

## Trabajo de Fin de Grado

Grado en Ingeniería Informática

Computación

---

# **Comparación del envejecimiento cerebral en pacientes preclínicos de la enfermedad de Alzhéimer y pacientes sanos**

---

*David Bernal Gómez*

### **Dirección**

Olatz Arbelaitz Gallego

24 de junio de 2023



---

# Resumen

---

La enfermedad del Alzheimer es una enfermedad neurodegenerativa progresiva que afecta principalmente a las personas mayores y se caracteriza por la pérdida gradual de las funciones cognitivas, incluyendo la memoria, el lenguaje, la atención y la capacidad para realizar actividades diarias. Es la causa más común de demencia en todo el mundo, afectando a más de 50 millones de personas en la actualidad.

Los síntomas de la enfermedad del Alzheimer son el resultado de la muerte de las células cerebrales y la pérdida de conexiones neuronales, lo que lleva a un deterioro cognitivo progresivo y una disminución de la función cerebral. Actualmente no existe cura para la enfermedad, y los tratamientos disponibles sólo pueden mejorar temporalmente algunos de los síntomas.

Por todo esto, la detección temprana del Alzheimer es crucial para poder ralentizar su progreso y mejorar la calidad de vida del paciente.

En este contexto, el siguiente trabajo tiene como objetivo detectar diferencias en la evolución de las distintas áreas cerebrales con respecto a la edad entre personas sanas y personas con alguna patología neurodegenerativa. El estudio y análisis de estas diferencias puede proporcionar información valiosa para el diagnóstico y la monitorización de la enfermedad del Alzheimer, y eventualmente conducir a mejores tratamientos y estrategias de prevención.



---

# Índice de contenidos

---

<b>Índice de contenidos</b>	<b>III</b>
<b>Índice de figuras</b>	<b>V</b>
<b>Índice de tablas</b>	<b>VII</b>
<b>Índice de algoritmos</b>	<b>IX</b>
<b>1 Introducción</b>	<b>1</b>
1.1. Contexto . . . . .	1
1.2. Motivación . . . . .	2
1.3. Objetivo . . . . .	3
<b>2 Planificación</b>	<b>5</b>
2.1. Diagrama EDT . . . . .	5
2.2. Desarrollo Estimado . . . . .	6
2.3. Gestión de riesgos . . . . .	7
2.4. Desarrollo realizado . . . . .	8
<b>3 Fundamentos</b>	<b>9</b>
3.1. Enfermedad de Alzheimer . . . . .	9
3.1.1. Definición . . . . .	9
3.1.2. Posibles causas . . . . .	9
3.1.3. Síntomas . . . . .	10
3.1.4. Fases de la enfermedad . . . . .	11
3.1.5. Detección . . . . .	11
3.2. El cerebro . . . . .	12
3.2.1. Áreas cerebrales . . . . .	12
3.2.2. Las neuronas . . . . .	13
3.3. Aprendizaje automático . . . . .	14
3.3.1. Técnica de Regresión . . . . .	14
3.3.2. Normalización y estandarización . . . . .	17
3.3.3. Bootstrapping e intervalos de confianza . . . . .	18

3.4. Matriz de confusión . . . . .	18
<b>4 La base de datos</b>	<b>21</b>
4.1. Contenido de la base de datos . . . . .	21
4.2. Análisis del contenido . . . . .	23
4.3. Preprocesado . . . . .	25
4.4. Nomenclatura . . . . .	26
<b>5 Proceso de implementación</b>	<b>27</b>
5.1. Entorno de trabajo . . . . .	27
5.2. Aproximación con regresión Ridge . . . . .	27
5.3. Comparación de curvas . . . . .	28
5.4. Ordenación de áreas en función de su diferencia . . . . .	30
5.5. Cuestiones tras la primera aproximación . . . . .	34
5.6. Resultados de las nuevas aproximaciones . . . . .	35
5.7. Curvas a partir de muestras Bootstrap . . . . .	37
5.8. Resultados de la aproximación con Bootstrapping . . . . .	44
5.9. Clasificación de pacientes usando las curvas . . . . .	45
5.9.1. Primera aproximación de clasificación . . . . .	46
5.9.2. Segunda aproximación de clasificación . . . . .	47
5.9.3. Resultados de la clasificación . . . . .	48
5.10. Clasificación con las mejores áreas y uso del umbral . . . . .	50
<b>6 Conclusiones y trabajo futuro</b>	<b>53</b>
6.1. Análisis de resultados y Conclusiones . . . . .	53
6.2. Trabajo futuro . . . . .	56
<b>Apéndice</b>	<b>57</b>
<b>Bibliografía</b>	<b>59</b>

---

# Índice de figuras

---

2.1. Diagrama EDT . . . . .	6
2.2. Diagrama Gantt con los períodos planificados . . . . .	7
3.1. Efecto de las proteínas Beta-amiloide y tau en el cerebro . . . . .	10
3.2. División del cerebro en lóbulos . . . . .	13
3.3. La neurona y sus partes . . . . .	13
3.4. Diferencia entre clasificación y regresión . . . . .	15
3.5. Ejemplo del efecto de $\lambda$ en la regresión Ridge . . . . .	16
4.1. Proporción de ejemplos con y sin punción lumbar . . . . .	23
4.2. Proporción de ejemplos según sus patologías, para todos los ejemplos de la bd . . . . .	24
4.3. Proporción de ejemplos según sus patologías, solo para los ejemplos utilizados . . . . .	24
4.4. Histograma con la distribución de los ejemplos por rangos de edad . . . . .	25
5.1. Curvas de las 3 materias del área número 5 para el grupo de pacientes sin patologías . . . . .	28
5.2. Curvas de Ant_temp_med_L_WM_VOL.5 . . . . .	29
5.3. Áreas ordenadas por diferencia de GM . . . . .	31
5.4. Áreas ordenadas por diferencia de WM . . . . .	32
5.5. Áreas ordenadas por diferencia de CSF . . . . .	32
5.6. Las 5 áreas con mayor diferencia de cada materia . . . . .	33
5.7. Comparación visual de las diferencias de rango en las curvas del área 18 . . . . .	35
5.8. Ejemplo de curvas con sus intervalos para el área 1 . . . . .	37
5.9. Áreas ordenadas por diferencia total en curvas bootstrap. Materia GM . . . . .	40
5.10. Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia GM . . . . .	40
5.11. Áreas ordenadas por diferencia total en curvas bootstrap. Materia WM . . . . .	41
5.12. Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia WM . . . . .	41
5.13. Áreas ordenadas por diferencia total en curvas bootstrap. Materia CSF . . . . .	42
5.14. Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia CSF . . . . .	42
5.15. Las 5 áreas con mayor diferencia en los tramos sin solapamiento. . . . .	43

5.16. Resultados de la clasificación en los 3 métodos de la segunda aproximación . . .	49
--	----

---

# Índice de tablas

---

2.1. Duración estimada . . . . .	6
2.2. Duración real en comparación con la planificación . . . . .	8
3.1. Ejemplos de una matriz de confusión . . . . .	19
4.1. Proporción de ejemplos con y sin punción lumbar . . . . .	23
4.2. Proporción de ejemplos con y sin patologías . . . . .	24
4.3. Cantidad de ejemplos por rangos de edad . . . . .	25
5.1. Ordenación de las 15 áreas cerebrales con mayor diferencia de GM . . . . .	30
5.2. Ordenación de las 15 áreas cerebrales con mayor diferencia de WM . . . . .	30
5.3. Ordenación de las 15 áreas cerebrales con mayor diferencia de CSF . . . . .	31
5.4. Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de GM	36
5.5. Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de WM . . . . .	36
5.6. Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de CSF . . . . .	36
5.7. Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia GM . . . . .	38
5.8. Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia WM . . . . .	39
5.9. Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia CSF . . . . .	39
5.10. Tasa de aciertos en la clasificación . . . . .	48
5.11. Matriz de confusión de la primera aproximación . . . . .	48
5.12. Matrices de confusión de la segunda aproximación . . . . .	48
5.13. Tasa de aciertos en la clasificación con las 15 áreas de mayor diferencia . . . . .	50
5.14. Matriz de confusión con método de acumulación para las 15 áreas con mayor diferencia . . . . .	50
5.15. Matriz de confusión con método de media para las 15 áreas con mayor diferencia	50
5.16. Matriz de confusión con método de contador para las 15 áreas con mayor diferencia . . . . .	51
5.17. Tasa de aciertos en la clasificación con las 15 áreas de mayor diferencia . . . . .	51

5.18. Matriz de confusión del método distancia media con umbral 0.75 . . . . .	51
6.1. Las áreas cerebrales con mayor diferencia según las pruebas de este trabajo . .	54
6.2. Clasificación de las áreas en los distintos lóbulos cerebrales . . . . .	55

---

# Lista de Algoritmos

---

5.1. Calculo de diferencia entre dos curvas . . . . .	29
5.2. Clasificación de pacientes usando curvas . . . . .	45



---

# Capítulo 1

## Introducción

---

### 1.1. Contexto

En 1906, el doctor Alois Alzheimer, un neuropatólogo y psiquiatra alemán, describió por primera vez la enfermedad que hoy en día lleva su nombre. [1]

El descubrimiento de la enfermedad de Alzheimer fue resultado del estudio de una paciente que ingresó en el Hospital de Frankfurt en 1901, a la edad de 51 años. La paciente presentaba un cuadro clínico muy llamativo que incluía delirios celotípicos, pérdida rápida y progresiva de memoria, alucinaciones, desorientación temporal y espacial, paranoia, trastornos de la conducta y un grave trastorno del lenguaje. La paciente fue identificada como Auguste Deter y fue estudiada por el doctor Alois Alzheimer hasta su muerte en 1906.

Después del fallecimiento de la paciente, durante la autopsia observó que la corteza cerebral era más estrecha de lo normal, además de poseer una acumulación de placas de proteína beta amiloide y ovillos neurofibrilares en las neuronas provocados por la proteína tau.

A pesar de las investigaciones del doctor Alois, durante la primera mitad del siglo XX la enfermedad se consideraba parte del envejecimiento normal en un ser humano. No fue hasta la década de los 60 cuando la comunidad científica declaró la enfermedad de Alzheimer como un tipo de demencia, y no como simple envejecimiento.

Desgraciadamente, después de años de investigaciones, aún no se ha logrado una cura contra la enfermedad. Tampoco se sabe cuál es la causa exacta, aunque se barajan algunas hipótesis:

- Origen genético: la mayoría de pacientes que sufren la enfermedad, han tenido algún familiar con Alzheimer. De hecho, si la enfermedad se presenta de forma temprana

(antes de los 60 años), es muy probable que se deba a esto.

- **Proteínas amiloide y tau:** es la hipótesis que más fuerza tiene a día de hoy. Se cree que el desarrollo de la enfermedad se debe a una excesiva acumulación de estas proteínas, en forma de placas y ovillos neurofibrilares.

