes sMCI se define el tiempo de censura como tiempo entre la primera y última visita.

---

##### Algorithm 7 Cálculo de tiempo de conversión

---

1. Se cargan los datos preprocesados del grupo pMCI en un DataFrame, df
  2. Se obtiene el primer diagnóstico de demencia para cada ID
  3. Se calcula la diferencia de fechas entre primera vista y primera demencia
- 

---

##### Algorithm 8 Cálculo de tiempo de censura

---

1. Se cargan los datos preprocesados del grupo sMCI en un DataFrame, df
  2. Se obtiene la última para cada ID
  3. Se calcula la diferencia de fechas entre primera vista y última visita
- 

Se combinará esta información junto con los perfiles AT(N) para comprobar el efecto que tienen estos marcadores en la conversión a demencia.

El código empleado para calcular los tiempos de conversión, tiempo de censura y construir los histogramas, se encuentra en `"/Demo/t\_conversion\_censura/Histogramas.py"`. Como se indica en el anexo D

## 4.2. Modelos de Progresión de Enfermedad: Fase Prodrómica

Tras elaborar el estudio transversal de la población, se realizará un estudio longitudinal de los pacientes. El estudio consistirá en, primero, elaborar todos los modelos posibles con todas las combinaciones posibles de marcadores neuropsicológicos. Tras esto, se evaluarán las mejores combinaciones de marcadores. Por último, se evaluará la robustez del mejor y se construirá la historia natural de cada paciente usando ese modelo.

Para construir los modelos se partirá de la implementación de Leaspy empleada por Jorge Bengoa en su TFG [6] para la fase preclínica de la enfermedad y se adaptará para la fase prodrómica. Para ello será necesario cambiar, también en estos como se realiza el preprocesado de la información de ADNImerge. El preprocesado pasará a ser el mismo que se aplica para el estudio transversal, teniendo en cuenta los requisitos que exige leaspy en su entrada.

Una vez modificado el código, en primer lugar se probará una versión sin paralelizar y con uno número limitado de combinaciones para comprobar que funcione. Tras esto, se probará en Curie una versión paralelizada con un número limitado de combinaciones de marcadores. Por último, tras asegurar que todas las versiones funcionan, se elaborarán todas las combinaciones en Magerit.

### 4.2.1. Paralelización de Tareas

Debido a la gran cantidad de modelos a realizar, se emplearán dos máquinas externas, Curie y Magerit [4]. Estas máquinas contienen varios procesadores capaces de trabajar en paralelo, reduciendo enormemente el tiempo que se tardará en elaborar todos los modelos.

#### Paralelización empleando el estandar MPI y la arquitectura maestro-esclavo

A la hora de paralelizar el código es necesario solucionar el problema de la comunicación y coordinación entre diferentes procesos. Para solucionar el problema de la comunicación entre procesos se empleará el estándar MPI (*Message Passing Interface*). El estándar MPI define la sintaxis de los mensajes que se enviará entre los procesos, así como la semántica de la librería que se emplee para la comunicación. Dado que se trabajará en Python, se utilizarán los paquetes de mpi4py.[9]

Una vez resuelto el problema de la comunicación, el siguiente paso es definir como se coordinarán los diferentes procesos. Para ello, se empleará una arquitectura maestro-esclavo, en la que un proceso (el maestro) actuará como coordinador de

los demás. Los demás procesos inicialmente se encontrarán en espera. Llegado un momento, el maestro enviará a cada proceso esclavo la tarea a realizar así como la información necesaria y activará una bandera indicando que el proceso está ocupado. Al finalizar cada esclavo, estos enviarán al maestro el resultado de la ejecución de su tarea y se desactivará la bandera que indica que el proceso esta ocupado. [6]

Podemos encontrar el código en la carpeta *multileaspy*, como se ve en el anexo D.

### Paralelización en Linux

Tanto Curie como Magerit funcionan en Linux. Para ejecutar el código en las máquinas es necesario crear los procesos primero, luego ejecutar el script en cada uno de ellos.

En Curie se emplea directamente uno de los mandatos de MPI:

- `mpirun -np ntasks python script.py`

En el caso de Magerit-3 se usará un archivo `.sh`. En él se ejecutarán los mandatos para activar anaconda y cargar un entorno, crear y eliminar un directorio para archivos temporales y ejecutar el código, con la ayuda del gestor de tareas que incluye Magerit, y permite programar su ejecución y recibir notificaciones y avisos de su estado.[4]

#### 4.2.2. Leaspy

Como se mencionó anteriormente Leaspy es un paquete de Python para la elaboración de DPMs. Se utilizará para la elaboración de los modelos, usando la instancia de curvas de crecimiento logístico; así como para la representación de resultados.

En primer lugar, es necesario conocer las características que deben tener los datos de entrada al modelo. Hay que detallar también como instanciarlo y configurarlo de acuerdo a nuestras necesidades.

Los datos de entrada deben cumplir las siguientes características [32]:

- Deben contar con una columna llamada **ID** que contenga el índice del individuo.
- Debe contar con una columna **TIME** que contenga la edad del individuo en el momento de la visita.
- Debe haber una columna por variable a estudiar.
- Cada fila se debe corresponder con una visita.
- Los valores de las variables deben ser normalizados entre 0 y 1.
- Los valores de las variables deben, de media, aumentar con el tiempo para cada individuo.
- Se recomiendan al menos dos visitas por individuo para el entrenamiento.
- Las columnas **pueden** contener NaN.

Muchas de estas condiciones ya se aseguraron durante el preprocesado de los datos, pero para poder utilizar el paquete todavía es necesario sustituir el nombre de una de las columnas y normalizar los datos de forma que todos se encuentren entre 0 y 1 de forma ascendente. Antes de la normalización se seguirán los siguientes pasos:

1. Obtener los "DataFrames" (df) con los datos y separar las primeras visitas
2. Separar el df en dos, uno conteniendo los datos de test y otros los de entrenamiento.
3. Identificar los diferentes grupos posibles dentro de los df. Por ejemplo: sMCI, pMCI o demencia.
4. Valorar el número de muestras que posee cada grupo. Se generan nuevos valores de forma aleatoria hasta que todos los grupos tengan el mismo número de muestras.

Este último paso es especialmente importante. No todos los grupos tendrán el mismo número de muestras asociadas. Dado que la transformación emplea la probabilidad acumulada de encontrar un valor concreto del marcador, esto llevará a mayores sesgos y a que algunos valores queden sobrerrepresentados. Al igualar el número de muestras de forma aleatoria, creando, en esencia visitas virtuales, reducimos el impacto de estos sesgos.

Una vez hecho esto, se puede aplicar la transformación. Se calcula la función de probabilidad acumulada de cada marcador. Para ello, se calculan las probabilidades de obtener los diferentes valores que ha tomado el marcador. Tras esto, mediante una interpolación lineal se calcula la función completa. Por último, se emplea la función de distribución para sustituir los valores de los marcadores por su probabilidad acumulada, entre 0 y 1. Es importante guardar los valores que había tomado el marcador antes de la transformación, así como sus probabilidades para poder deshacer la transformación mediante el proceso inverso.

#### 4.2.3. Construcción del modelo

Una vez los datos se ajustan al formato exigido por el paquete Leaspy, se calcularán y guardarán los modelos de la población de la población general para todas las combinaciones de marcadores a estudiar. Una vez ya se hayan elaborado, se guardarán para evitar volver a calcularlos en el futuro. Para elaborar los modelos se empleará una estrategia de bootstrapping, es decir, se tomará un 80 % de la población al azar y se empleará para elaborar un modelo. A este grupo se le llamará grupo de *training*. Después se tomará un 80 % diferente y se elaborará otro modelo nuevo. Se elaborarán diez modelos de la población general diferentes con esta estrategia.

Una vez se tienen los modelos de la población, se realiza la personalización de los modelos para cada individuo. Esa personalización no se realizará solo para el grupo de *training*, también para el 20 % restante, el grupo de *test*. Esto permite comprobar como se comporta el modelo ante individuos que nunca ha visto. Además, como se indico en la sección 2.3.2 se empleará un clasificador KDE-Bayes para identificar a

que grupo (sMCI, pMCI o Demencia) pertenece cada individuo. Se puede encontrar más información de estos clasificadores en el anexo C.

Al ser estas operaciones las más complejas de realizar son las que se paralelizarán, como se desarrolla en la sección 4.2.1.

#### 4.2.4. Evaluación de los Modelos

Una vez se han elaborado y personalizado los modelos se calcularán medidas estadísticas que permitan conocer como se ajusta cada modelo a la realidad. Para ello se calcularán los siguientes parámetros de cada modelo, separando entre grupo de test y de training:

- Coeficiente de Correlación entre el modelo y los datos reales de la reserva cognitiva
- Coeficiente de Correlación entre el modelo y los datos reales de la edad
- Tasa de sMCI identificados correctamente
- Tasa de pMCI identificados correctamente
- Tasa de demencias identificadas correctamente
- Área bajo la curva entre modelo y la realidad

Después se elaborará la media y desviación típica entre los diez modelos elaborados para cada combinación de marcadores. Además, se calculará también el error absoluto medio (Median Absolute Error, MAE) de cada modelo para cada marcador. Por último, se añadirá al código el cálculo del criterio que se usará para elegir el mejor modelo. El criterio será el siguiente:

$$criterio_{test} = \%pmci_{test} \times reserva_{test} \quad (4.1)$$

$$criterio_{train} = \%pmci_{train} \times reserva_{train} \quad (4.2)$$

$$criterio_{total} = \sqrt{criterio_{test} \times criterio_{train}} \quad (4.3)$$

### 4.3. Comprobación de robustez ante datos faltantes (NaN)

Una vez comparados los modelos en condiciones normales, se comprueba la robustez de los modelos ante datos faltantes y visitas NaN (Not a Number). Para ello, el código que se adaptó previamente contiene una función capaz de introducir nuevas visitas con valores NaN, o eliminar visitas viejas. Su funcionamiento es el siguiente. En primer lugar, la función identifica los NaN ya presentes para eliminarlos. Con esto, se garantiza que la cantidad de NaN final en el marcador sea la deseada y que se repartan uniformemente. Una vez hecho esto se calcula el número de visitas con NaN a añadir para cada paciente. Tras esto, se calcula donde los índices de las visitas para distribuirlas uniformemente. Por último se añaden las visitas con valores NaN

en todos los marcadores. Una vez se han modificado las visitas, se procede a realizar un modelo nuevo de la misma manera que se realizó previamente.

Una vez más, el código será una versión adaptada del código de Jorge Bengoa [?], pero en este caso no será necesario realizar ninguna modificación, ya que se realizaron en las secciones anteriores.

#### 4.4. Historia Natural

Una vez valorados tanto cual es el mejor modelo como la resistencia a valores NaN se pasará a calcular la historia natural de los pacientes y valorar como de correcta es dicha historia natural. La historia natural define las trayectorias de los marcadores y la evolución de la enfermedad a lo largo del tiempo para cada paciente.

Para elaborar la historia natural primero se procederá a la personalización de la trayectoria de la población general para cada paciente. Además, se obtendrá el tiempo de corte para cada paciente a partir del cual se considera que entran en demencia. Usando este tiempo, se pasará de hablar de la edad del paciente a su tiempo de conversión,  $\tau$ , que es el tiempo que queda hasta el inicio de la demencia,  $\tau < 0$ , o que ha pasado  $\tau > 0$  desde el inicio.

Una vez calculado todo esto, se empleará para elaborar las trayectorias de los marcadores a lo largo del tiempo para cada paciente. Se elaborará también un histograma que muestre los tiempos de conversión de cada paciente al inicio del estudio en función del grupo al que pertenecen. Por último, se empleará los tiempos de conversión para comprobar si las estimaciones son correctas. Es decir, un sMCI tiene que tener un  $\tau < 0$  en su última visita; un pMCI tiene que empezar con  $\tau < 0$  en la primera y acabar con  $\tau > 0$ ; y un paciente con demencia tiene que tener  $\tau > 0$  en la primera visita.

Como en los apartados anteriores, se trabajará partiendo de la versión para la fase preclínica de Jorge Bengoa [6]. Se tendrán que modificar no solo nombres de variables, sino también incluir el preprocesado realizado anteriormente, modificar nombres de las gráficas, y verificar las funciones que calculan los tiempos, donde también hay que modificar los nombres de los grupos a buscar en el DataFrame. Por último, como su código buscaba también comparar con Grace, será necesario comentar lo referente a Grace y adaptar las funciones que necesiten los datos de los dos. Además, para el cálculo de las tasas de éxito, se usará un código de Matlab, pensado inicialmente para RPDPM y modificado ligeramente para aceptar los datos exportados a .mat desde python. Para ello, será necesario incluir también código para exportar esta información en los scripts de Python.