El desarrollo de este trabajo se basa sobre todo en la segunda hipótesis: la presencia de las proteínas beta amiloide, tau (y fosfo tau), aumenta las posibilidades de padecer alzheimer.

### 1.2. Motivación

La detección temprana y precisa de la enfermedad del Alzheimer es fundamental para ralentizar su progreso y mejorar la calidad de vida del paciente. Actualmente, existen diversas técnicas para el diagnóstico de la enfermedad, incluyendo pruebas neuropsicológicas, biomarcadores en líquido cefalorraquídeo y neuroimagen estructural y funcional.

La neuroimagen estructural junto con la resonancia magnética, ha sido ampliamente utilizada para evaluar la atrofia cerebral en la enfermedad del Alzheimer. En particular, la medición del volumen del hipocampo ha demostrado ser un marcador útil de la enfermedad del Alzheimer temprana y puede ayudar en la detección temprana y el diagnóstico de la enfermedad.

Por otro lado, la neuroimagen funcional, como la resonancia magnética funcional, ha sido utilizada para investigar los cambios en la actividad cerebral en la enfermedad del Alzheimer. Estudios previos han demostrado que la actividad cerebral en regiones específicas del cerebro, como la corteza prefrontal y el lóbulo temporal, se altera en la enfermedad del Alzheimer.

Para la obtención de biomarcadores, la punción lumbar es una técnica utilizada en el diagnóstico de la enfermedad del Alzheimer y otras enfermedades neurodegenerativas. Consiste en la extracción de una muestra de líquido cefalorraquídeo de la columna vertebral, que se analiza para detectar la presencia de biomarcadores de la enfermedad.

Los biomarcadores del líquido cefalorraquídeo, como la proteína beta-amiloide y la proteína tau, se han utilizado como indicadores de la enfermedad del Alzheimer. Los niveles elevados de beta-amiloide en el líquido cefalorraquídeo se han asociado con la presencia de placas amiloides en el cerebro, mientras que los niveles elevados de proteína tau se han asociado con la presencia de ovillos neurofibrilares en el cerebro.

Aunque la punción lumbar puede proporcionar información valiosa para el diagnóstico de la enfermedad del Alzheimer, es un procedimiento invasivo que puede estar asociado con ciertos riesgos y complicaciones, como dolor de cabeza, sangrado o infección.

A pesar de todas estas técnicas de detección, el diagnóstico de la enfermedad solo puede realizarse con total seguridad post-mortem, al realizar la autopsia del cerebro.

Es por ello que agregando técnicas de estadística y aprendizaje automático a los métodos actuales [2], se podría llevar a cabo una detección más temprana de la enfermedad, lo que

permitiría seguir un tratamiento que favorezca su calidad de vida y retrase la evolución del Alzheimer [3]. Además, esta detección podría llevarse a cabo de manera mucho menos invasiva, gracias a que se evitaría la técnica de la punción lumbar para el diagnóstico.

### **1.3. Objetivo**

El objetivo de este proyecto es analizar la evolución del envejecimiento de las diferentes áreas cerebrales en pacientes con la presencia de biomarcadores de Alzheimer y pacientes sanos, con el fin de identificar patrones y características específicas que puedan ayudar en su detección temprana y tratamiento.

Para lograr este objetivo, se utilizará una base de datos con la información de cientos de pacientes de control (sanos) y con alguna característica que indique la posibilidad de la enfermedad.

Se utilizarán técnicas de estadística y aprendizaje automático para examinar los distintos ejemplos y comparar los resultados entre pacientes enfermos y sanos. Se aproximarán los datos con una curva mediante métodos de regresión. Estas aproximaciones facilitarán el proceso de comparación entre los distintos grupos de pacientes, además de ser un método muy visual que permitirá identificar la evolución de las distintas áreas a medida que el grupo de edad avanza.

El análisis se centrará en la identificación de cambios en el volumen de las diferentes áreas cerebrales, así como en su evolución con respecto a la edad y en la correlación de estos cambios con la presencia de biomarcadores (presencia de las proteínas beta amiloide, tau y fosfo tau).

En última instancia, el objetivo es contribuir al avance del conocimiento en el diagnóstico y tratamiento de la enfermedad de Alzheimer, y proporcionar nuevas herramientas para la detección temprana y la monitorización de la progresión de la enfermedad analizando únicamente los volúmenes cerebrales, sin necesidad de recurrir a una comprobación de biomarcadores invasiva, como lo es la punción lumbar.



---

## Capítulo 2

# Planificación

---

### 2.1. Diagrama EDT

Con el objetivo de establecer un orden y un mejor seguimiento durante el desarrollo del trabajo, se ha decidido descomponerlo en bloques de trabajo. Cada bloque agrupa un conjunto de tareas con características comunes. Los bloques que componen este trabajo son los siguientes:

- Paquete **planificación**: Recoge las tareas relacionadas con la planificación y seguimiento y control del proyecto, entre las que se encuentran la elaboración de éste capítulo o reuniones de seguimiento con la tutora.
- Paquete **desarrollo del trabajo**: Recoge las tareas relacionadas con la investigación, implementación, obtención y análisis de resultados. Es el grueso del trabajo, por lo que tendrá una dedicación de tiempo mayor.
- Paquete **documentación**: Se incluyen las tareas de elaboración de la memoria y preparación de la defensa. Junto con el paquete anterior, son los más importantes.

En la figura 2.1 se muestra el diagrama EDT con la descomposición de tareas a realizar, de tal manera que el desarrollo del trabajo sea más ordenado y estructurado.



Figura 2.1: Diagrama EDT

Tarea	Duración Estimada (en horas)	Duración del paquete (en horas)
1. Planificación	10	40
2. Seguimiento y control	30	
3. Investigación	70	140
4. Implementación	60	
5. Análisis resultados	10	
6. Memoria	70	120
7. Defensa	50	
Total	300	300

Tabla 2.1: Duración estimada

## 2.2. Desarrollo Estimado

Con el objetivo de llevar un mejor seguimiento y organización del proyecto, se ha estimado un tiempo para cada tarea. En la tabla 2.1 se presenta la duración estimada para el desarrollo de las tareas, así como los períodos de tiempo en los que se llevarán a cabo, en la figura 2.2. Hay que destacar que las tareas no tienen por qué comenzar y terminar exactamente en el momento que se indica en el diagrama, simplemente deberá desarrollarse en el periodo indicado.

Como se mencionó en la sección anterior, los paquetes de desarrollo y documentación son los más extensos, por lo que se estima una mayor duración y período de realización.

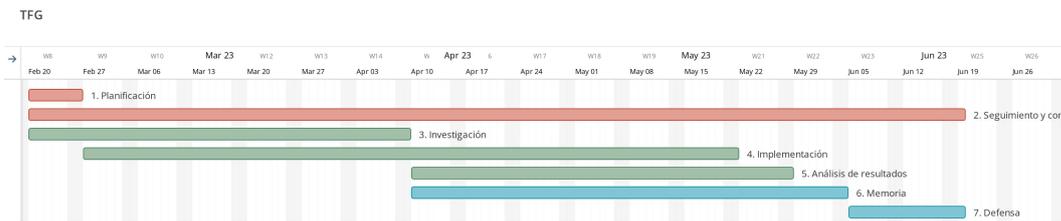


Figura 2.2: Diagrama Gantt con los períodos planificados

## 2.3. Gestión de riesgos

Conviene analizar los posibles riesgos que puedan surgir durante el desarrollo del proyecto, con el objetivo de prevenirlos o subsanarlos, en caso de producirse.

1. **Retrasos en los plazos de las tareas.** Cabe la posibilidad de producirse retrasos en algunas de las tareas, especialmente en los puntos 4 (Implementación) y 6 (memoria), en parte debido a que paralelamente se están cursando algunas asignaturas. Por ello se ha planificado dedicar más tiempo del que podría necesitarse en un primer momento.
2. **Incomprensión de la información.** Debido a la falta de experiencia en el área de la neurología, ciertos términos o datos pueden resultar difíciles de comprender. Por eso la tarea 3 (Investigación) abarca prácticamente el total desarrollo del proyecto. Esto permite mayor flexibilidad y tiempo para poder estudiar los conceptos sin problema. Adicionalmente, se podrían llevar a cabo reuniones extra con la tutora que podrían ayudar.
3. **Desviación del objetivo principal del proyecto.** El área de trabajo es tan amplia que existe la posibilidad de desviarse del objetivo principal, o no realizar correctamente las tareas indicadas. Reuniones con la tutora cada dos semanas aproximadamente servirán para comprobar que el trabajo se está desarrollando de manera correcta.
4. **Pérdida del trabajo.** También debe de tenerse en cuenta la posibilidad de perder parte del desarrollo del proyecto o su totalidad. Por si acaso, se almacenará la información de forma local en el ordenador personal. También se guardará todo en la nube, además de tener realizar una copia de seguridad con la implementación antes de realizar algún cambio importante.
5. **Errores en la implementación.** Errores en la implementación de los programas es uno de los riesgos más importantes, ya que puede desencadenar otros problemas. Por lo tanto se comprobarán todos los avances realizados, además de mostrarlos en las reuniones y comentar sus resultados.

## 2. PLANIFICACIÓN

---

Tarea	Duración Estimada (en horas)	Duración real (en horas)	Diferencia (en horas)
1. Planificación	10	10	0
2. Seguimiento y control	30	30	0
3. Investigación	70	70	0
4. Implementación	60	75	15
5. Análisis resultados	10	5	-5
6. Memoria	70	75	5
7. Defensa	50	40	-10
8. Póster*	-	5	5
Total	300	310	10

**Tabla 2.2:** Duración real en comparación con la planificación

### 2.4. Desarrollo realizado

A lo largo del desarrollo del trabajo, se ha documentado el tiempo real invertido en cada una de las tareas. En la tabla 2.2 puede compararse el tiempo realizado con el tiempo estimado. Puede verse la aparición de una nueva tarea, denominada "8.Póster", no planificada, introducida dentro del paquete de documentación.

La elaboración del póster era una tarea que no se tuvo en cuenta al principio, por lo que tuvo que añadirse en la fase final del desarrollo. Esto provocó un pequeño aumento en el número de horas realizadas, que sumadas al total hacen 310 horas aproximadamente, 10 horas más que lo planificado en la fase inicial del proyecto.

Puede verse además un aumento significativo a la dedicación de las tareas *4.Implementación* y *6.Memoria*. Esto se debe a los malos resultados obtenidos en la fase final de implementación, los cuales trataron de ser mejorados mediante el uso de otras técnicas, como se relata al final el capítulo 5, para posteriormente documentarlos en la memoria.

---

## Capítulo 3

# Fundamentos

---

### 3.1. Enfermedad de Alzheimer

Para poder trabajar en el proyecto, es indispensable conocer más en profundidad la enfermedad de Alzheimer, cómo se origina, su detección y diagnóstico, así como sus síntomas más comunes, para hacernos una idea de qué regiones del cerebro pueden ser las más afectadas.

#### 3.1.1. Definición

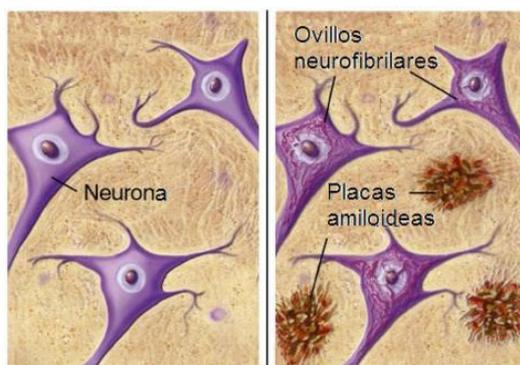
Según la asociación internacional de Alzheimer [4], se define como un tipo de demencia que causa problemas con la memoria, el pensamiento y el comportamiento. Los síntomas generalmente se desarrollan lentamente y empeoran con el tiempo, hasta que son tan graves que interfieren con las tareas cotidianas. La enfermedad hace que el cerebro encoja y que las neuronas cerebrales acaben muriendo.

Por lo general afecta a personas mayores de 65 años, pero pueden darse signos tempranos de la enfermedad en individuos menores de 65. Se estima que, de todos los casos de demencia que hay en el mundo, alrededor del 60 % son casos de Alzheimer.

#### 3.1.2. Posibles causas

Aún se desconocen las causas exactas que provocan la enfermedad de Alzheimer. Pero sí se sabe que puede deberse a que algunas proteínas no funcionan correctamente, lo que hace que las neuronas se dañen, pierdan conexiones y finalmente mueran.

A esto hay que añadirle también una combinación de factores genéticos, ambientales y de estilo de vida, que pueden propiciar la aparición de la enfermedad.



**Figura 3.1:** Efecto de las proteínas Beta-amiloide y tau en el cerebro

Hoy en día las investigaciones se centran en el estudio de dos proteínas [5], Beta-amiloide y Tau (figura 3.1 [6]):

- **Beta-amiloide:** la excesiva agrupación de estas proteínas en el cerebro parecen tener un efecto tóxico en las neuronas, que interrumpe su conexión. A dichas agrupaciones se les llama placas amiloides o placas seniles.
- **Tau:** la proteína tau se encarga de transportar nutrientes a las neuronas. En la enfermedad de Alzheimer, estas proteínas se organizan en forma de ovillos neurofibrilares, que interrumpen el transporte de nutrientes.

#### 3.1.3. Síntomas

La pérdida de memoria es el síntoma clave de la enfermedad de Alzheimer. Los primeros signos incluyen dificultad para recordar eventos o conversaciones recientes, pero la memoria empeora y se desarrollan otros síntomas a medida que la enfermedad avanza. [7]

Hacer más de una tarea a la vez resulta especialmente difícil. Llevar a cabo tareas que requieran el uso de los números resulta complicado en las últimas fases de la enfermedad.

La enfermedad de Alzheimer provoca un deterioro en la capacidad de tomar decisiones y juicios razonables en situaciones cotidianas. Por ejemplo, la persona puede tomar malas decisiones en ámbitos sociales o usar ropa que no es adecuada para las condiciones climáticas.

Las actividades rutinarias que requieren completar pasos en orden se convierten en un problema. Esto puede incluir por ejemplo, planear y elaborar una comida.

Los cambios que la enfermedad de Alzheimer produce en el cerebro pueden afectar el estado de ánimo y el comportamiento. Los problemas pueden incluir depresión, pérdida de interés, aislamiento social, desconfianza o delirios, entre otras.

### 3.1.4. Fases de la enfermedad

Aunque cada persona experimenta la enfermedad de manera distinta, por lo general se pueden diferenciar tres fases principales:

1. **Fase temprana:** Se experimentan síntomas leves, como olvidos, dificultad para recordar información reciente o pequeños cambios en la personalidad.
2. **Fase moderada:** Los síntomas empeoran y se comienza a tener problemas a la hora de realizar tareas cotidianas como vestirse o cocinar. Es la fase más prolongada de la enfermedad.
3. **Fase avanzada:** El paciente pierde la capacidad de comunicarse y de realizar actividades básicas como alimentarse o controlar la vejiga. Debido a la falta de movilidad, pueden presentarse otros problemas de salud, como problemas en la piel o infecciones respiratorias.

Para retrasar lo máximo posible el avance de la enfermedad, es muy importante detectarla en las fases más tempranas.

### 3.1.5. Detección

Es importante saber que no se puede diagnosticar Alzheimer con el 100 % de exactitud hasta que no se realiza una autopsia post-mortem. Sin embargo existen diversos métodos que permiten detectarlo con gran seguridad.

Primeramente, se llevan a cabo una serie de evaluaciones clínicas y neuropsicológicas. Suelen consistir en entrevistas con un profesional o pruebas de memoria, que permiten evaluar el grado de deterioro cognitivo. Adicionalmente pueden realizarse entrevistas con familiares o amigos, con el fin de encontrar detalles que puedan sospechar el padecimiento de la enfermedad.

Si los resultados de estas primeras evaluaciones dan indicios de la enfermedad, entonces se procede a hacer un diagnóstico más exhaustivo, en el que se incluyen la obtención de imágenes cerebrales mediante resonancia magnética (MRI) o tomografías por emisión de positrones (PET). Dichas imágenes muestran cambios en el cerebro que nos indican la presencia de la enfermedad. Aunque si el paciente aún se encuentra en una etapa muy temprana, puede que no se detecte.

También se utilizan biomarcadores, medidas biológicas que indican la presencia de la enfermedad. En este caso se comprueba la presencia elevada de proteínas beta amiloide y tau, mediante una punción lumbar de la que se extrae líquido cefalorraquídeo. Se trata de una prueba muy invasiva, pero puede ayudar en gran medida a la detección. [8]

También se realizan pruebas genéticas, ya que en algunos casos la aparición del Alzheimer es una cuestión hereditaria. Estas pruebas pueden ayudar a detectar mutaciones y predecir el riesgo de desarrollar la enfermedad.

## 3.2. El cerebro

La evolución del cerebro, así como sus áreas a lo largo de la vida difiere entre las personas sanas y aquellos que padecen Alzheimer. Además, es el órgano que más se ve afectado a medida que la enfermedad avanza.

El cerebro es un órgano complejo, ubicado dentro del cráneo, que gestiona la actividad del sistema nervioso. Forma parte del Sistema Nervioso Central (SNC) y constituye la parte más voluminosa y conocida del encéfalo. Dentro del cráneo, el cerebro flota en un líquido transparente, llamado líquido cefalorraquídeo, que cumple funciones de protección, tanto físicas como inmunológicas. [9]

Está compuesto por más de 100 mil millones de neuronas. Algunos grupos específicos de ellas, trabajando en conjunto, nos dan la capacidad para razonar y pensar, para experimentar sentimientos y para comprender el mundo. También nos dan la capacidad para recordar cantidades diversas de información, además de regular y mantener cada función vital del cuerpo, controlar los movimientos y procesar la información recibida a través de los sentidos. Son tantas las funciones que cumple lo que hace del cerebro uno de los órganos más importantes del cuerpo humano.

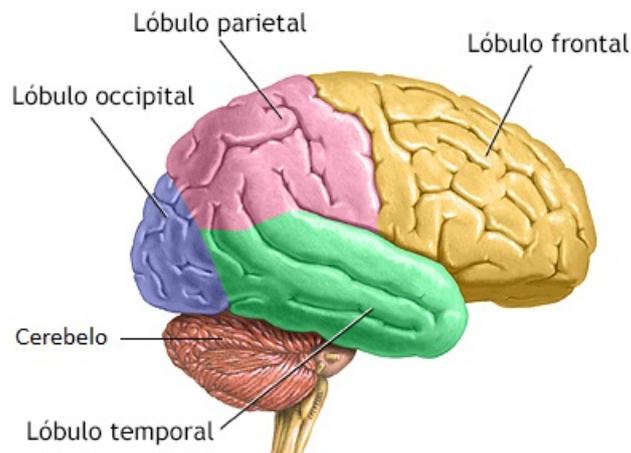
### 3.2.1. Áreas cerebrales

El cerebro consta de dos hemisferios: hemisferio derecho y hemisferio izquierdo. La parte derecha está relacionada con la expresión no verbal: percepción, conducta emocional, orientación espacial... Por otro lado, el hemisferio izquierdo está relacionado con la parte verbal: controla la expresión oral, el habla y la escritura. También cumple otras funciones como la capacidad de análisis, razonamiento lógico o la resolución de problemas numéricos. [10]

A su vez, podemos dividir el cerebro en 4 lóbulos [11], [12] y el cerebelo (figura 3.2, [13]):

- **Lóbulo Frontal:** controla el razonamiento, las emociones, parte del habla y de los movimientos.
- **Lóbulo Parietal:** controla las sensaciones del contacto, el dolor, la temperatura y partes del habla.
- **Lóbulo Temporal:** controla la memoria, sentido del oído, y partes del habla.
- **Lóbulo Occipital:** controla la visión.
- **Cerebelo:** parte trasera del cerebro y la responsable de la coordinación y el equilibrio.

Además, cada uno de estos lóbulos se subdivide en regiones más pequeñas, y complejas.



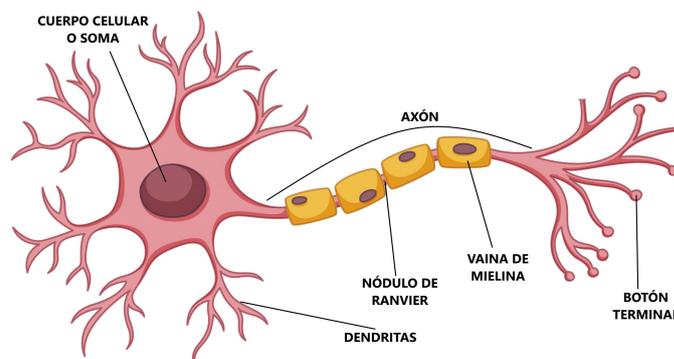
**Figura 3.2:** División del cerebro en lóbulos

### 3.2.2. Las neuronas

La neurona, es la unidad fundamental que forma el cerebro. Es la célula responsable de recibir información del mundo exterior y de hacer llegar las señales para la respuesta a estos desde los diferentes órganos.

Una neurona (figura 3.3 [14]) tiene un cuerpo celular o soma, que incluye el núcleo celular, y extensiones especiales denominadas axones y dendritas. Los axones y las dendritas permiten que las neuronas se comuniquen, incluso a través de largas distancias. Los impulsos eléctricos llegan a la neurona a través de las dendritas, se procesan en el soma y se transmite al axón, que se comunica con las dendritas de la siguiente neurona.

Un paciente con la enfermedad de Alzheimer sufre un deterioro de las neuronas muy importante, ya que éstas se ven afectadas tanto por la acumulación de placas amiloides que interrumpen sus conexiones, o por la formación de ovillos neurofibrilares que interrumpen el transporte de nutrientes.