Se puede encontrar el cálculo de los tiempos y guardado de la información, tanto para Python como para Matlab en el archivo *estimateConversionTime.py*. La elaboración de los histogramas se encuentra en *plotReserveDistribution.py* y la de las trayectorias en *plotTimeline.py*. Por último, el código de matlab usado para el cálculo de las métricas se encuentra en *distribution\\_tau\\_MCI2AD\\_RP.m*. Como

se indica en el anexo D

## 4.5. Pronostico

Como ya se mencionó anteriormente, la característica más útil de estos modelos es la capacidad de emplearlos para elaborar pronósticos sobre la evolución de la enfermedad. Más específicamente, se pueden utilizar los modelos para predecir el inicio de la enfermedad. Para evaluar la capacidad de pronóstico de los modelos elaborados con Leaspy se hará lo siguiente.

Antes del proceso de personalización se eliminarán los datos de las visitas en el plazo a pronosticar antes del proceso de personalización. En este caso, se intentará pronosticar primero para el último año. Posteriormente se aumentará la ventana de tiempo, año a año, hasta alcanzar los cinco. Tras esto, se realizará la personalización de los modelos, como se hizo anteriormente. Utilizando los modelos de cada paciente, se estimará el grupo al que pertenece el paciente.

Para valorar la calidad de la predicción, las predicciones se agruparán en función del tiempo pasado desde el último dato introducido al modelo, en intervalos de un año y con un mínimo de treinta pacientes. Se utilizarán como métricas para la valoración el área bajo la curva del clasificador, las tasas de detección de los diferentes grupos y el error absoluto medio de los marcadores en el plazo estimado. [36]

Para este primer experimento se compararán los resultados con los de un predictor constante, es decir, un algoritmo que asigna a la siguiente visita el mismo diagnóstico que el de la visita anterior; y con los de RPDPM.

Además se realizará un segundo experimento específicamente para valorar la capacidad de detectar la conversión a demencia. Se tomarán a los pacientes pMCI y se eliminarán todas las visitas con diagnóstico de demencia. De esta forma, se puede comprobar si el pronóstico es capaz de identificar correctamente el paso a demencia. Para valorar esta capacidad se utiliza la correlación entre el tiempo de conversión real y el estimado, y la capacidad de identificar a los pacientes como pMCI. [36]

Una vez más, se partirá del código de Jorge Bengoa, aunque se eliminará lo referente a Grace y lstm [6].

El código para elaborar las predicciones con el predictor constante se encuentra en *future\\_constant\\_prediction.py*. Por otra parte, las predicciones en leaspy se realizan en *future\\_leaspy\\_prediction.py*, y la visualización de todos los resultados en *future\\_prediction\\_compare.py*. Como se indica en el anexo D

# Capítulo 5

## Resultados

### 5.1. Estudio Transversal

#### 5.1.1. Estudio transversal de cada individuo en su primera visita

Las siguientes tablas detallan la información de los marcadores de los pacientes agrupados en sMCI, pMCI y demencia. Toda la información incluye los valores mínimos y máximos encontrados. La primera tabla contienen a todos los individuos del estudio. La segunda contiene unicamente a pacientes con medidas de CSF.

#### 5.1.2. Perfiles AT(N)

Se observa que un 56.8 % de los pacientes sMCI tenían un perfil A-T-. Entre los pMCI este perfil solo se encuentra en un 16.5 % de los pacientes, mientras que entre los pacientes con demencia la cantidad cae al 7.9 %. En cuanto a los pacientes con A+, se observa en un 38 % de los sMCI, un 81 % de los pMCI y un 89.3 % de los pacientes con demencia. Por último, un 68.7 % de las demencias tienen perfil A+T+, bajando en los pMCI al 63.2 % y cayendo a un 25.9 % de los sMCI.

La cantidad relativa de pacientes que pertenecen a cada perfil, dada en porcentajes se detalla en la siguiente tabla:

| Grupo Clínico | Criterio Seguido para A+ | A+     | A-T-   | A-T+   | A+T-   | A+T+   |
|---------------|--------------------------|--------|--------|--------|--------|--------|
| sMCI          | PTAU/ABETA > 0.028       | 38.0 % | 56.8 % | 5.2 %  | 12.0 % | 25.9 % |
|               | ABETA < 880              | 42.0 % | 46.3 % | 11.7 % | 22.5 % | 19.4 % |
| pMCI          | PTAU/ABETA > 0.028       | 81.0 % | 16.5 % | 2.6 %  | 17.7 % | 63.2 % |
|               | ABETA < 880              | 77.5 % | 11.7 % | 10.8 % | 22.5 % | 55.0 % |
| Demencia      | PTAU/ABETA > 0.028       | 89.3 % | 7.9 %  | 2.8 %  | 20.6 % | 68.7 % |
|               | ABETA < 880              | 86.0 % | 6.1 %  | 7.9 %  | 22.4 % | 63.6 % |

Tabla 5.3: Perfiles AT de los pacientes con muestras CSF, clasificados en sMCI, pMCI y demencia. Se valoran dos criterios diferentes para A+:  $pTau/A\beta_{42} > 0,028$  y  $A\beta_{42} < 880$ . En los dos casos se sigue el criterio  $pTau > 27pg/ml$  para valorar T+. [37]

#### 5.1.3. Tiempo de conversión y Tiempo de Censura

A continuación se pueden observar los histogramas de los tiempos de censura y tiempos de conversión de los pacientes sMCI y pMCI respectivamente.

Tabla 5.1: Estudio Transversal - Todos los sujetos

| Grupo          | sMCI                          | pMCI                          | Demencia                       | Significancia |
|----------------|-------------------------------|-------------------------------|--------------------------------|---------------|
| Pacientes      | 551                           | 369                           | 358                            | -             |
| Mujeres        | 218 (39.6%)                   | 150 (40.7%)                   | 157 (43.9%)                    | -             |
| APOE4          | 228 (41.4%)                   | 235 (63.7%)                   | 243 (67.9%)                    | ab            |
| Edad           | 73.0 (7.6)<br>(55.0 - 91.4)   | 73.9 (7.2)<br>(55.0 - 89.5)   | 74.6 (7.8)<br>(55.1 - 90.9)    | b             |
| Educación      | 15.9 (2.8)<br>(4 - 20)        | 16.0 (2.7)<br>(6 - 20)        | 15.2 (2.9)<br>(4 - 20)         | bc            |
| MMSE           | 27.8 (1.8)<br>(22.0 - 30.0)   | 27.0 (1.8)<br>(19.0 - 30.0)   | 23.2 (2.1)<br>(16.0 - 30.0)    | abc           |
| CDRSB          | 1.3 (0.9)<br>(0.0 - 6.0)      | 1.9 (0.9)<br>(0.5 - 5.0)      | 4.4 (1.7)<br>(1.0 - 10.0)      | abc           |
| R-Immediate    | 36.4 (10.6)<br>(13.0 - 68.0)  | 29.5 (8.3)<br>(7.0 - 67.0)    | 22.7 (7.1)<br>(1.0 - 46.0)     | abc           |
| ADAS13         | 14.7 (5.9)<br>(0.67 - 36.33)  | 20.5 (6.3)<br>(4.0 - 39.67)   | 29.8 (7.9)<br>(12.67 - 54.67)  | abc           |
| FAQ            | 2.2 (3.3)<br>(0.0 - 24.0)     | 5.2 (4.8)<br>(0.0 - 22.0)     | 13.2 (6.9)<br>(0.0 - 30.0)     | abc           |
| LDELTOTAL      | 6.7 (3.2)<br>(0.0 - 20.0)     | 4.0 (3.2)<br>(0.0 - 18.0)     | 1.4 (2.0)<br>(0.0 - 11.0)      | abc           |
| PACCTrailsB    | -4.8 (3.4)<br>(-13.6 - 2.8)   | -7.9 (3.4)<br>(-17.8 - 1.9)   | -14.1 (3.1)<br>(-23.6 - -5.2)  | abc           |
| TRABSCOR       | 107.4 (57.2)<br>(6.0 - 300.0) | 136.6 (75.0)<br>(0.0 - 300.0) | 195.9 (87.6)<br>(42.0 - 300.0) | abc           |
| A $\beta$      | -                             | -                             | -                              | -             |
| TAU            | -                             | -                             | -                              | -             |
| pTAU           | -                             | -                             | -                              | -             |
| pTAU/A $\beta$ | -                             | -                             | -                              | -             |
| FDG            | -                             | -                             | -                              | -             |

Tabla 5.2: Estudio Transversal - Solo sujetos con CSF

| Grupo          | sMCI                                | pMCI                                | Demencia                            | Significancia |
|----------------|-------------------------------------|-------------------------------------|-------------------------------------|---------------|
| Pacientes      | 324                                 | 231                                 | 214                                 | -             |
| Mujeres        | 131 (40.4%)                         | 91 (39.4%)                          | 90 (42.1%)                          | -             |
| APOE4          | 137 (42.3%)                         | 146 (63.2%)                         | 149 (69.6%)                         | ab            |
| Edad           | 72.5 (7.5)<br>(55.0 - 91.4)         | 73.5 (7.3)<br>(55.0 - 88.4)         | 74.4 (8.1)<br>(55.9 - 90.3)         | b             |
| Educación      | 16.0 (2.8)<br>(8 - 20)              | 16.0 (2.8)<br>(6 - 20)              | 15.4 (2.9)<br>(4 - 20)              | bc            |
| MMSE           | 28.0 (1.7)<br>(23.0 - 30.0)         | 27.1 (1.8)<br>(23.0 - 30.0)         | 23.4 (2.0)<br>(19.0 - 27.0)         | abc           |
| CDRSB          | 1.3 (0.7)<br>(0.5 - 5.5)            | 1.9 (1.0)<br>(0.5 - 5.0)            | 4.4 (1.7)<br>(1.0 - 10.0)           | abc           |
| R-Immediate    | 36.9 (10.6)<br>(13.0 - 68.0)        | 29.5 (8.0)<br>(11.0 - 59.0)         | 22.7 (6.9)<br>(1.0 - 46.0)          | abc           |
| ADAS13         | 14.2 (5.9)<br>(3.0 - 36.33)         | 20.3 (6.4)<br>(7.0 - 38.0)          | 29.7 (7.9)<br>(15.33 - 54.67)       | abc           |
| FAQ            | 2.1 (3.4)<br>(0.0 - 21.0)           | 4.9 (4.6)<br>(0.0 - 22.0)           | 13.1 (7.0)<br>(0.0 - 28.0)          | abc           |
| LDELTOTAL      | 7.1 (2.9)<br>(0.0 - 14.0)           | 4.1 (3.3)<br>(0.0 - 11.0)           | 1.4 (1.8)<br>(0.0 - 8.0)            | abc           |
| PACCTrailsB    | -4.5 (3.3)<br>(-13.1 - 2.8)         | -7.8 (3.5)<br>(-15.7 - 0.5)         | -14.1 (3.0)<br>(-21.2 - -6.7)       | abc           |
| TRABSCOR       | 106.3 (55.6)<br>(33.0 - 300.0)      | 132.9 (73.9)<br>(38.0 - 300.0)      | 198.0 (87.0)<br>(42.0 - 300.0)      | abc           |
| $A\beta$       | 1073.6 (438.0)<br>(210.9 - 1700.0)  | 736.2 (322.8)<br>(267.2 - 1700.0)   | 664.0 (317.5)<br>(212.3 - 1700.0)   | abc           |
| TAU            | 257.0 (113.2)<br>(97.89 - 851.8)    | 344.0 (137.0)<br>(101.2 - 816.9)    | 366.7 (146.8)<br>(80.0 - 851.6)     | ab            |
| pTAU           | 24.4 (12.7)<br>(8.21 - 103.0)       | 34.5 (15.6)<br>(8.72 - 92.08)       | 36.5 (15.8)<br>(8.0 - 94.86)        | ab            |
| pTAU/ $A\beta$ | 0.029 (0.024)<br>(0.006 - 0.192)    | 0.055 (0.032)<br>(0.008 - 0.237)    | 0.064 (0.034)<br>(0.008 - 0.245)    | abc           |
| FDG            | n = 282<br>1.3 (0.1)<br>(0.8 - 1.7) | n = 179<br>1.2 (0.1)<br>(0.9 - 1.6) | n = 166<br>1.0 (0.1)<br>(0.6 - 1.5) | abc           |

En primer lugar se puede observar que los entre los pMCI el tiempo de conversión de la mayor parte de los pacientes, 78,32% ( $n = 289/369$ ), convierte en los primeros cuatro años. Además, se observa que aproximadamente un 72,7% ( $n = 168/231$ ) de los pacientes pMCI con medidas de líquido cefalorraquídeo (CSF) tienen patología amiloide (A+). Observando únicamente a los pacientes pMCI con CSF, un 76,2% (176/231) convirtió a demencia en menos de cuatro años. De ellos, un 76,1% (134/176) tenían patología amiloide. Por otra parte, solo un 31,2% (101/324) de los pacientes sMCI con medidas de CSF son A+.