**Figura 3.3:** La neurona y sus partes

### 3.3. Aprendizaje automático

El aprendizaje automático es el proceso mediante el cual se usan modelos matemáticos de datos para aprender sin necesidad de programar toda la casuística directamente, de manera que la computadora aprende en función de los datos que recibe. [15]

Se considera un subconjunto de la inteligencia artificial. El aprendizaje automático usa algoritmos para identificar patrones en los datos, y esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones o clasificación. Con más experiencia y datos, los resultados del aprendizaje automático son más precisos, de forma muy similar a cómo los humanos mejoran con más práctica.

La adaptabilidad del aprendizaje automático lo convierte en una excelente opción en escenarios en los que los datos siempre cambian, ya que el modelo se adapta a la naturaleza de los datos, con adaptaciones mínimas en el código. De lo contrario se debería de implementar un programa distinto en función del tipo de datos que se tenga.

Existen varios tipos de modelo de aprendizaje automático, entre los que destacan: [16]

- **Aprendizaje supervisado:** los datos están etiquetados, lo que permite al modelo hacer una predicción y compararla con la etiqueta a fin de perfeccionar la predicción
- **Aprendizaje no supervisado:** los datos no están etiquetados, por lo que se extrae información y patrones previamente desconocidos, agrupando los datos en conjuntos u observando cómo se distribuyen.
- **Aprendizaje por refuerzo:** el modelo observa el entorno y toma aquellas decisiones que minimizan el error o maximizan una recompensa.

#### 3.3.1. Técnica de Regresión

Dentro del aprendizaje supervisado, se nos presentan dos técnicas principales [17]: clasificación y regresión (figura 3.4). La clasificación consiste en entrenar un modelo para asignar una etiqueta de clase, como por ejemplo: dado un paciente junto con los resultados de algunas pruebas, determinar si tiene Alzheimer o no.

Por otra parte, la regresión consiste en la predicción de un valor dependiente de una o más características de entrada (llamadas variables predictoras). Un ejemplo de ello puede ser el siguiente: dado un paciente junto con los resultados de algunas pruebas, predecir o averiguar su volumen cerebral a la edad de 80 años.

Existen diversos algoritmos de regresión, el más sencillo de todos es la regresión lineal. La regresión lineal es una técnica estadística que se utiliza para analizar la relación entre dos variables, una variable independiente y otra variable dependiente. En la regresión lineal se asume que la relación entre dos variables es lineal, y que el conjunto de datos puede ser aproximado con una línea recta definida por la ecuación:

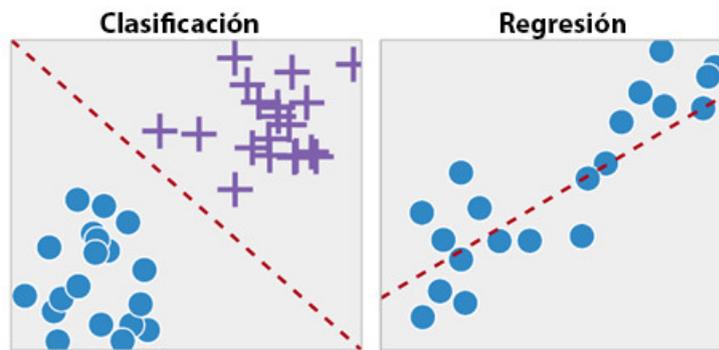


Figura 3.4: Diferencia entre clasificación y regresión

$$y = m \cdot x + b \quad (3.1)$$

Donde  $y$  es la variable dependiente,  $x$  es la variable independiente,  $m$  es la pendiente y  $b$  el punto de intersección en el eje  $y$ .

Para encontrar la recta que mejor aproxima los datos se utiliza el método de mínimos cuadrados, que consiste en minimizar la suma de las distancias entre los puntos de datos y la recta.

Cuando se quiere una aproximación que se ajuste mejor a los datos, o si se ve que los datos pueden no tener una relación lineal, en ese caso se deben utilizar otras técnicas de regresión no lineales, las cuales aproximan los datos mediante una curva más compleja, como una parábola, un polinomio o una exponencial.

### 3.3.1.1. Regresión Ridge

En ocasiones podemos destinar un subconjunto de los datos para obtener la curva de regresión (conjunto de entrenamiento), y el resto de datos para comprobar lo buena que es dicha aproximación (conjunto de test). Puede darse el caso de que la curva (o modelo) se ajuste perfectamente al conjunto de entrenamiento, sin embargo no se ajuste de manera adecuada al conjunto de test. En ese caso decimos que tenemos un modelo sobreajustado al conjunto de entrenamiento, y con un alto valor de varianza. Este modelo no es adecuado para su uso.

Una de las soluciones a este problema es la regresión Ridge [18]. Ridge es un tipo de regresión que trata de encontrar un modelo que se ajuste menos al conjunto de entrenamiento, con el objetivo de reducir la varianza.

Generalmente, en los métodos de regresión se obtiene la curva óptima minimizando la suma de distancias al cuadrado. La regresión Kernel añade a esta ecuación una penalización proporcional al cuadrado del valor de las variables dependientes.

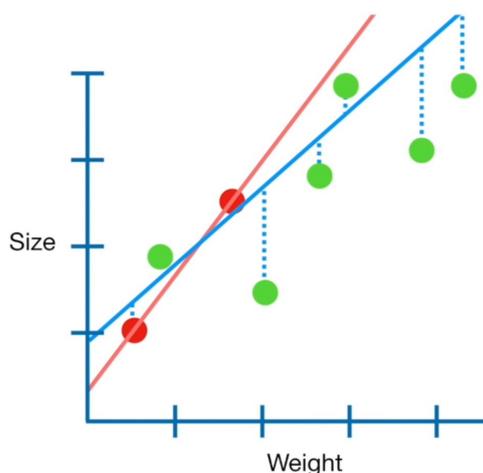


Figura 3.5: Ejemplo del efecto de  $\lambda$  en la regresión Ridge

$$y = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (3.2)$$

Donde:

- $\sum_{i=1}^n (y_i - f(x_i))^2$  es la suma de distancias punto-recta al cuadrado.
- $\lambda$  es la proporción de penalización.
- $\sum_{j=1}^p \beta_j^2$  es la suma del cuadrado de las variables dependientes.
- $\beta_j$  es la  $j$ -ésima variable dependiente.

Esta penalización está multiplicada por un valor  $\lambda$  positivo, que indica la fuerza de dicha penalización. Mayores valores de lambda hacen que el modelo sea menos sensible a los datos. Se puede escoger un valor óptimo de lambda de forma manual o con técnicas como la validación cruzada.

En la figura 3.5 [19], puede observarse un ejemplo básico de cómo afecta  $\lambda$  a la generación de la curva, en ese caso una regresión lineal. Se realiza la regresión Ridge sobre un conjunto de entrenamiento (puntos rojos). La línea roja no tiene penalización ( $\lambda = 0$ ), la línea azul tiene un valor positivo de  $\lambda$ , que aunque aumenta la distancia punto-recta en el conjunto de entrenamiento, en el conjunto de test (puntos verdes) aporta mejores resultados.

Además de evitar el sobreajuste, la regresión Ridge ofrece otra ventaja: aporta mejores resultados cuando el conjunto de datos es multicolineal, es decir, cuando existen varios valores de la variable independiente para los mismos valores de la variable dependiente.

### 3.3.2. Normalización y estandarización

La normalización de datos [20] es un proceso utilizado para ajustar los valores de un conjunto de datos a una escala común, con el fin de eliminar las diferencias en la magnitud de las variables. El objetivo de la normalización es hacer que las variables sean comparables y para mejorar la precisión de los análisis estadísticos y modelos predictivos que se basan en los datos. La normalización de datos se realiza al transformar los valores originales de las variables a nuevos valores que están en una escala común. Algunos de los más utilizados son

Dividir cada valor por la suma total de valores:

$$x_{i-norm} = \frac{x_i}{\sum_{j=1}^n x_j} \quad (3.3)$$

Normalización Min-Max:

$$x_{i-norm} = \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \quad (3.4)$$

En ambos casos, el rango de los valores normalizados se transforma a una escala [0, 1].

Al normalizar los datos, se logra una distribución de valores en una escala común, lo que facilita la comparación y la identificación de patrones y tendencias en los datos. Además, la normalización ayuda a reducir la influencia de valores atípicos o extremos, lo que puede distorsionar los resultados de análisis estadísticos o modelos predictivos.

Por otra parte, la estandarización es un proceso similar, ajusta el conjunto de datos para que tengan una media de 0 y desviación típica de 1. Es un proceso muy útil para comparar diferentes conjuntos de datos, o para preprocesarlos si se quieren utilizar para algún tipo de análisis estadístico.

El método de estandarización se lleva a cabo mediante la siguiente fórmula:

$$x_{i-estandarizado} = \frac{(x_i - \mu)}{\sigma} \quad (3.5)$$

Donde  $\mu$  es la media de la muestra y  $\sigma$  la desviación estándar.

#### 3.3.3. Bootstrapping e intervalos de confianza

La técnica de Bootstrapping [21] es un método estadístico que se utiliza comúnmente para la estimación de un estadístico de interés (como por ejemplo, la media) basado en muestras aleatorias.

Para llevarlo a cabo, se generan múltiples muestras aleatorias con reemplazo a partir de una muestra original, cuyo tamaño es igual al tamaño de la muestra original. Se calcula el estadístico para cada una de las muestras, y tras repetirlo varias veces, se puede obtener una distribución empírica del estadístico.

El procedimiento de bootstrapping resulta de gran utilidad para obtener intervalos de confianza. Un intervalo de confianza es un rango de valores que se utiliza para estimar un parámetro desconocido de una población basándose en una muestra de datos. Aporta un margen de error dentro del cuál se espera que se encuentre el verdadero valor del parámetro con un determinado nivel de confianza.

El intervalo de confianza se encuentra en un rango de valores. Por ejemplo:

*“La media de una variable es [10, 15] con un intervalo de confianza del 95 %”*

Esto significa que el valor real de la media se encuentra entre los valores 10 y 15 con una probabilidad del 95 %, es decir, si nosotros tomamos varias muestras y calculamos la media de cada una, el 95 % de las veces el valor de la media se encontrará en el rango [10, 15].

El uso de intervalos de confianza resulta de gran utilidad cuando existe una gran incertidumbre en los datos de la muestra, ya que permiten evaluar la fiabilidad de los resultados que nos aporta una muestra.

#### 3.4. Matriz de confusión

La matriz de confusión [22] es una herramienta utilizada en el campo de la estadística y el aprendizaje automático para evaluar el rendimiento de un modelo de clasificación. Consiste en una tabla que muestra la cantidad de predicciones correctas e incorrectas realizadas por el modelo en relación con las clases reales de los datos.

Por lo general, las filas de la matriz representan los valores reales de los ejemplos, mientras que las columnas representan los valores predichos por el modelo clasificador. de tal manera que cada celda de la matriz representa una combinación de etiqueta real con etiqueta predicha.

En el caso de una clasificación binaria, es decir cuando se clasifica entorno a dos clases, la matriz de confusión tiene cuatro componentes:

- Verdaderos positivos (TP): representa el número de casos que el modelo clasificó correctamente como positivos.

		Predicción	
		Negativo	Positivo
Real	Negativo	TN	FP
	Positivo	FN	TP

**Tabla 3.1:** Ejemplos de una matriz de confusión

- Falsos positivos (FP): indica el número de casos que el modelo clasificó como positivos de manera incorrecta.
- Verdaderos negativos (TN): muestra el número de casos correctamente clasificados como negativos.
- Falsos negativos (FN): indica el número de casos que el modelo clasificó incorrectamente como negativos.

La tabla 3.1, muestra un ejemplo de una matriz de confusión para un modelo de clasificación binaria.

La matriz de confusión permite también obtener algunas métricas de evaluación que permiten comprobar lo bueno que es el modelo de clasificación, como puede ser la tasa de aciertos.

La tasa de aciertos o *Accuracy*, representa la proporción de resultados verdaderos (tanto positivos como negativos) con respecto al número total de casos. Se calcula de la siguiente manera:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \quad (3.6)$$

A partir de la matriz de confusión pueden obtenerse muchas otras métricas, como la precisión, *Recall*, *F1 score*, *Specitifty*... Lo que hacen de esta matriz una forma muy completa de evaluar un clasificador.



---

## Capítulo 4

# La base de datos

---

Para llevar a cabo este trabajo, se proporcionó una base de datos por parte de CITA-Alzhéimer. En ella, se recoge la información de dos proyectos: PGA y DEBA.

El proyecto PGA (Proyecto Gipuzkoa Alzhéimer) tiene como objetivo la caracterización de la enfermedad de Alzhéimer en la fase preclínica, es decir, buscar características y patrones comunes en individuos antes de llevar a cabo un diagnóstico, investigando los biomarcadores que afectan a la enfermedad y los factores de riesgo.

Por otro lado, el proyecto DEBA pretende mejorar la detección y busca posibles vías para la prevención de la enfermedad. Han recabado información de voluntarios mayores de 60 años en la localidad de Deba (Gipuzkoa).

Gracias a ambos proyectos, se ha logrado una base de datos de un total de 537 individuos, en los que se incluye información relevante como la edad, presencia de biomarcadores o volumen cerebral.

### 4.1. Contenido de la base de datos

Se utilizarán los ejemplos de ambos proyectos indistintamente. Para cada persona se disponen de la siguiente información:

1. **cmp\_code**: Código del paciente. Un índice que sirve para identificar el ejemplo en la base de datos
2. **Project**: indica el proyecto del que proviene el ejemplo (*PGA* O *DEBA*)
3. **Gender**: género del paciente (0 para mujer, 1 para hombre)

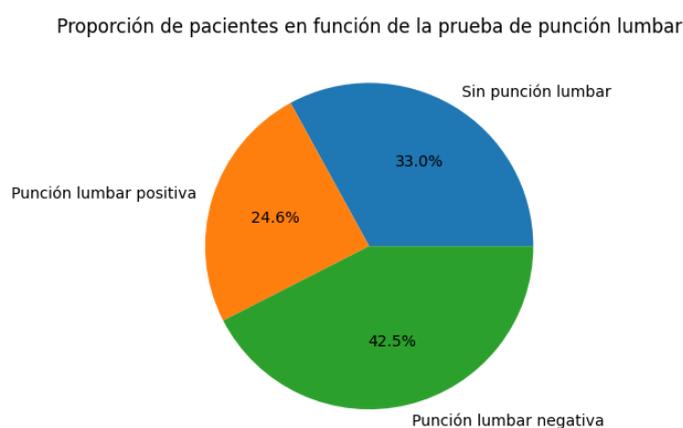
4. **Age:** edad del paciente. Se tienen edades comprendidas entre los 40 y 90 años. En la base de datos, este valor es continuo. Para ciertas cuestiones, se llevará un redondeo al entero más cercano
5. Resultado de la punción lumbar:
  - **NA:** no se ha llevado a cabo.
  - **0:** se ha llevado a cabo y no presenta biomarcadores.
  - **1:** se ha llevado a cabo y presenta biomarcadores.
6. Presencia de biomarcadores determinantes para el diagnóstico del Alzheimer:
  - beta-amiloide (CSF\_AB42\_Cutoff\_LUD)
  - tau (CSF\_TTAU\_Cutoff\_LUD)
  - fosfo-tau (CSF\_PTAU\_Cutoff\_LUD).

Estas columnas pueden tener los siguientes valores:

- **9999:** no se ha llevado a cabo punción lumbar, valor desconocido.
  - **0:** no presenta el biomarcador.
  - **1:** presenta el biomarcador.
7. **TIV:** Volumen intracerebral total, es decir, el volumen total del cerebro del paciente
  8. Cantidad de materia gris (GM = Gray Matter), materia blanca (WM = White Matter) y líquido cefalorraquídeo (CSF = cerebrospinal fluid), de un total de 68 áreas cerebrales. Por lo tanto, se disponen de un total de 204 columnas con esta información: 68 áreas × 3 materias.

Punción lumbar	Cantidad	Proporción
No	177	33 %
Si (positiva)	132	24,5 %
Si (negativa)	228	42,5 %
Total	537	100 %

**Tabla 4.1:** Proporción de ejemplos con y sin punción lumbar



**Figura 4.1:** Proporción de ejemplos con y sin punción lumbar

## 4.2. Análisis del contenido

Del total de 537 personas, 177 no se han sometido a la prueba de punción lumbar, mientras que los 360 restantes sí que lo han hecho. De los 360 individuos sometidos a la punción, 228 han dado resultado negativo en biomarcadores, es decir, no poseen ninguna de las 3 proteínas indicadoras de un posible desarrollo de Alzheimer. Las 132 personas restantes, han dado positivo en la prueba de punción (ver tabla 4.1 y la figura 4.1).

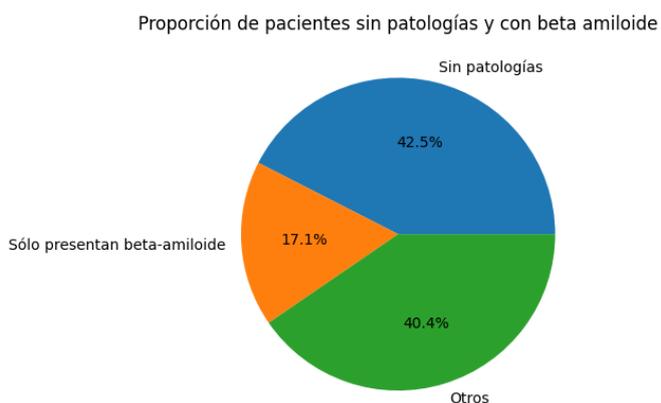
Los profesionales de CITA-Alzheimer comentaron que de todas las proteínas, la beta-amiloide es la más importante para el diagnóstico del Alzheimer, es por eso que a petición suya, para el trabajo se utilizarán los datos de las personas consideradas sanas (aquellas que se han sometido a la punción lumbar y no presentan ningún biomarcador) y personas con presencia de beta-amiloide. De los 537, tan solo 92 presentan solo dicha proteína. Hay que destacar que esta cantidad representa a las personas que tienen solo beta amiloide, y no presentan ninguna de las otras dos proteínas (tau y fosfo-tau). Esta información se muestra de manera gráfica en la tabla 4.2, la figura 4.2 con todos los ejemplos y figura 4.3 solo con los ejemplos utilizados).

También es importante mencionar, que en el grupo *Otras/desconocido* se encuentran los pacientes que no pertenecen a los grupos anteriores (sin patologías/con beta-amiloide) y aquellos cuya información se desconoce por no haberse sometido a la punción lumbar.

Ya que se va a hacer un estudio en función de la edad de los pacientes, conviene ver cómo están distribuidos los rangos de edad en la base de datos. Definiendo rangos de edad de 5 años, los rangos con mayor número de ejemplos son:

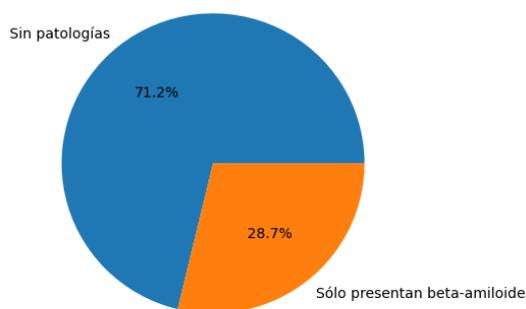
Patología	Cantidad	Proporción
Sin patologías	228	42,5 %
Sólo beta-amiloide	92	17,1 %
Otras/desconocido	217	40,4 %
Total	537	100 %

**Tabla 4.2:** Proporción de ejemplos con y sin patologías



**Figura 4.2:** Proporción de ejemplos según sus patologías, para todos los ejemplos de la bd

Proporción de pacientes sin patologías y con beta amiloide en los ejemplos utilizados



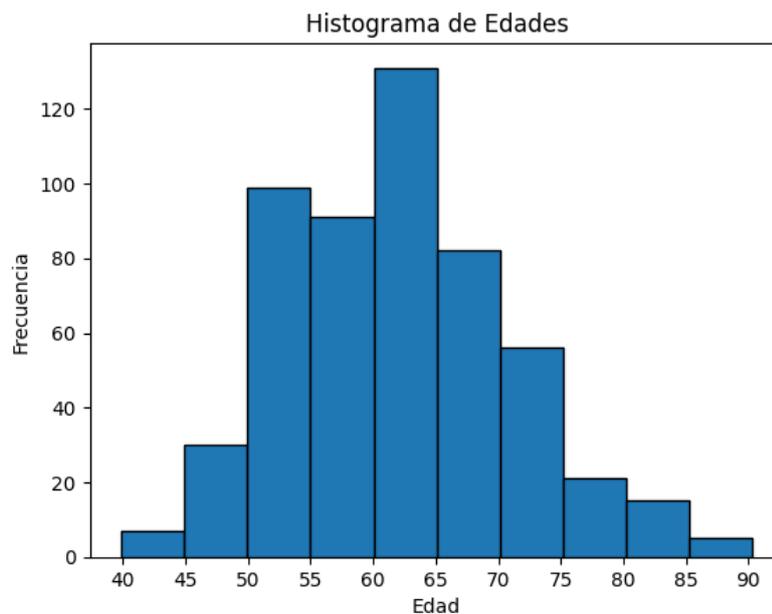
**Figura 4.3:** Proporción de ejemplos según sus patologías, solo para los ejemplos utilizados

1. Rango (60, 65]: 130 ejemplos
2. Rango (50, 55]: 99 ejemplos
3. Rango (55, 60]: 84 ejemplos

Por otra parte, existen rangos de edad en la que apenas hay datos. Por ejemplo, en el rango (40, 45) únicamente existen 7 ejemplos, y en el rango (85, 90] se tienen 5. Esta información se debe tener en cuenta a la hora de trabajar con los datos, ya que un escaso número de ejemplos en ciertos rangos puede provocar resultados menos fiables y más sobreajustados. En la tabla 4.3 y la figura 4.4 se puede observar el número de ejemplos disponibles distribuidos por rangos de edad de 5 años.