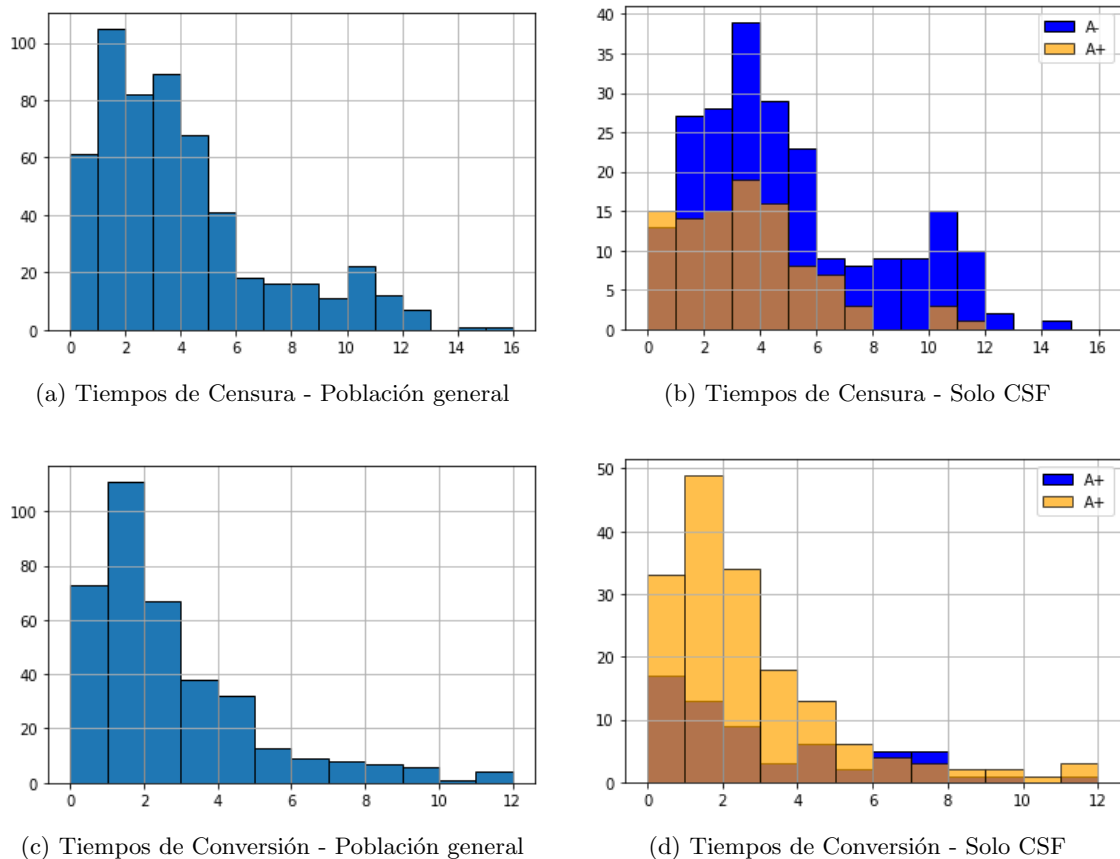


Figura 5.1: Comparación de tiempos de censura y conversión en diferentes poblaciones

## 5.2. Modelos Población General

Los resultados se mostrarán en la siguiente tabla, que contiene los marcadores que usa el modelo, correlación de la reserva cognitiva, la capacidad de clasificar los pacientes en grupos y el área bajo la curva para los pacientes de test. Además, se compararán con los mejores modelos obtenidos en RPDPM y Leaspy en otras publicaciones similares. Para simplificar la tabla, se emplearán las siguientes abreviaturas para cada marcador:

- **RF**: RAVLT forgetting
- **RI**: RAVLT immediate

- **RL**: RAVLT learning
- **F**: FAQ
- **M**: MMSE
- **C**: CDRSB
- **D**: mPACCtrailsB
- **T**: mPACCtrailsB
- **TR**: TRABSCOR

En primer lugar, se recoge en una tabla los resultados para los datos de test de la reserva cognitiva, área bajo la curva, el criterio de selección del modelo y las tasas de acierto de diagnóstico de sMCI, pMCI y demencia.

|                               | Marcadores      | Reserva | %sMCI  | %pMCI  | %Demencia | AUC   | Criterio |
|-------------------------------|-----------------|---------|--------|--------|-----------|-------|----------|
| Leaspy                        | M, C            | 0.90    | 87.71  | 78.04  | 79.48     | 0.94  | 67.80    |
|                               | RL, F, M, C     | 0.91    | 86.77  | 76.75  | 74.16     | 0.94  | 67.50    |
|                               | F, C, D         | 0.91    | 86.25  | 79.40  | 76.49     | 0.94  | 67.08    |
| Leaspy<br>(Hyaee<br>Moon)[33] | RF, M, C        | 0.895   | 94.83  | 76.757 | 72.436    | 0.966 | 68.7     |
|                               | F, M, C, D, TR  | 0.871   | 92.119 | 78.649 | 69.487    | 0.954 | 68.48    |
|                               | RL, F, C, D, TR | 0.888   | 92.881 | 77.027 | 69.872    | 0.956 | 68.40    |
| RPDPM                         | RF, RL, M, C    | 0.89    | 90.71  | 82.98  | 74.62     | 0.95  | 73.65    |
|                               | RF, RI, M, C    | 0.89    | 91.16  | 83.43  | 74.03     | 0.96  | 73.22    |
|                               | RI, RL, M, C    | 0.90    | 91.79  | 82.84  | 75.52     | 0.95  | 72.90    |

Tabla 5.4: Comparación de las mejores combinaciones de marcadores

Se mostrará también el error absoluto medio (Mean Absolute Error, MAE) de los marcadores comunes a los mejores modelos

|       | Leaspy |             |         | RPDPM        |              |              |
|-------|--------|-------------|---------|--------------|--------------|--------------|
|       | M, C   | RL, F, M, C | F, C, D | RF, RL, M, C | RF, RI, M, C | RI, RL, M, C |
| MMSE  | 1,479  | 1,554       | -       | 1,587        | 1,553        | 1,560        |
| CDRSB | 0,690  | 0,833       | 0,852   | 0,636        | 0,669        | 0,687        |

Tabla 5.5: Comparación del MAE de las mejores combinaciones de marcadores. Resultados de los datos de test

### 5.3. Comprobación de robustez ante datos faltantes (NaN)

Además, se comprueba la robustez ante NaN para el mejor modelo construido

|                        | % NaN añadidos | Reserva | %sMCI | %pMCI | %Demencia | AUC  |
|------------------------|----------------|---------|-------|-------|-----------|------|
| Leaspy<br>(M, C)       | -              | 0.90    | 87.71 | 78.04 | 79.48     | 0.94 |
|                        | 20             | 0.89    | 88.44 | 77.23 | 80.78     | 0.94 |
|                        | 40             | 0.87    | 87.39 | 78.67 | 78.96     | 0.93 |
| RPDPM<br>(RL,<br>M, C) | -              | 0.89    | 90.7  | 80.7  | 75.2      | 0.97 |
|                        | 20             | 0.87    | 90,2  | 68,8  | 74,8      | 0.95 |
|                        | 40             | 0.88    | 90,2  | 71,5  | 72,2      | 0.94 |

Tabla 5.6: Comparación de la robustez a datos NaN añadidos de la mejor combinación de marcadores para los modelos realizados en leaspy y un modelo de referencia en RPDPM

Así como el MAE de sus marcadores para cada caso:

|       | Leaspy (M, C) |       |       | RPDPM (RL, M, C) |       |       |
|-------|---------------|-------|-------|------------------|-------|-------|
|       | -             | 20    | 40    | -                | 20    | 40    |
| MMSE  | 1,479         | 1,505 | 1,509 | 1,520            | 1,463 | 1,338 |
| CDRSB | 0,690         | 0,698 | 0,705 | 0,603            | 0,577 | 0,567 |

Tabla 5.7: Comparación del MAE Para diferentes valores de NaN. Mejor modelo de Leaspy y mejor modelo de RPDPM. Resultados de los datos de test

## 5.4. Historia Natural

En primer lugar se mostrarán las trayectorias de los marcadores del modelo:

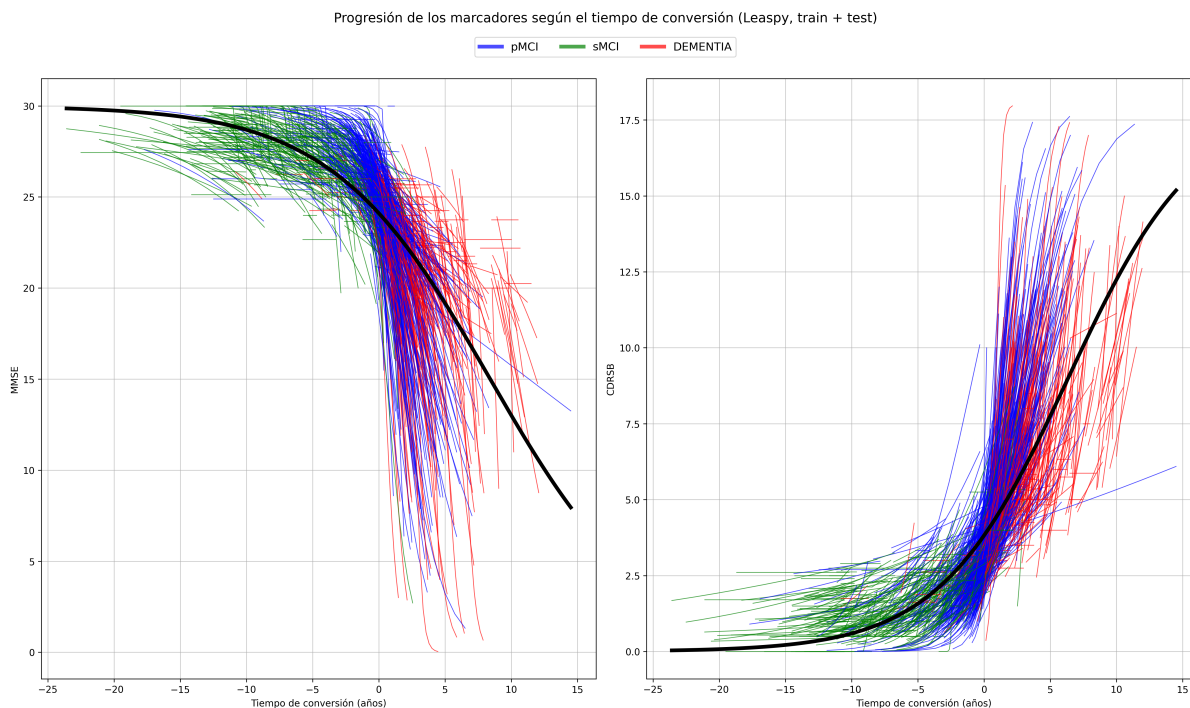


Figura 5.2: Trayectorias de los marcadores - Lespy - train+test

|       | sMCI  | pMCI  | Demencia | clinical cutoff | Leaspy cutoff |
|-------|-------|-------|----------|-----------------|---------------|
| MMSE  | 27.81 | 27.02 | 23.22    | 20              | 24.11         |
| CDRSB | 1.35  | 1.89  | 4.38     | 4.5             | 3.82          |