Rango	Cantidad	Proporción
(40, 45]	7	1,30 %
(45, 50]	30	5,59 %
(50, 55]	99	18,44 %
(55, 60]	90	16,76 %
(60, 65]	130	24,21 %
(65, 70]	84	15,64 %
(70, 75]	55	10,24 %
(75, 80]	21	3,91 %
(80, 85]	15	2,79 %
(85, 90]	6	1,12 %
Total	537	100 %

**Tabla 4.3:** Cantidad de ejemplos por rangos de edad



**Figura 4.4:** Histograma con la distribución de los ejemplos por rangos de edad

### 4.3. Preprocesado

Además de la selección, se ha llevado a cabo un preprocesamiento de datos, que consiste en la normalización y estandarización de los mismos.

Primeramente se han normalizado todos los volúmenes en función del volumen intracraneal total de cada individuo, de tal manera que ahora tenemos en cada área, la proporción de materia que tiene el individuo.

Por otra parte, también se han estandarizado los datos, para que sigan una distribución de media 0 y desviación típica 1. Unos datos correctamente estandarizados facilitan el análisis y la aplicación de técnicas estadísticas.

#### 4.4. Nomenclatura

Las columnas que contienen la información de los volúmenes de las 68 áreas tienen la siguiente nomenclatura:

$$\text{nombre\_area} + \underbrace{L/R}_1 + \underbrace{GM/WM/CSF}_2 + \underbrace{VOL.XX}_3$$

1. L para las áreas del hemisferio izquierdo, R para las áreas del hemisferio derecho.
2. GM si indica materia gris, WM para materia blanca, CSF para líquido cefalorraquídeo.
3. XX es número de área asignado.

Por ejemplo, el área de nombre:

*Cerebellum\_R\_GM\_VOL.18*

Hace referencia a la cantidad de materia gris que se encuentra en el área derecha del cerebelo, cuyo número de área es el 18.

---

## Capítulo 5

# Proceso de implementación

---

### 5.1. Entorno de trabajo

Para desarrollar este trabajo, se decidió utilizar el lenguaje de programación python por las siguientes razones:

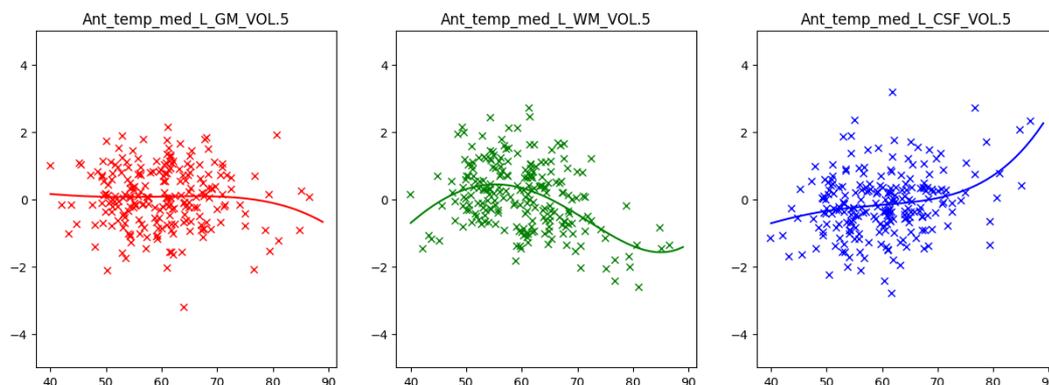
Primeramente por ser un lenguaje con una sintaxis simple que facilita la rápida comprensión del código, además de ser un lenguaje que se ha utilizado mucho a lo largo de la carrera.

Por otra parte, posee una amplia variedad de librerías y herramientas específicamente diseñadas para la estadística, minería de datos o machine learning. Algunas de ellas son Numpy para la manipulación de matrices, Pandas para trabajar con grandes cantidades de datos, scikit-learn para aplicar técnicas de machine learning, o matplotlib para elaborar gráficas y mostrar resultados.

Además, gracias al uso de Google Colab, una herramienta que permite ejecutar código python en el navegador, se ha podido compartir el proceso de implementación con la tutora semanalmente y sin ningún problema.

### 5.2. Aproximación con regresión Ridge

Tras llevar a cabo el proceso de normalización y estandarización los datos de los volúmenes, el siguiente paso es aproximar todos los datos de cada área y grupo de pacientes en función de la edad. Dicha aproximación se ha llevado a cabo mediante la regresión Ridge, un tipo de regresión que ofrece muy buenos resultados cuando existe colinealidad en el conjunto de datos. La variable predictora será la edad, y se predecirá el volumen de materia gris, blanca o líquido cefalorraquídeo de cada área cerebral.



**Figura 5.1:** Curvas de las 3 materias del área número 5 para el grupo de pacientes sin patologías

Se aproximará en un rango de edad inicial de 40 a 90 años, en la que para cada punto de la curva corresponde a uno de los años del rango de predicción.

Tras este proceso de aproximación, se obtienen un total de 204 curvas, 3 (Materia Gris, Materia Blanca y líquido Cefalorraquídeo) para cada una de las 68 áreas.

Por ejemplo, en la figura 5.1, se ven las curvas resultantes para las 3 materias del área número 5, de nombre `Ant_temp_med_L`. El rango de predicción se sitúa entre las edades [40, 90]

### 5.3. Comparación de curvas

Una vez se disponen de todas las curvas para un grupo de pacientes, pueden calcularse las del segundo grupo, con el objetivo de compararlas. En este caso, se obtienen las curvas para dos grupos de pacientes:

- Pacientes **sin patologías**: aquellos que han dado negativo en la prueba de punción lumbar.
- Pacientes **con beta-amiloide**: han dado positivo en el biomarcador beta-amiloide, y negativo en tau y fosfo-tau.

Por lo tanto, hasta el momento se disponen de un total de 408 curvas, 204 para cada uno de los dos grupos de pacientes.

En la imagen 5.2 se pueden observar de manera gráfica las diferencias entre las curvas del área 5 materia WM. La cantidad de materia blanca hasta los 70 años es inferior en el grupo beta-amiloide, aunque esta tendencia se invierte a partir de esa misma edad, donde el grupo sin patologías es ahora el que tiene una menor cantidad de materia blanca.

Sin embargo, es necesaria algún tipo de métrica para evaluar las diferencias entre dos curvas. En este caso, como todas las curvas tienen la misma longitud (se encuentran en el

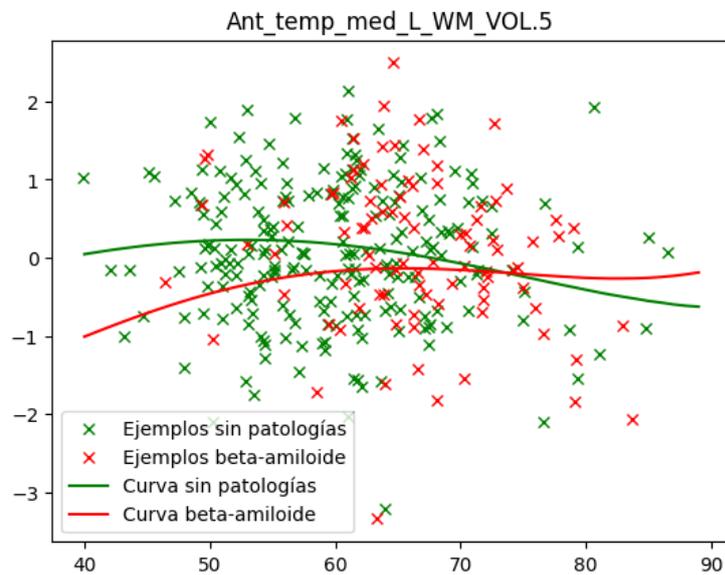


Figura 5.2: Curvas de Ant\_temp\_med\_L\_WM\_VOL.5

Cálculo de diferencia entre dos curvas

---

```

1 input: curva1, curva2
2 output: diferencia
3 diferencia = 0
4 for 1 to tamaño_curvas
5     diferencia += abs(curva1[i]-curva2[i])
6 rof
7 return diferencia

```

---

Algoritmo 5.1: Calculo de diferencia entre dos curvas

mismo rango), se puede utilizar el acumulado de distancias en cada punto. La distancia se calcula como el valor absoluto de la resta de dos puntos de la curva (ver algoritmo 5.1).

Por ejemplo, para el área anterior (Ant\_temp\_med\_L\_WM\_VOL.5), la diferencia obtenida es de 15,83 unidades.

De esta manera, cuanto mayor sea el valor de la diferencia en un área, mayor será la diferencia en la evolución entre los dos grupos de pacientes.

Posición	Nombre	Diferencia
1	Cerebelum_R_GM_VOL.18	39,429
2	Amygd_L_GM_VOL.3	36,982
3	Cerebelum_L_GM_VOL.17	36,011
4	Amygd_R_GM_VOL.4	35,841
5	Parahipp_L_GM_VOL.9	33,783
6	Mid_Inf_Temp_L_GM_VOL.13	31,038
7	Lat_Temp_Ventr_L_GM_VOL.47	29,708
8	C_Callosum_R_GM_VOL.46	28,894
9	Mid_Front_R_GM_VOL.30	28,034
10	Accumb_R_GM_VOL.38	27,912
11	Post_Cingul_L_GM_VOL.27	27,333
12	Ant_Orbit_Gyr_L_GM_VOL.55	27,319
13	Post_Cingul_R_GM_VOL.28	27,213
14	GM_VOL_hipp_L	27,168
15	Third_Ventr_R_GM_VOL.50	26,953

**Tabla 5.1:** Ordenación de las 15 áreas cerebrales con mayor diferencia de GM

Posición	Nombre	Diferencia
1	Caudate_R_WM_VOL.36	32,219
2	Parahipp_L_WM_VOL.9	28,841
3	Third_Ventr_R_WM_VOL.50	27,617
4	Third_Ventr_L_WM_VOL.49	27,447
5	Pallidum_L_WM_VOL.43	27,301
6	Brainstem_L_WM_VOL.19	25,455
7	Pre_Centr_Gyr_L_WM_VOL.51	25,372
8	Lingual_Gyr_L_WM_VOL.65	24,735
9	Post_Cingul_L_WM_VOL.27	24,009
10	Lingual_Gyr_R_WM_VOL.66	23,055
11	Brainstem_R_WM_VOL.20	22,352
12	Straight_Gyr_L_WM_VOL.53	21,841
13	Cuneus_R_WM_VOL.68	21,686
14	WM_VOL_hipp_L	21,643
15	Caudate_L_WM_VOL.35	20,797