Tabla 5.8: Comparación de las mejores combinaciones de marcadores

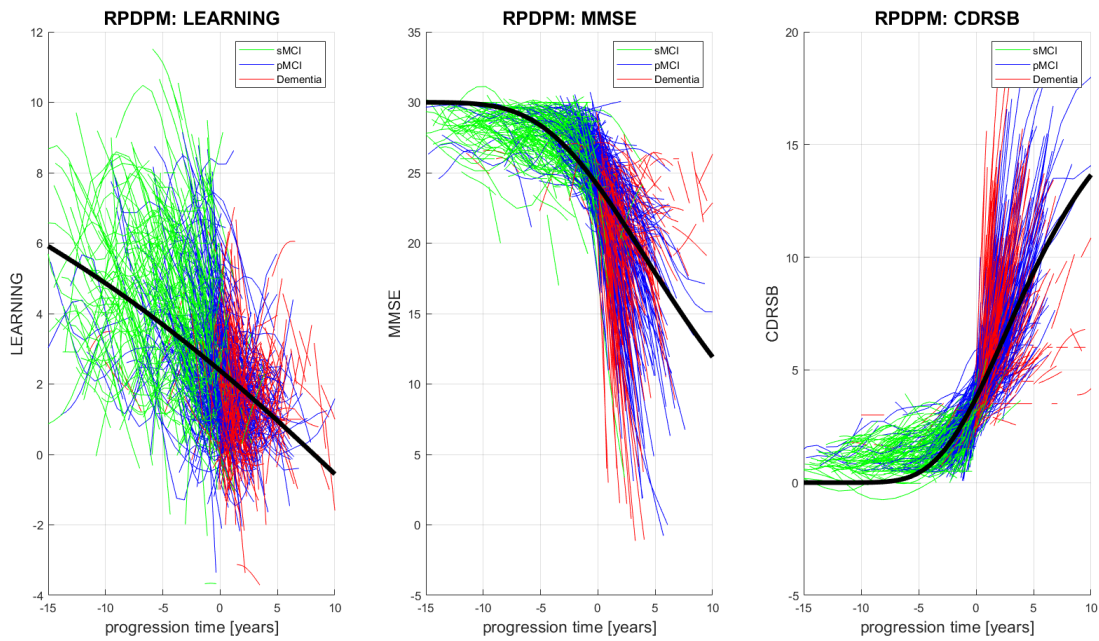


Figura 5.3: Trayectorias de los marcadores - RPDPM - train+test

Pasando a los histogramas de los grupos y su reserva:

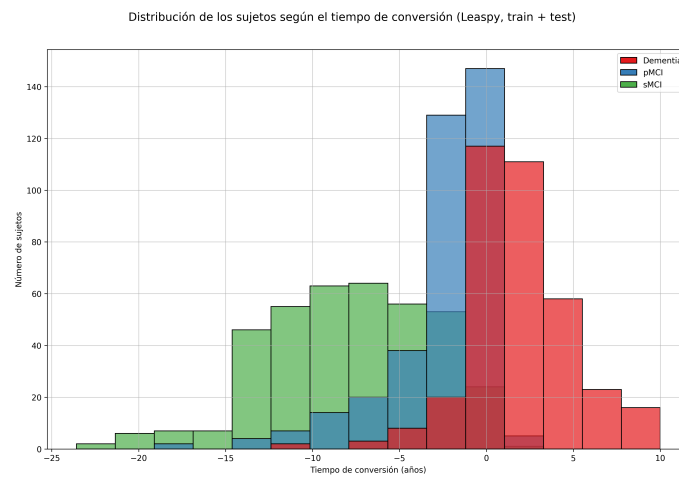


Figura 5.4: Reserva cognitiva de individuos sMCI, pMCI y Demencia - train+test - Leaspy

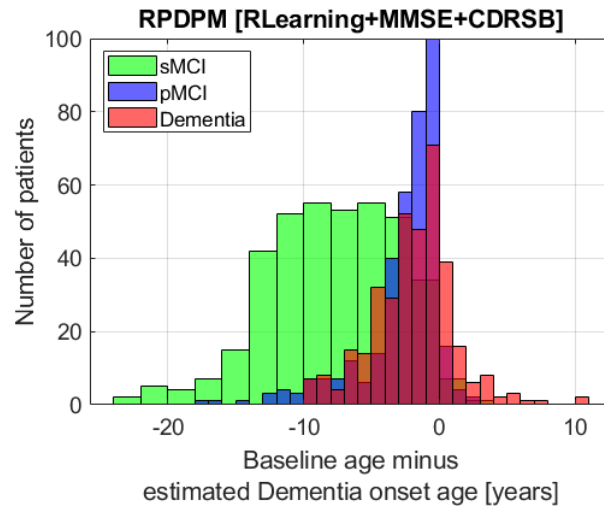


Figura 5.5: Reserva cognitiva de individuos sMCI, pMCI y Demencia - train+test - RPDPM

Por último, en cuanto a la evaluación de la exactitud de la predicción:

|        | Marcadores | %sMCI | %pMCI | %Demencia | %sMCI exitoso | %pMCI exitoso | Demencia exitoso |
|--------|------------|-------|-------|-----------|---------------|---------------|------------------|
| Leaspy | M, C       | 69.6  | 99.2  | 100.0     | 97.9          | 94.0          | 86.3             |
| RPDPM  | RL, M, C   | 49.4  | 98.8  | 96.2      | 82.8          | 92.6          | 90.8             |

Tabla 5.9: Evaluación del corte al inicio de la enfermedad para cada marcador. Comparación entre Leaspy y la medida real

### 5.5. Pronostico

En primer lugar, se muestran los resultados del MAE del primer experimento:

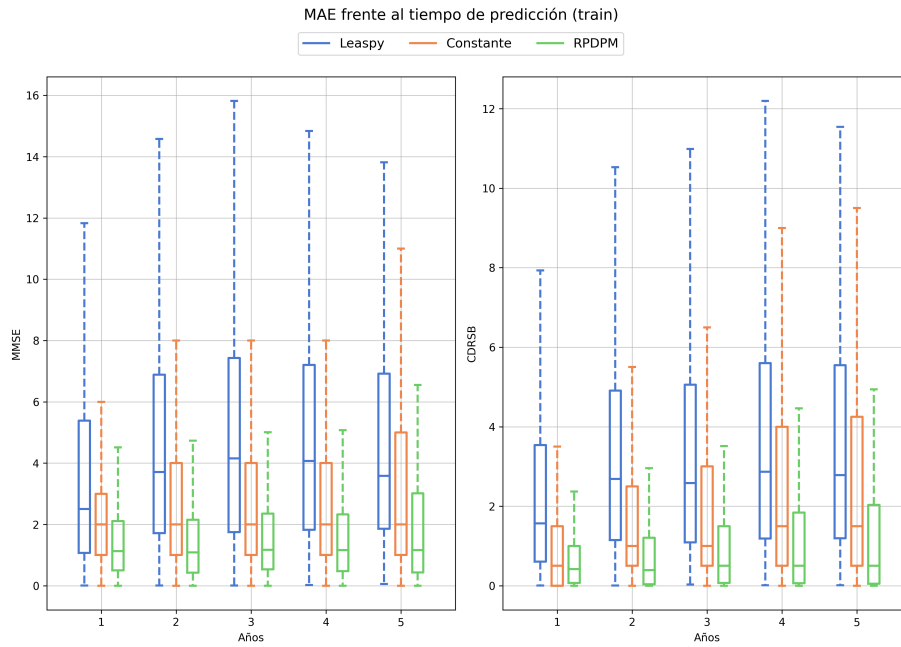


Figura 5.6: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - train

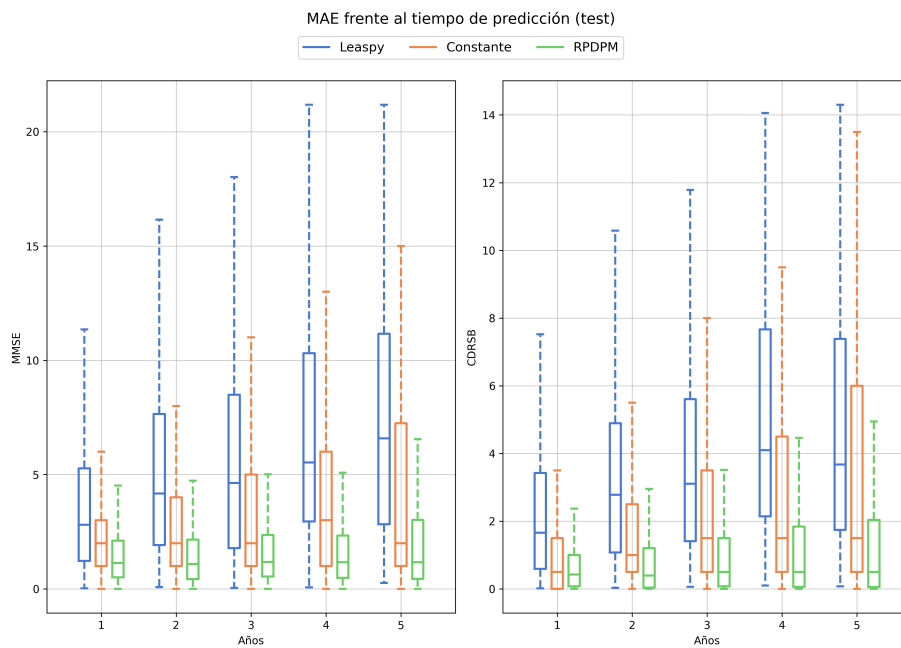


Figura 5.7: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - test

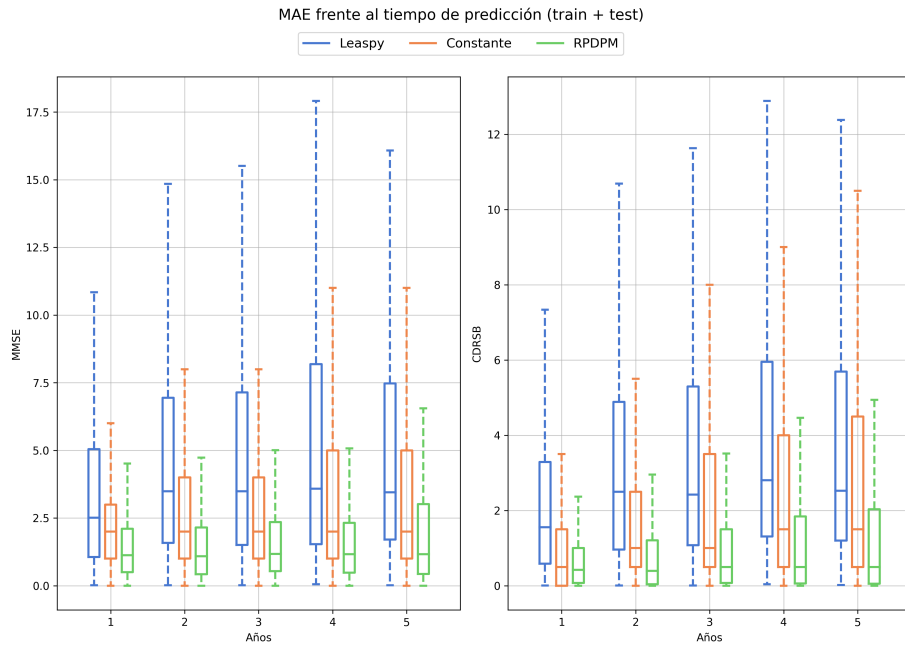


Figura 5.8: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - MAE - Todos los pacientes

Los resultados del resto de métricas de calidad para el primer experimento son los siguientes:

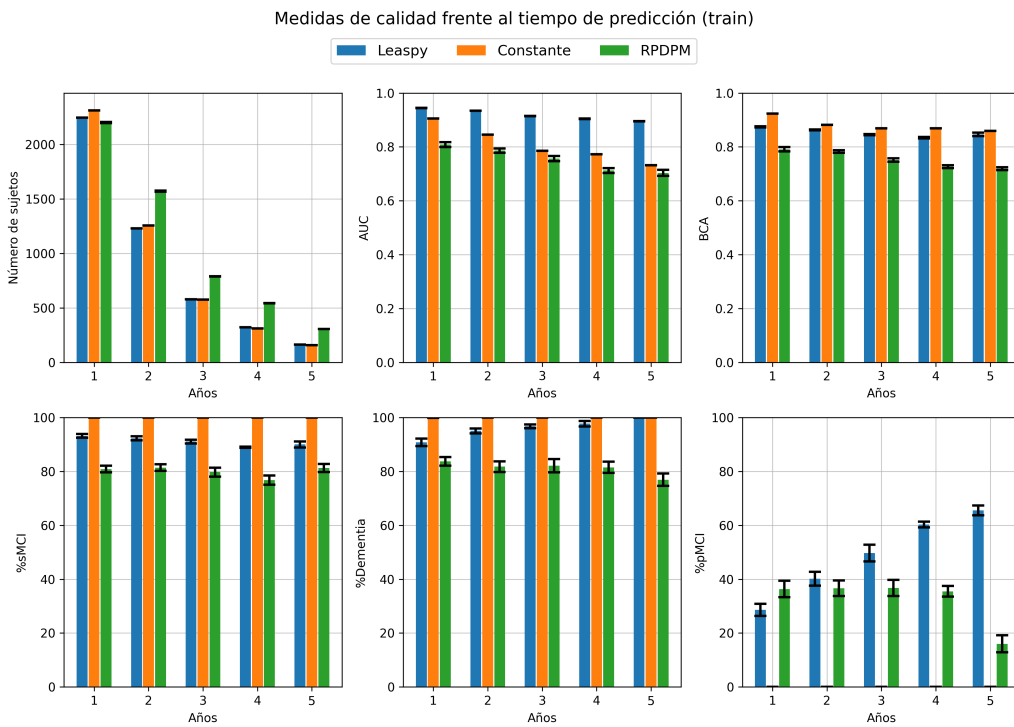


Figura 5.9: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Train

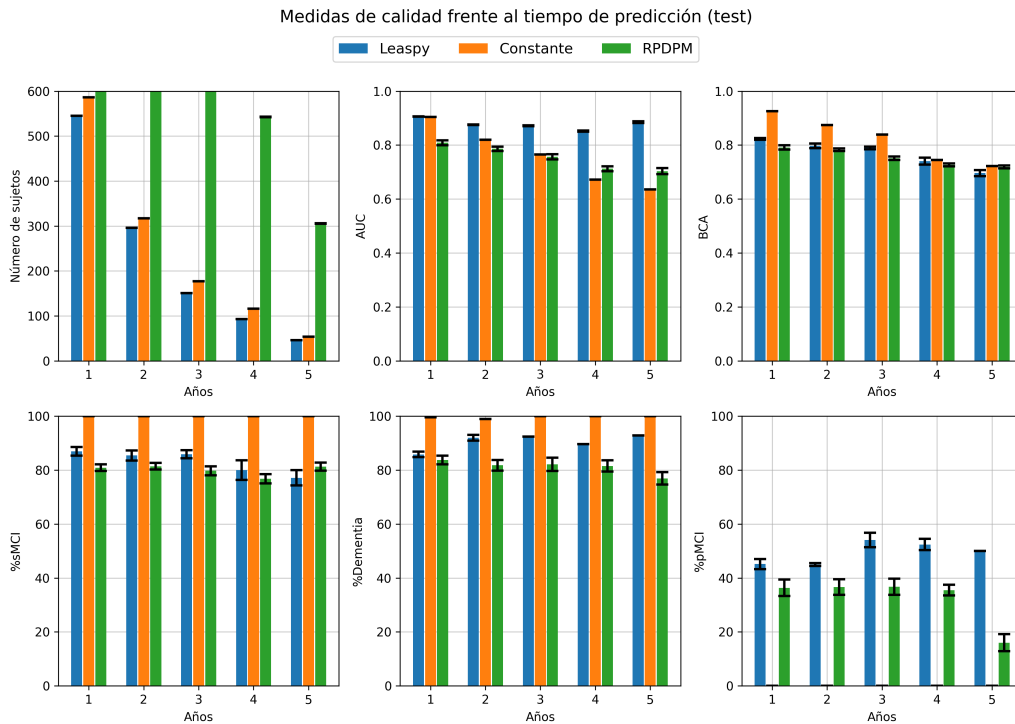


Figura 5.10: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Test

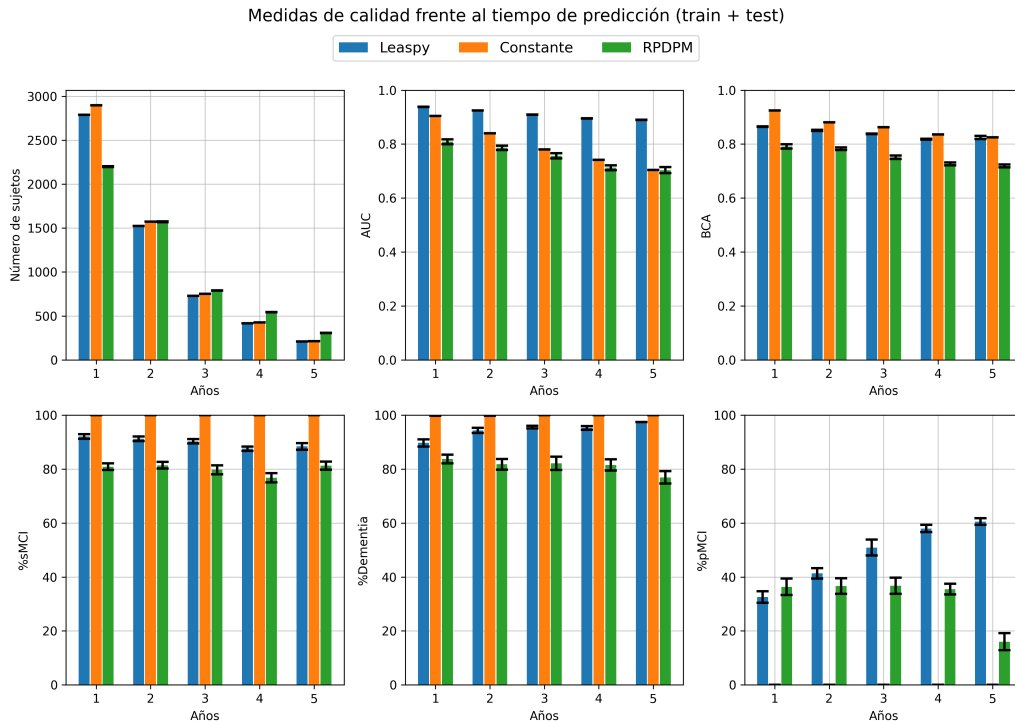


Figura 5.11: Pronóstico, todos los grupos de diagnóstico, comparado con un predictor constante - Medidas de Calidad - Todos los pacientes

En cuanto al segundo experimento, los resultados son los siguientes:

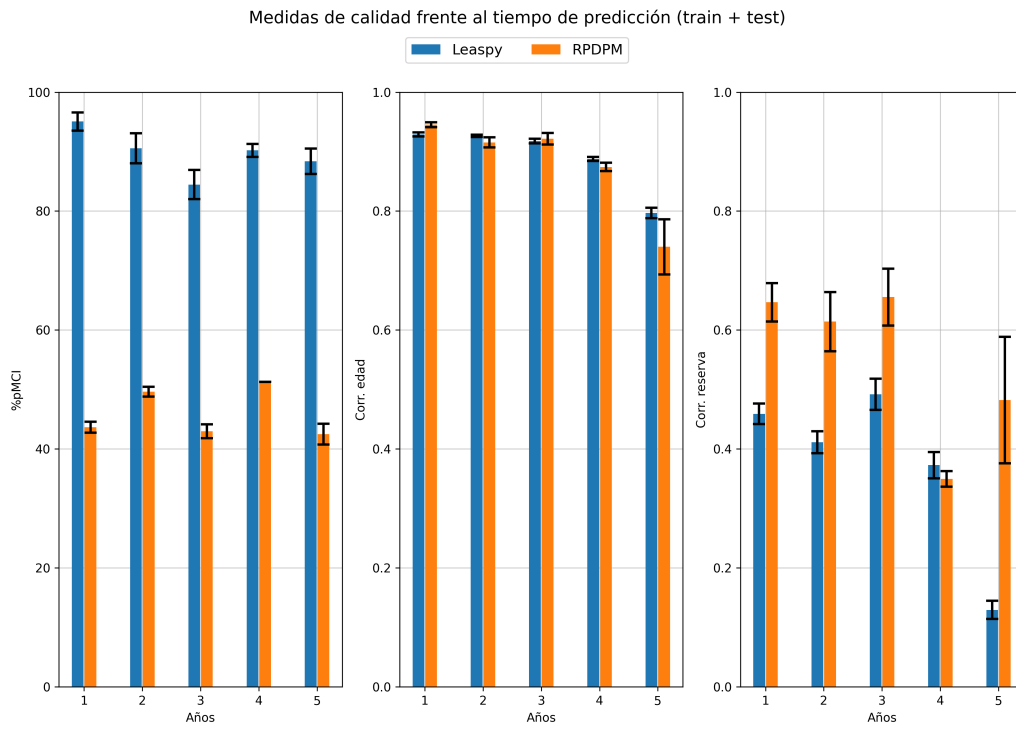


Figura 5.12: Pronostico, experimento de detección de conversión - Medidas de Calidad - Todos los pacientes

# Capítulo 6

## Discusión

### 6.1. Estudio Transversal

Observando los valores de los diferentes marcadores y factores de riesgo presentes en las tablas 5.1 y 5.2 podemos comprobar que hay diferencias significativas entre la presencia de APOE4 de los pacientes sMCI y pMCI, y entre los sMCI y los pacientes de demencia. Se observa también diferencias significativas en la edad entre pacientes sMCI y pacientes de demencia; y en la educación entre pacientes sMCI y pacientes de demencia y entre pacientes pMCI y pacientes de demencia. En cuanto a los marcadores neuropsicológicos, hay diferencias significativas entre todos los grupos para todos los marcadores, indicando su validez para elaborar el diagnóstico. Por último, en cuanto a los marcadores CSF se observa diferencias significativas entre todos los grupos para el ABETA y el FDG, y entre sMCI y pMCI, y entre sMCI y demencia para la proteína Tau.

Al comparar los perfiles AT(N) de los diferentes grupos, presentes en la tabla 5.3 se puede observar con facilidad que los pacientes con A+ convierten más a demencia que aquellos con A-. Además, la baja aparición de perfiles A-T+ parece indicar que la mayor parte de la muestra padecería AD frente a otras patologías como la demencia vascular. Por último, se observa que los pacientes con demencia o que convertirán a demencia tienen T+, reflejando la importancia de este marcador como indicador de los síntomas.

En cuanto a los tiempos de conversión, en los histogramas de las figuras 5.1c y 5.1d se observa que los pacientes con A+ tienden a convertir más rápido que los pacientes con A-, convirtiendo la mayoría de ellos en los primeros cuatro años. Se observa en los histogramas de los tiempos de censura, en las figuras 5.1a y 5.1b, que todos los pacientes tienden a convertir en los primeros 5 años. Juntando todo esto, se puede deducir una relación entre la presencia de A+ y la velocidad de conversión de a demencia.

### 6.2. Modelos Población General

En primer lugar, se puede observar en la tabla 5.4 como la nueva implementación de Leaspy parece obtener unos modelos con un criterio algo más bajo que la implementación anterior, y bastante más baja que la de RPDPM. Sí se puede observar

que la mejor combinación de marcadores para esta implementación es además la que menos marcadores exige, simplificando su implementación en práctica clínica. Además, como cabría esperar, el CDRSB que además es el marcador utilizado para el diagnóstico, se encuentra presente en todos los mejores modelos.

En cuanto a métricas concretas de cada modelo, se puede observar como los nuevos modelos obtienen en general mejor detección de pMCI y demencias, pero peor de sMCI. Tienen también una mejor correlación de la reserva cognitiva y un peor área bajo la curva. Cabe también resaltar que las diferencias con los otros modelos de Leaspy se deben con toda probabilidad a la variabilidad propia de estos algoritmos no deterministas. En cuanto a la comparación con RPDPM, la detección de sMCI y pMCI es peor, pero detecta mejor la demencia. Leaspy tiene mejor correlación de la reserva cognitiva, pero por área bajo la curva. Los resultados de la comparación con RPDPM son muy similares a los de los mejores modelos anteriores. Por último, en cuanto al MAE, en la tabla 5.5 se observa que Leaspy comete en general menos error con los valores de MMSE, pero mayor error con los valores de CDRSB.

### 6.3. Comprobación de robustez ante datos faltantes (NaN)

Centrando la atención en la tabla 5.6, se observa que al aumentar la cantidad de NaN en el modelo de Leaspy el área bajo la curva y los porcentajes de detección apenas varían. Sin embargo, la correlación con la reserva cae ligeramente. Por otro lado en el caso de RPDPM se observa que los porcentajes de detección de sMCI y Demencia caen ligeramente, y el de pMCI cae de 80 % a un 70 %. Se produce una ligera caída también en el área bajo la curva y la reserva. Por esta razón, se puede concluir que Leaspy parece ser más robusto especialmente en la detección de pMCIs, que es a la vez el grupo más interesante y difícil de detectar.

En cuanto a el MAE, en la tabla 5.6 se observa que el el modelo de Leaspy tiende a crecer ligeramente para los dos marcadores. Por otra parte en RPDPM observamos que decrece ligeramente. En definitiva, la estimación de los valores concretos de los marcadores parece ser más robusta en RPDPM.