**Tabla 5.2:** Ordenación de las 15 áreas cerebrales con mayor diferencia de WM

#### 5.4. Ordenación de áreas en función de su diferencia

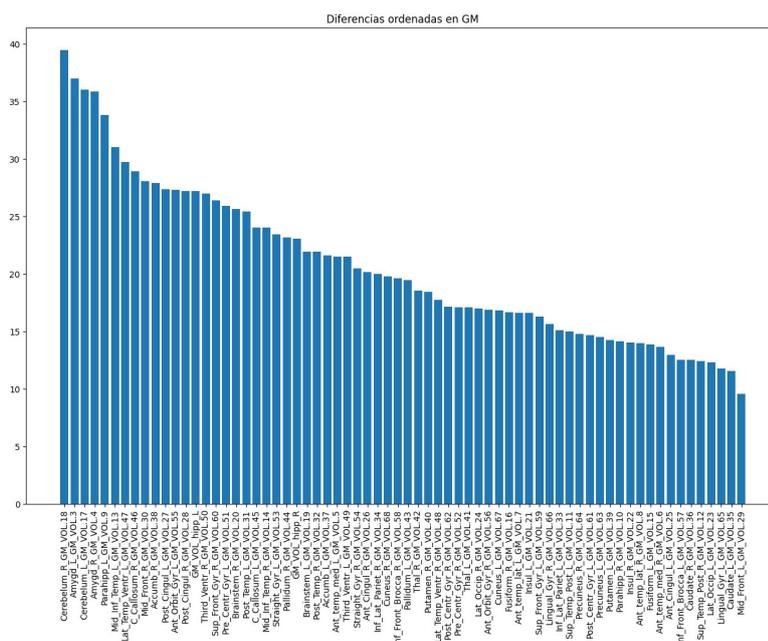
Ahora que se dispone de una métrica que permita observar las diferencias entre las curvas de los dos grupos de pacientes, podemos ordenar las áreas en función de este valor, y así encontrar las áreas con una mayor diferencia de evolución, las cuales podrían aportar mayor información para entender la enfermedad.

Los valores de diferencia de cada área ordenados se encuentran en la tablas 5.1 (para GM), 5.2 (para WM) y 5.3 (para CSF). Puede verse como áreas con mayor diferencia difieren en función del tipo de materia, las tres ordenaciones son distintas. Además, los valores de diferencias son mucho mayores cuando hablamos del líquido cefalorraquídeo, mientras que en las materias gris y blanca las diferencias son por lo general menores, especialmente en ésta última. Esto es un resultado de esperar, ya que el líquido cefalorraquídeo ocupa el volumen que las materias blanca y gris dejan al reducirse con la edad.

#### 5.4. Ordenación de áreas en función de su diferencia

Posición	Nombre	Diferencia
1	Straight_Gyr_R_CSF_VOL.54	43,745
2	CSF_VOL_hipp_R	38,091
3	Ant_Orbit_Gyr_L_CSF_VOL.55	34,332
4	Ant_Orbit_Gyr_R_CSF_VOL.56	33,958
5	Ant_Cingul_L_CSF_VOL.25	29,868
6	Lat_Temp_Ventr_R_CSF_VOL.48	28,527
7	Third_Ventr_R_CSF_VOL.50	28,278
8	Cuneus_L_CSF_VOL.67	27,348
9	Third_Ventr_L_CSF_VOL.49	27,11
10	Caudate_R_CSF_VOL.36	26,938
11	Post_Centr_Gyr_R_CSF_VOL.62	26,842
12	Thal_L_CSF_VOL.41	26,412
13	Ant_Cingul_R_CSF_VOL.26	26,269
14	Pallidum_L_CSF_VOL.43	25,786
15	Inf_Front_Brocca_L_CSF_VOL.57	25,399

**Tabla 5.3:** Ordenación de las 15 áreas cerebrales con mayor diferencia de CSF



**Figura 5.3:** Áreas ordenadas por diferencia de GM

Además, en las figuras 5.3 (para GM), 5.4 (para WM), 5.5 (para CSF) puede observarse manera gráfica la ordenación con todas las áreas y los valores acumulados para cada una de ellas. Cabe destacar que no se observa un salto notable que ayude a seleccionar el conjunto de áreas significativamente diferentes en ninguna de las materias, el descenso es regular en todas las gráficas.

En la figura 5.6 pueden verse gráficamente las 5 áreas con mayor diferencia de cada materia. Las mayores diferencias están claramente en los extremos, donde se dispone de una menor cantidad de datos.

## 5. PROCESO DE IMPLEMENTACIÓN

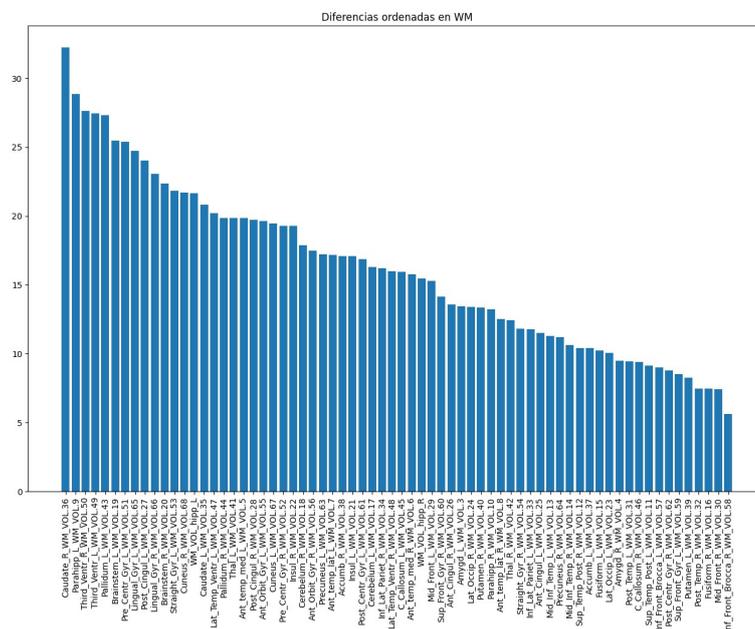


Figura 5.4: Áreas ordenadas por diferencia de WM

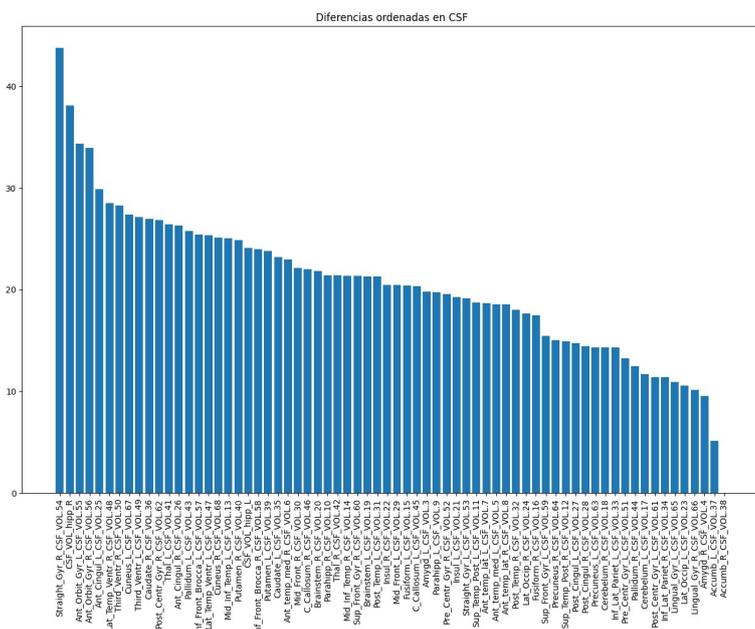


Figura 5.5: Áreas ordenadas por diferencia de CSF

En los resultados se observa que no todas las materias tienen la misma ordenación de áreas, lo que indica que las 3 materias evolucionan de distinta manera en pacientes sin patologías y en pacientes con el biomarcador beta-amiloide.

## 5.4. Ordenación de áreas en función de su diferencia

Top 5 áreas con con mayor diferencia entre sanos (verde) y beta-amiloide (rojo)

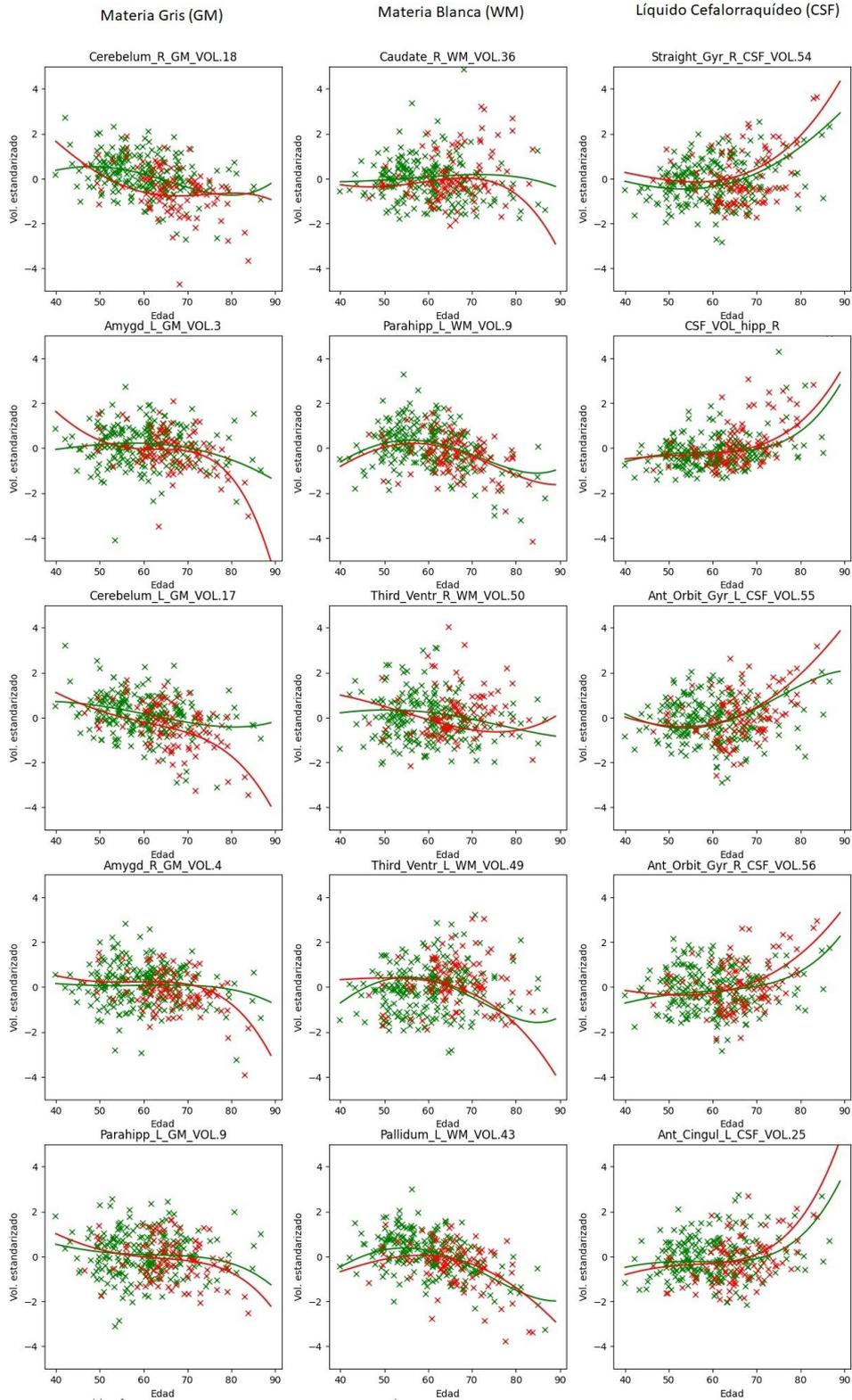


Figura 5.6: Las 5 áreas con mayor diferencia de cada materia

## 5.5. Cuestiones tras la primera aproximación

Después de observar los resultados, se presentan algunas cuestiones a tener en cuenta, de cara a futuras pruebas.