### 6.4. Historia Natural

Centrando la atención en las trayectorias de los marcadores, en la figura 5.2, lo primero que se observa es que, como se esperaba, decrecen o crecen (dependiendo de si son marcadores crecientes o decrecientes) más rápido en los pacientes de los grupos pMCI que en los sMCI, y más rápido en los pacientes de demencia que en los pMCI. Además, si observamos el valor del marcador para la trayectoria media de la población en el inicio de la enfermedad, se observa que, aunque se aleja ligeramente del cutoff clínico, es aproximadamente correcto. En el caso del MMSE, el real es 20 frente a 24 que estima leaspy. En el de CDRSB, el real es 4,5 frente a 3,8 que estima Leaspy. Además, comparando con los resultados de RPDPM de la figura 5.3, se puede observar como el modelo de Leaspy diferencia mejor las trayectorias del marcador

CDRSB de los pacientes pMCI y demencia. En cuanto al cutoff, en las gráficas se observa que RPDPM devuelve un cutoff muy similar al ofrecido por Leaspy. Por último, observando únicamente los resultados de RPDPM se puede observar que el marcador adicional que incluye su mejor modelo, el Learning, ofrece trayectorias que se alejan mucho de la trayectoria de la población general. Si permiten diferenciar al sMCI de pMCI y demencia, pero no diferencian bien pMCI de demencia. En base a esto, y teniendo en cuenta que es preferible valorar la menor cantidad de marcadores posibles en el modelo para facilitar su implementación clínica en un futuro, se puede considerar que el modelo de Leaspy es ligeramente mejor, al menos en lo que se refiere a las trayectorias.

Por otra parte, observando el histograma de los tiempos de conversión, en la figura 5.4 se puede observar que como es de esperar los pacientes sMCI concentran su reserva en la primera visita mucho antes del tiempo de corte. Los pacientes pMCI concentran su reserva en el momento de la visita alrededor del tiempo de corte. Por último, los pacientes de demencia la concentran después. Comparando con RPDPM, se observa en la figura 5.5 que en RPDPM la reserva de los pacientes de demencia al inicio del estudio es muy similar a la de los pacientes pMCI, indicando posibles errores en la estimación de RPDPM. Por tanto, se puede considerar mejor la estimación de Leaspy.

En último lugar, si observamos la tasa de éxito en la detección de la enfermedad, en la tabla 5.9 se observa que los pacientes que ha catalogado como sMCI y pMCI a menudo se han clasificado correctamente. Sin embargo, al mirar a la demencia, se observa que comete más errores, siendo su tasa de éxito del 86 % frente a los 97 % y 94 % de los otros dos grupos. Esto se puede deber a la menor cantidad de datos que se disponen de pacientes con demencia, que por razones morales solo se les realiza un seguimiento por dos años.

## 6.5. Pronostico

En primer lugar, se valorará los resultados del primer experimento. Observando el MAE de los marcadores, en la figura 5.8 se observa que, en general, leaspy comete más error, y con mayores incertidumbres, que simplemente asumir que el valor se repite en el tiempo. Además, se comprueba que tanto el error como su desviación aumenta ligeramente con el tiempo. Este incremento es mucho más rápido para los grupos de test, figura 5.7, que nunca ha visto el algoritmo antes de la personalización, a diferencia de los de train (figura 5.6). Estos grupos también poseen una mayor desviación. Si por otra parte se compara con los resultados de RPDPM, se observa que Leaspy tiene un MAE mucho más alto, con una desviación mucho más grande.

En cuanto a las métricas de calidad del experimento general, los resultados se encuentran en la figura 5.9 para el grupo de training, la figura 5.10 para el de test, y la figura 5.11 para el conjunto de los dos grupos. Es importante mencionar primero que el número de sujetos a estudiar desciende enormemente con el tiempo. En cuanto al área bajo la curva, se observa que al usar leaspy decae mucho más despacio con el tiempo en comparación con el predictor constante. Además, leaspy

tiene un AUC mucho más uniforme entre los grupos de test y train, mientras que el predictor constante cae mucho más para grupos de test. En lo que se refiere a la precisión de la predicción, es muy similar entre los dos predictores para los grupos de train. Para los grupos de test, sin embargo, leaspy empieza siendo algo peor que el constante, pero cae mucho más despacio, lo que lleva a que se igualen a partir del cuarto año. Comparando las mismas métricas con los resultados de RPDPM, se observa como las medidas de AUC y precisión de Leaspy son mejores que las de RPDPM, y además caen más despacio.

Evaluando la capacidad de detección de sMCI, pMCI y demencias, se observa que el predictor constante, como ya se sabía, siempre acierta los sMCI y las demencias, pero nunca los pMCI. Es por este motivo que se busca usar otros predictores como leaspy. En cuanto a los resultados de leaspy, la capacidad de detectar pMCI aumenta con el tiempo en el grupo de train, y ligeramente en el de test. La capacidad de detectar sMCI, por su parte, disminuye, mientras que la de demencias permanece relativamente constante. Además, se observa que una vez más, las métricas son peores en los grupos de test. Si comparamos por otra parte las medidas de Leaspy con las de RPDPM se observa como RPDPM tiene menores tasas de detección de sMCI y demencia, y además, las de pMCI cae más y más rápido con el tiempo.

En cuanto al segundo experimento, figura 5.12, para detectar unicamente la capacidad de conversión, se observa que la tasa de detección de pMCI es relativamente elevada, y decae lentamente con el tiempo. Lo mismo aplica a la correlación con la edad. Sin embargo, al valorar la correlación con la reserva cognitiva, se observa que es bastante baja, y va disminuyendo con el tiempo, hasta caer en picado al estimar a cinco años. Se concluye por tanto que aunque leaspy permite conocer si el individuo convertirá o no, no puede elaborar predicciones muy exactas acerca de cuando convertirá el paciente. Haciendo la comparación con RPDPM se observa que RPDPM tiende a acertar mejor la edad de conversión, pero a identificar peor a los pMCI.

## Capítulo 7

# Conclusiones

### 7.1. Conclusión

Una vez finalizado el estudio, centrandó la atención primero en el estudio transversal, se observa que como ya indicaban otros artículos [21, 20, 37] hay una relación entre la presencia de patología amiloide y el desarrollo de la enfermedad. Sin embargo, la presencia de un grupo minoritario de A-T+ tanto entre los pacientes pMCI y los de demencia confirma que entre los pacientes de AD existen otros pacientes de demencia que no se corresponden a esta enfermedad. Además, los histogramas de la figura 5.1 confirman que la progresión a demencia de ese grupo, o al menos de los A-, es más lenta. Además, el estudio de significancia de los marcadores de las tablas 5.1 y 5.2 confirma la edad y educación como factores de riesgo [13].

En cuanto a la comparación de Leaspy con sus otras alternativas se observa que da unos resultados similares, requiriendo de menos marcadores, lo que lo hace más conveniente en la práctica clínica. Además, aunque algunos de los resultados del mejor modelo puedan ser ligeramente peores que los de los modelos de RPDPM usados, la mejor robustez proporcionada por Leaspy justificaría su uso frente a RPDPM. Por otra parte, los buenos resultados de la historia natural indican que si se hiciesen modelos de la población general separados para A-T+ y para A+T+ se podrían construir historias naturales de los dos grupos por separado, y comprobar si, como se espera, hay diferencias en la progresión. Por último, los buenos resultados de Leaspy en el pronóstico, similares al predictor constante de referencia en las métricas generales, indican que en general, podría ser una buena herramienta para elaborar predicciones cualitativas sobre si los pacientes convertirán o no, aunque es difícil asegurar cuando en periodos largos de tiempo.

### 7.2. Desarrollos futuros

Tras confirmar que existe un grupo que podría no pertenecer a la AD, y comprobar que Leaspy es una buena herramienta para el estudio de la enfermedad, el siguiente paso debería ser elaborar modelos por separado de este grupo A-T+ y de los pacientes del grupo A+T+. Esto permitiría conocer si hay diferencias en la historia natural entre los dos grupos. Además, sería interesante añadir los biomarcadores, como las medidas de ABETA o Tau a la lista de marcadores con los que

construir los modelos.

Con la misma finalidad, se debería prestar especial atención a los marcadores en sangre que empiezan a sugerir algunos estudios. Su implantación podría permitir alejar las medidas neuropsicológicas del diagnóstico, y para este estudio, permitiría tener más pacientes con medidas de biomarcadores, aumentando el tamaño de la muestra si se quieren incluir también estos marcadores.

Por último, se deberían incluir marcadores de riesgo vascular (V), ya que permitirían diferenciar mejor entre los casos de AD y las demencias de origen vascular, provocadas por micro infartos cerebrales. Además, se podrían añadir también otros factores de riesgo como la ingesta de alcohol o el consumo de tabaco, que aumentan no solo el riesgo de enfermedades vasculares, también parecen estar relacionadas con el riesgo de demencia.

## Apéndice A

# Anexo: Geometría de Riemann

[24]

A la hora de realizar DPMs se parte de la hipótesis de que todos los datos de interés se encuentran en un subespacio del espacio estudiado. Las trayectorias de los marcadores se encuentran en curvas dentro del subespacio, cuyos puntos son las diferentes observaciones. Para trabajar con esta información, es necesario por tanto introducir la geometría riemanniana.

La geometría de Riemann permite, no solo definir estos espacios, también operar con ellos para obtener puntos, curvas y distancias. Históricamente, la geometría de Riemann se ha utilizado para estudiar espacios topológicos diferenciables en  $\mathbb{R}^n$  llamados variedades. Las variedades están definidas por los parámetros que permiten transformarlas al espacio Euclídeo. Además, con las variedades se pueden realizar transportes paralelos de curvas, desplazándolas a otra región de la variedad.

### A.1. Variedades

Una variedad  $M \subset \mathbb{R}^n$  es un espacio topológico para el cual cada punto presenta un vecindario (conjunto que rodea a un espacio) que es homeomórfico al espacio Euclídeo. Es decir, existen funciones continuas de este espacio al lineal. Se puede derivar de el espacio lineal a su región correspondiente, lo que lleva a estructuras localmente diferenciables. Si son continuas, se considera globalmente diferenciable. Por último, para una variedad diferenciable, se define el espacio tangente  $T_p M$  como el espacio que contiene todas las posibles derivadas de  $M$  en un punto  $p$ .

### A.2. Métricas y Variedades Riemannianas

Se considera una variedad diferenciable  $M$  tal que a cada punto  $p \in M$  se le asocia un producto escalar  $g_p$  en el campo vectorial del espacio tangente  $T_p M$ , el cual varía suavemente de punto a punto. A la colección  $g_M = (g_p)_{p \in M}$  se la llama métrica de la variedad. Esto actúa como una generalización del producto escalar para las variedades. A la métrica equipada con la métrica  $(M, g_M)$  se la llama Variedad de Riemann. Gracias a esto, se pueden introducir conceptos como las distancias.

### A.3. Geodiscos

Los geodiscos son los equivalentes en geometría Riemanniana a las líneas rectas. Se trata de curvas que representan la distancia más corta entre dos puntos de una variedad.

### A.4. Mapeado Exponencial

Se considera un punto  $p \in M$ , una velocidad  $v \in T_p M$  y un geodisco  $\gamma$  que cumple  $\gamma(t) = p$  y  $\dot{\gamma} = v$ . Se puede demostrar que el geodisco se puede reescribir como  $\gamma = \text{Exp}_{p,t}(v) : t \mapsto \text{Exp}_{p,t}(v)(t)$ . El mapeado exponencial asocia el vector  $v$  al punto alcanzado por el geodisco en  $t+1$ . Define  $v \in T_p \mapsto \text{Exp}_p(v) = \text{Exp}_{p,t}(v)(t+1)$

### A.5. Transporte Paralelo

Dada una variedad  $M$  y una curva diferenciable  $\gamma : I \subset \mathbb{R} \rightarrow M$ , un campo de vectores  $X$  es paralelo a lo largo de  $\gamma$  si  $\frac{DX}{dt} = 0$ . Dado  $w_0 \in T_{\gamma(t_0)}M$ , se puede demostrar que existe un campo de vectores único  $w(t)$  paralelo a lo largo de  $\gamma$  de tal forma que  $w(t_0) = w_0$ . Esto se corresponde al transporte de  $w_0$  a lo largo de  $\gamma$  de tal forma que  $w(t_0)$  siga siendo paralelo a  $w_0$ .

## Apéndice B

# Anexo: Contraste de Hipótesis. P-valor. ANOVA. Chi-cuadrado [43]

### B.1. Contraste de Hipótesis y P-valor

Se define hipótesis estadística como una conjetura respecto a una o dos poblaciones. Dado que el tamaño de las muestras es limitado, la toma de decisiones se debe realizar sabiendo la probabilidad de que la verdad o falsedad de las hipótesis sea errónea. Existe una probabilidad de que se rechace la hipótesis aún siendo verdadera si así lo indica la muestra.

En la prueba de hipótesis es importante conocer el concepto de **hipótesis nula** ( $H_0$ ). La hipótesis nula se refiere a la hipótesis a probar. Su rechazo conduce a la aceptación de la **hipótesis alternativa** que se opone directamente a la hipótesis nula. Habitualmente se estructura la prueba de tal forma que la hipótesis alternativa es la teoría a probar.

Al realizar una prueba de hipótesis hay una probabilidad de cometer dos tipos de error:

- El **Error de Tipo I** se produce al rechazar una hipótesis nula que, en realidad, es verdadera. La probabilidad de cometer este error se llama nivel de significancia o tamaño de la prueba y se denota como  $\alpha$
- El **Error de Tipo II** se produce al no rechazar una hipótesis nula que, en realidad, es falsa. La probabilidad de cometer este error se denota como  $\beta$  y no es posible calcularlo si no se tiene una hipótesis alternativa específica.

Los procedimientos de prueba deben intentar evitar los dos tipos de error. Una forma de reducir la probabilidad de cometer cualquiera de los dos errores es aumentar el tamaño de la muestra. Ambos errores se encuentran estrechamente relacionados y, por lo general, reducir el riesgo de uno de ellos sin aumentar la muestra aumenta el riesgo de cometer el otro.

Se define la potencia de la prueba como la probabilidad de rechazar la hipótesis nula cuando una alternativa específica es verdadera. Se puede calcular como  $1 - \beta$

y se utiliza para comprobar diferentes tipos de prueba.

Por último, el método del P valor permite asignar una probabilidad para decidir si se rechaza o no la hipótesis. Se define como el nivel de significancia más bajo para el que estadístico de la prueba es significativo. El P valor indica la posibilidad de obtener un conjunto de datos entre las muestras de una distribución. Por lo tanto, P valores muy pequeños refutan la hipótesis.

## B.2. Análisis de Varianza (ANOVA)

Para la realización de la prueba de hipótesis con medias de la población uno de los métodos más empleados es el del análisis de varianza (ANOVA). El objetivo del análisis de varianza es identificar si las diferencias entre muestras de poblaciones diferentes se deben a variaciones aleatorias o diferencias entre las propias poblaciones. Se llaman factores a las variables independientes que se manipulan entre las poblaciones, así se puede realizar análisis de varianza de un solo factor o de dos o más factores.

Para ANOVA de un solo factor se emplean las siguientes suposiciones:

- Las  $k$  poblaciones se encuentran normalmente distribuidas
- Todas las poblaciones tienen la misma varianza
- Si todas las variaciones entre poblaciones se deben a efectos aleatorios, sus medias serán iguales. Por lo tanto, la hipótesis nula a probar será que las medias de las  $k$  muestras son iguales. La hipótesis alternativa es por lo tanto que al menos dos son diferentes.

Partiendo de esto, las observaciones se pueden escribir las observaciones como:

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad (\text{B.1})$$

Siendo  $\mu_i$  la media de la  $i$ -ésima muestra y  $\epsilon_{ij}$  la desviación con respecto a la media de la observación  $j$  de la muestra  $i$ . Se puede considerar por tanto que  $\epsilon$  representa el error aleatorio.

A menudo se realiza sobre esta ecuación la sustitución  $\mu_i = \mu + \alpha_i$ :

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (\text{B.2})$$

La sustitución se encuentra condicionada a que  $\sum_{i=0}^k \alpha_i$  de tal forma que  $\mu$  sea una media general de todas las muestras y  $\alpha_i$ , denominada efecto, represente las diferencias entre ellas. De esta manera, la hipótesis nula se puede reescribir como que todos los efectos sea iguales a 0.

La prueba de la hipótesis se realiza mediante la comparación de diferentes varianzas de los datos:

- En primer lugar, se utiliza la suma total de los cuadrados (SST)

- En segundo lugar, la suma de los cuadrados del grupo (SSB)
- Y por último la suma de los cuadrados del error (SSE)

Que se relacionan por la siguiente ecuación:

$$SST = SSB + SSE \quad (\text{B.3})$$

Partiendo de esto se puede llegar que para una  $H_0$  verdadera:

$$s_1^2 = \frac{SST}{k-1} = \sigma^2 \quad (\text{B.4})$$

Por otro lado, si  $H_1$  es verdadero aparecerá un término adicional relacionado con  $\sum \alpha_i^2$  que mide variaciones sistemáticas.

### B.3. Chi Cuadrado

Para evaluar la independencia de variables categóricas la prueba más habitual es la prueba Chi cuadrado. En una prueba chi cuadrado, la hipótesis nula,  $H_0$  a valorar será que dos variables son independientes. Para realizar la prueba se siguen los siguientes pasos:

1. Se construye la tabla de contingencia de  $r \times l$  celdas. La tabla contiene en sus celdas el número de observaciones (frecuencia de observación) de dos variables juntas. A los totales de cada fila o columna se les llaman frecuencias marginales.
2. Se calculan las frecuencias esperadas de cada celda, obteniendo la probabilidad de que una observación pertenezca a la celda (producto de las probabilidades de los totales de su fila y columna) y multiplicando por el número de observaciones.
3. Se calcula el  $\chi^2$  utilizando la información de la tabla de contingencia, las frecuencias esperadas y la siguiente ecuación:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \quad (\text{B.5})$$

Donde  $k$  es el número de celdas,  $o_i$  es la frecuencia observada y  $e_i$  la frecuencia esperada de cada una de las celdas de la tabla de contingencia.

4. Se obtiene el grado de libertad de la prueba  $v = (r-1)(l-1)$  y junto con la significancia  $\alpha$  se consulta en la tabla de la distribución el valor crítico  $\chi_\alpha^2$
5. Si  $\chi^2 > \chi_\alpha^2$  se rechaza la hipótesis nula



## Apéndice C

# Anexo: Clasificadores KDE-Bayes

Para la detección de si un individuo pertenece a MCI o demencia mediante la información modelos, se utilizan clasificadores de Bayes-KDE. Estos consisten, en primer lugar, en un Estimador de Densidad Kernel (KDE) que estima las densidades de probabilidad de cada característica para cada clase. A este, se le une un clasificador de Bayes que estima la probabilidad de que los datos nuevos del modelo pertenezcan a la clase. A cada individuo se le asigna el grupo más probable (MCI o Demencia).

### C.1. Clasificadores de Bayes

La inferencia Bayesiana consiste en, dado un modelo de probabilidad, conseguir una función de distribución de probabilidad que se ajuste un conjunto dado de datos y permita elaborar predicciones para nuevas observaciones. [12]

Como su propio nombre indica, los Clasificadores de Bayes se basan en la aplicación del Teorema de Bayes, cuya ecuación es la siguiente:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)} \quad (\text{C.1})$$

Siendo:

- $p(A)$  es la probabilidad de que ocurra el evento A
- $p(B)$  es la probabilidad de que ocurra el evento B
- $p(A|B)$  es la probabilidad de que habiendo ocurrido A ocurra B
- $p(B|A)$  es la probabilidad de que habiendo ocurrido B ocurra A

En este caso concreto, uno de los eventos pasaría a ser que el individuo pertenezca a una de las clases, mientras que el otro sería, por ejemplo, obtener un determinado valor en un marcador.

Para su implementación en Python se puede emplear el paquete *sklearn* que contiene varias clases para la elaboración de clasificadores de Bayes ingenuos con diferentes modelos [35]

## C.2. Estimación de Densidad Kernel (KDE)

El clasificador de Bayes por si solo no es suficiente para agrupar los individuos, es necesario conocer la funciones de densidad de probabilidad de las características a estudiar. Para ello se utilizará un KDE. Se trata de procesos no paramétricos para obtener funciones de densidad o elaborar clasificadores.

La ecuación fundamental de la KDE es la siguiente: [15]

$$\hat{f}_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^N K_\lambda(x_0, x_i) \quad (\text{C.2})$$

Donde:

- $\hat{f}_X(x_0)$  Es la estimación de la densidad en un punto.
- $\frac{1}{N\lambda}$  Es el factor de suavizado. N es el número de observaciones. Por otra parte,  $\lambda$  es el ancho de banda, y determina el detalle y suavizado de la estimación
- $K_\lambda(x_0, x_i)$  es la función núcleo (Kernel), que mide la similitud entre un punto de referencia  $x_0$  y las observaciones. Habitualmente es una Gaussiana.

Una vez más, para su implementación en Python se puede usar el paquete sklearn que contiene clases dedicadas a KDE, permitiendo elegir el núcleo y ancho de banda.[35]

## Apéndice D

# Anexo: Estructura de Carpetas

La estructura de carpetas empleada en el proyecto es la siguiente. No se incluyen indicados los archivos de resultados, aunque se encontrarán presentes en la demo:

```
Demo/
|
+ -- DataFrames/
| |
| + -- adnimerge_datos.csv
| + -- DEM.csv
| + -- DEM_CSF.csv
| + -- pMCI.csv
| + -- pMCI_CSF.csv
| + -- sMCI.csv
| + -- sMCI_CSF.csv
|
+ -- Tablas/
| |
| + -- Tablas_aux/
| | |
| | + -- __pycache__/
| | + -- clasificar_grupos.py
| | + -- construir_tablas.py
| | + -- preprocesado.py
| | + -- significancia.py
| |
| + -- Tabla1_latexCode.txt
| + -- TablaAT_latexCode.txt
| + -- Tablas.py
|
+ -- t_conversion_censura/
| |
| + -- aux_histogramas/
| | |
| | + -- __pycache__/
| | + -- busqueda_en_df.py
| | + -- calculo_fechas.py
```

```
| | + -- calculo_tiempos.py
| |
| + -- Histogramas.py
|
+ -- multileaspy/
| |
| + -- aux_leaspyProd/
| |   + -- __pycache__/
| |   + -- addNaNUniform.py
| |   + -- bayes_train.py
| |   + -- build_leaspy.py
| |   + -- build_model_leaspy.py
| |   + -- get_diagnose.py
| |   + -- get_dps.py
| |   + -- getAUC.py
| |   + -- getDataADNI.py
| |   + -- getMae.py
| |   + -- getOnset.py
| |   + -- getSubVisitMatrix.py
| |   + -- getTrueOnset.py
| |   + -- leaspy_alg.py
| |   + -- percentile_interp.py
| |   + -- removeVisits.py
| |
| + -- data/
| |   + -- adnimerge_datos.csv
| |
| + -- multi_Leaspy_Magerit.sh
| + -- multileaspy_faseProdromica.py
| + -- multileaspy_par_mpi.py
|
+ -- historia_natural/
| |
| + -- aux_leaspyProd/
| |   + -- __pycache__/
| |   + -- addNaNUniform.py
| |   + -- bayes_train.py
| |   + -- build_leaspy.py
| |   + -- build_model_leaspy.py
| |   + -- get_diagnose.py
| |   + -- get_dps.py
| |   + -- getAUC.py
| |   + -- getDataADNI.py
| |   + -- getMae.py
| |   + -- getOnset.py
| |   + -- getSubVisitMatrix.py
| |   + -- getTrueOnset.py
| |   + -- leaspy_alg.py
```

```

|     | + -- percentile_interp.py
|     | + -- removeVisits.py
|     |
| + -- aux_timeline/
|     | + -- __pycache__/_
|     | + -- cutoffTable.py
|     | + -- getConversionTime.py
|     | + -- getConversionTimeLSTM.py
|     | + -- getTimeline.py
|     | + -- getTrayLeaspy.py
|     | + -- group_classification.py
|     | + -- plotLongShortTray.py
|     | + -- ReserveAnalisis.py
|     | + -- reserveHistogram.py
|     | + -- timelineEstimate.py
|     | + -- trayFit.py
|     | + -- trayFitSpline.py
|     | + -- trayTableLeaspy.py
|     |
| + -- data/
|     | + -- historia_natural/
|     | | + -- age_onset_leaspy.mat
|     | | + -- age_onset_leaspy.pkl
|     | | + -- df.mat
|     | |
|     | + -- leaspy_7_8/
|     | | + -- dd.xlsx
|     | | + -- leaspy_btstrp.pkl
|     | | + -- leaspy_btstrp1.json
|     | | + -- leaspy_btstrp2.json
|     | | + -- leaspy_btstrp3.json
|     | | + -- leaspy_btstrp4.json
|     | | + -- leaspy_btstrp5.json
|     | | + -- leaspy_btstrp6.json
|     | | + -- leaspy_btstrp7.json
|     | | + -- leaspy_btstrp8.json
|     | | + -- leaspy_btstrp9.json
|     | | + -- leaspy_btstrp10.json
|     | |
|     | + -- adnimerge_datos.csv
|     |
| + -- graficas/
| + -- matlab-metricas/
|     | + -- data/
|     | | + -- age_onset_leaspy.mat
|     | | + -- data_MCItoAD_model_multi.mat
|     | | + -- df.mat
|     | | + -- MCItoAD_3__7__8.mat

```

```

|   |   |   + -- tSample_tTray_3_7_8_tau_MCI2AD_10_10y.mat
|   |   |
|   |   + -- rpdpm/
|   |   |   + -- analyze_t_mae_auc.m
|   |   |   + -- analyze_t_mae_auc_1_1.m
|   |   |   + -- bayes_train.m
|   |   |   + -- build_model_RPDPM.m
|   |   |   + -- ...
|   |
| + -- RPDPM/
|   |   + -- data/
|   |   |   + -- data_MCItoAD_model_multi.mat
|   |   |   + -- MCItoAD_3__7__8.mat
|   |   |   + -- tSample_tTray_3_7_8_tau_MCI2AD_10_10y.mat
|   |   |
|   |   + -- rpdpm/
|   |   |   + -- otros archivos...
|   |   |
|   |   + -- analysis_RPDPM_tau_MCI2AD.m
|   |   + -- distribution_tau_MCI2AD_RP.m
|   |   + -- get_tau_anyMode_MCItoAD_v1.m
|   |   + -- plot_timeline_MCItoAD_RP_v1.m
|   |
| + -- data.pkl
| + -- data1.pkl
| + -- data.qs
| + -- df.csv
| + -- diagnostico_estimado.csv
| + -- estimateConversionTime.py
| + -- plotReserveDistribution.py
| + -- plotTimeline.py
| + -- tray.pkl
|
+ -- pronostico/
  + -- leaspy+constante/
    |   + -- aux_const/
    |   |   + -- analyze_t_mae_auc.py
    |   |   + -- analyze_tau.py
    |   |   + -- ...
    |   |
    |   + -- aux_leaspy/
    |   |   + -- _pycache_/
    |   |   + -- addNaNUniform.py
    |   |   + -- ...
    |   |
    |   + -- data/
    |   |   + -- leaspy_7_8/
    |   |   |   + -- dd.xlsx

```

```
| | | + -- ...
| | + -- adnimerge_datos.csv
| | + -- tau.mat
| |
| + -- graficas/
| + -- tmp/
| + -- future_constant_prediction.py
| + -- future_leaspy_prediction.py
| + -- future_prediction_compare.py
|
+ -- RPDPM/
  + -- data/
    | + -- data_MCItoAD_model_multi.mat
    | + -- ...
    |
  + -- rpdpm/
    | + -- analyze_t_mae_auc.m
    | + -- ...
    |
  + -- show_convertTime_MCItoAD_prognosis.m
  + -- show_MCItoAD_prognosis.m
```



## Apéndice E

# Anexo: Informe de Turnitin

Turnitin es una herramienta de detección de plagio ampliamente utilizada para evaluar la originalidad de documentos [42]. Esta herramienta fue utilizada para verificar la originalidad de este documento. Tras el análisis del documento se comprueba que cumple los estándares académicos exigidos y que el documento se encuentra libre de plagio. Se puede ver el informe en la siguiente imagen:

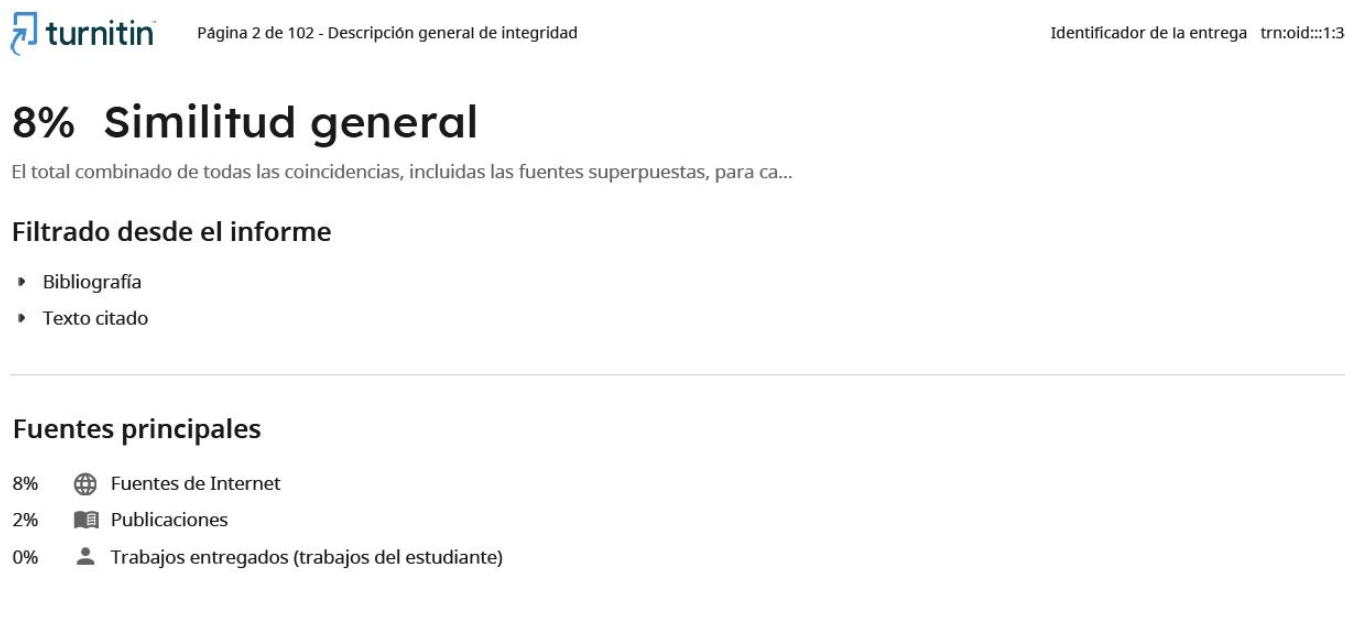


Figura E.1: Informe de similitud de Turnitin[42]



# Bibliografía

- [1]
- [2] Adni - alzheimer's disease neuroimaging initiative. <https://adni.loni.usc.edu>.
- [3] Centos - about. <https://www.centos.org/about/>.
- [4] Cesvima - centro de supercomputación y visualización de madrid. <https://www.cesvima.upm.es>.
- [5] Alzheimer's Disease International (ADI). Dementia statistics, 2025. Página consultada sobre cifras clave de demencia a nivel mundial: prevalencia, incidencia, costes y diagnóstico.
- [6] Jorge Bengoa. Métodos paramétricos y redes neuronales recurrentes en el diagnóstico de la enfermedad de alzheimer, julio 2025.
- [7] Marina Bentivoglio and Gigliola Grassi Zucconi. *Cuando el Cerebro Envejece: Mitos y Certezas sobre un Proceso Universal (e Inevitable)*. EMSE EDAPP, Barcelona, 2018.
- [8] Jorge L Bernal-Rusiel, Douglas N Greve, Martin Reuter, Bruce Fischl, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66:249–260, 2013.
- [9] Lisandro Dalcin. Mpi for python, 2025.
- [10] Organización Mundial de la Salud. Demencia. Nota descriptiva en línea, mar 2025.
- [11] Michael C Donohue, Reisa A Sperling, David P Salmon, Dorene M Rentz, Rema Raman, Ronald G Thomas, Michael Weiner, Paul S Aisen, et al. The preclinical alzheimer cognitive composite: measuring amyloid-related decline. *JAMA neurology*, 71(8):961–970, 2014.
- [12] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- [13] Pedro Gil Gregorio. *Neurodegeneración: Alzheimer, Parkinson y ELA*. EMSE EDAPP, 2018.
- [14] Colin Harrison. Xming: X server for windows, 2025.

- [15] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.
- [16] INE. Inebase. consul - [serie] indicadores urbanos - total nacional, esperanza de vida al nacer (años), total - ine. instituto nacional de estadística.
- [17] INE. Población residente por fecha, sexo y edad (desde 1971).
- [18] Instituto Nacional de Estadística (INE). Enfermedades crónicas diagnosticadas por sexo y edad. población con discapacidad. <https://www.ine.es/jaxi/Tabla.htm?tpx=66954>, 2025. Consulta realizada el 9 de septiembre de 2025.
- [19] Instituto Nacional de Estadística (INE). Estadística de defunciones según la causa de muerte. Últimos datos, 2025. Consulta realizada el 9 de septiembre de 2025.
- [20] Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of alzheimer’s disease: Alzheimer’s association workgroup. *Alzheimer’s & Dementia*, 20(8):5143–5169, 2024.
- [21] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer’s disease. *Alzheimer’s & dementia*, 14(4):535–562, 2018.
- [22] Young Ju Kim, Si Eun Kim, Alice Hahn, Hyemin Jang, Jun Pyo Kim, Hee Jin Kim, Duk L Na, Juhee Chin, and Sang Won Seo. Classification and prediction of cognitive trajectories of cognitively unimpaired individuals. *Frontiers in Aging Neuroscience*, 15:1122927, 2023.
- [23] Tim Kosse. Filezilla: The free ftp solution, 2025.
- [24] Igor Koval. *Learning multimodal digital models of disease progression from longitudinal data: methods & algorithms for the description, prediction and simulation of Alzheimer’s disease progression*. PhD thesis, Institut polytechnique de Paris, 2020.
- [25] Igor Koval. Leaspy documentation, 2025.
- [26] Kelly L Lange, Mark W Bondi, David P Salmon, Douglas Galasko, Dean C Delis, Ronald G Thomas, and Leon J Thal. Decline in verbal memory during preclinical alzheimer’s disease: examination of the effect of apoe genotype. *Journal of the International Neuropsychological Society*, 8(7):943–955, 2002.
- [27] Muriel Deutsch Lezak. *Neuropsychological assessment*. Oxford University Press, USA, 2004.
- [28] Nicholas Marriott. tmux wiki, 2025.
- [29] Ann M Mayo. Use of the functional activities questionnaire in older adults with dementia. *Hartford Inst Geriatr Nurs*, 13(2), 2016.
- [30] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, Marc Modat, M Jorge Cardoso, S’ebastien Ourselin, and Lauge Sørensen. Robust parametric modeling of Alzheimer’s disease progression. *NeuroImage*, 225:117460, 2021.

- [31] Ministerio de Sanidad, Consumo y Bienestar Social. Plan integral de alzheimer y otras demencias (2019–2023). Publicación oficial en línea (PDF), 2019. Documento aprobado por el Ministerio de Sanidad, Consumo y Bienestar Social de España.
- [32] Disease Progression Modelling. Tp2 leaspy beginner. [https://disease-progression-modelling.github.io/pages/notebooks/disease\\_course\\_mapping/TP2\\_leaspy\\_beginner.html](https://disease-progression-modelling.github.io/pages/notebooks/disease_course_mapping/TP2_leaspy_beginner.html), 2025.
- [33] Hyae Sung Moon. Optimization for alzheimer’s disease prediction with ai: Leaspy-driven disease progression model, febrero 2024.
- [34] OECD. *Addressing Dementia: The OECD Response*. OECD Health Policy Studies. OECD Publishing, Paris, 2015.
- [35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. *Scikit-learn: Machine Learning in Python*, 2011.
- [36] Carlos Platero and Jorge Bengoa. Benchmarking parametric models of disease progression for early detection of cognitive decline. *Computer Methods and Programs in Biomedicine*, page 109162, 2025.
- [37] Carlos Platero, Jussi Tohka, and Bryan Strange. Estimating dementia onset: At (n) profiles and predictive modeling in mild cognitive impairment patients. *Current Alzheimer Research*, 20(11):778–790, 2023.
- [38] Martin Prince, Anders Wimo, Maëlen Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. *World Alzheimer report 2015. The global impact of dementia: an analysis of prevalence, incidence, cost and trends*. PhD thesis, Alzheimer’s Disease International, 2015.
- [39] Stroke Engine. Trail making test (tmt), n.d.
- [40] Jing Ee Tan, Esther Strauss, and Elisabeth Sherman. Clinical dementia rating. In *encyclopedia of clinical neuropsychology*, pages 803–806. Springer, 2018.
- [41] Simon Tatham. Putty: A free ssh and telnet client, 2025.
- [42] LLC Turnitin. Turnitin. <https://www.turnitin.com>, 2025. Accessed: 2025-05-05.
- [43] Ronald E Walpole et al. Probabilidad y estadística para ingenieros y ciencias. 2012.