En los rangos extremos de edad, el número de ejemplos que tenemos es escaso (ver figura 4.4). Esto puede provocar que las aproximaciones de las curvas en los rangos [40, 45], [45, 50], y [85, 90] se sobreajusten demasiado a los datos, y den unos valores de diferencia menos fiables. Además, como uno de los objetivos es detectar el Alzheimer en casos preclínicos, los especialistas de CITA-Alzheimer consideran adecuado prescindir de estos rangos, ya que por lo general o son edades muy tempranas para desarrollar la enfermedad, o edades muy avanzadas para ser un caso preclínico.

Por otra parte, no se ha explorado todo el potencial del método Ridge. El parámetro  $\lambda$  permite aumentar o reducir la sensibilidad de la curva a los datos. Un mayor valor de  $\lambda$  aporta una menor sensibilidad.

Por lo tanto, para abordar estas cuestiones, se han llevado a cabo nuevas aproximaciones, modificando ciertos elementos:

- Generar nuevas curvas en un rango de edad más reducido, concretamente en [50, 85].
- Ampliar el valor de  $\lambda$  a 1, 5 y 10, reduciendo de esta manera la sensibilidad a los datos.

Gracias a este análisis con los nuevos grupos de curvas, se podrá analizar cómo afecta la modificación de los parámetros de rango y  $\lambda$  a la ordenación de las áreas según su diferencia en evolución para distintos pacientes.

Por lo tanto, se generarán los siguientes grupos de curvas:

- **Grupo 1:** predicción en el rango [40, 90] y  $\lambda = 0,5$ .
- **Grupo 2:** predicción en el rango [50, 85] y  $\lambda = 0,5$
- **Grupo 3:** predicción en el rango [50, 85] y  $\lambda = 1$ .
- **Grupo 4:** predicción en el rango [50, 85] y  $\lambda = 5$ .
- **Grupo 5:** predicción en el rango [50, 85] y  $\lambda = 10$ .

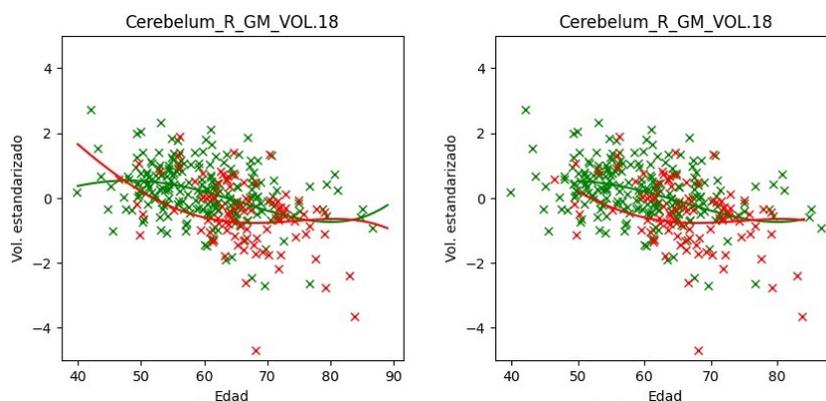


Figura 5.7: Comparación visual de las diferencias de rango en las curvas del área 18

## 5.6. Resultados de las nuevas aproximaciones

Tras llevar a cabo las nuevas aproximaciones modificando los parámetros del rango y  $\lambda$ , se observan bastantes cambios en el orden de las áreas cerebrales con mayor diferencia. En las tablas 5.4 (para GM), 5.5 (para WM) y 5.6 (para CSF), se muestran los órdenes de todos los grupos de curvas generados para las 3 materias. Se muestran solo las 15 áreas con mayor diferencia. Se ha prescindido mostrar el valor de la diferencia, ya que lo verdaderamente relevante es evaluar si existen saltos muy significativos, y en este caso nos lo hay. Las gráficas de todas las áreas ordenadas se encuentran en el apéndice. Para facilitar la lectura, se han sombreado de color verde las áreas que coinciden en la misma posición en 4 o más grupos, y en azul las áreas que coinciden en 4 o más grupos, pero no en la misma posición.

Puede verse como la reducción del rango de predicción afecta significativamente al orden de las curvas. Esto se debe a que los rangos eliminados eran aquellos dónde las diferencias eran mayores, porque había una cantidad reducida de datos que influían demasiado en la curva. En la figura 5.7 se observa el caso del área Cerebellum\_R\_GM\_VOL.18. En la gráfica izquierda, se muestran las curvas generadas para el rango de predicción [40, 90], mientras que la derecha muestra las curvas para el rango [50, 85]. Claramente, en los rangos eliminados las diferencias eran mayores, lo cuál afecta al cambio de orden.

Por otra parte, la modificación del valor de  $\lambda$  apenas ha provocado cambios. Si que es verdad que existen ligeras variaciones, pero por lo general las 5 primeras áreas se mantienen idénticas en los grupos 2, 3, 4 y 5. Esto nos indica que incluso con un factor de regularización  $\lambda$  bajo, la curva es capaz de generalizar bien, sin sobre-ajustarse demasiado al conjunto de datos.

Por lo tanto, tras esta segunda aproximación podemos concluir que, por una parte el rango sobre el que se realiza la predicción es muy determinante a la hora de obtener la ordenación, dado que el número de puntos existentes difiere mucho dependiendo del rango de edades utilizado. Por otro lado, el parámetro  $\lambda$  de la regularización Ridge no afecta demasiado. En futuras pruebas utilizaremos el rango de predicción [50, 85], considerado de mayor interés por los médicos de CITA-Alzheimer, y  $\lambda = 1$ . Son los valores del Grupo 3 de curvas, el que ha obtenido un mayor número de coincidencias con respecto al resto de grupos.

## 5. PROCESO DE IMPLEMENTACIÓN

Posición	Grupo1	Grupo2	Grupo3	Grupo4	Grupo5
1	VOL.18	VOL.18	VOL.18	VOL.18	VOL.18
2	VOL.3	VOL.17	VOL.17	VOL.17	VOL.17
3	VOL.17	VOL.46	VOL.46	VOL.46	VOL.46
4	VOL.4	VOL.55	VOL.55	VOL.55	VOL.55
5	VOL.9	VOL.60	VOL.60	VOL.45	VOL.45
6	VOL.13	VOL.45	VOL.45	VOL.60	VOL.50
7	VOL.47	VOL.50	VOL.50	VOL.50	VOL.60
8	VOL.46	VOL.9	VOL.9	VOL.20	VOL.20
9	VOL.30	VOL.20	VOL.20	VOL.9	VOL.19
10	VOL.38	VOL.53	VOL.53	VOL.19	VOL.9
11	VOL.27	VOL.42	VOL.42	VOL.53	VOL.53
12	VOL.55	VOL.13	VOL.13	VOL.54	VOL.54
13	VOL.28	VOL.54	VOL.19	VOL.13	VOL.13
14	hipp_L	VOL.3	VOL.54	VOL.42	VOL.42
15	VOL.50	VOL.19	VOL.31	VOL.44	VOL.44

**Tabla 5.4:** Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de GM

Posición	Grupo1	Grupo2	Grupo3	Grupo4	Grupo5
1	VOL.36	VOL.36	VOL.36	VOL.36	VOL.36
2	VOL.9	VOL.35	VOL.35	VOL.35	VOL.35
3	VOL.50	VOL.66	VOL.66	VOL.66	VOL.66
4	VOL.49	VOL.68	VOL.68	VOL.19	VOL.19
5	VOL.43	VOL.50	VOL.50	VOL.50	VOL.50
6	VOL.19	VOL.19	VOL.19	VOL.51	VOL.51
7	VOL.51	VOL.65	VOL.65	VOL.68	VOL.17
8	VOL.65	VOL.43	VOL.51	VOL.17	VOL.68
9	VOL.27	VOL.20	VOL.20	VOL.20	VOL.20
10	VOL.66	VOL.51	VOL.43	VOL.43	VOL.43
11	VOL.20	VOL.17	VOL.17	VOL.61	VOL.61
12	VOL.53	VOL.67	VOL.67	VOL.65	VOL.65
13	VOL.68	VOL.9	VOL.61	VOL.49	VOL.49
14	hipp_L	VOL.61	VOL.9	VOL.67	VOL.67
15	VOL.35	VOL.49	VOL.49	VOL.18	VOL.18

**Tabla 5.5:** Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de WM

Posición	Grupo1	Grupo2	Grupo3	Grupo4	Grupo5
1	VOL.54	VOL.48	VOL.48	VOL.48	VOL.48
2	hipp_R	VOL.47	VOL.47	VOL.47	VOL.47
3	VOL.55	VOL.55	VOL.36	VOL.36	VOL.36
4	VOL.56	VOL.36	VOL.55	VOL.55	VOL.54
5	VOL.25	hipp_R	hipp_R	VOL.54	VOL.55
6	VOL.48	VOL.56	VOL.54	VOL.56	VOL.35
7	VOL.50	VOL.54	VOL.56	VOL.35	VOL.56
8	VOL.67	VOL.3	VOL.3	hipp_R	hipp_R
9	VOL.49	hipp_L	VOL.35	VOL.3	VOL.3
10	VOL.36	VOL.35	hipp_L	VOL.50	VOL.46
11	VOL.62	VOL.13	VOL.13	VOL.46	VOL.50
12	VOL.41	VOL.50	VOL.50	hipp_L	VOL.13
13	VOL.26	VOL.25	VOL.25	VOL.13	hipp_L
14	VOL.43	VOL.46	VOL.46	VOL.25	VOL.25
15	VOL.57	VOL.45	VOL.45	VOL.45	VOL.45

**Tabla 5.6:** Ordenación por grupos de las 15 áreas cerebrales por con mayor diferencia de CSF

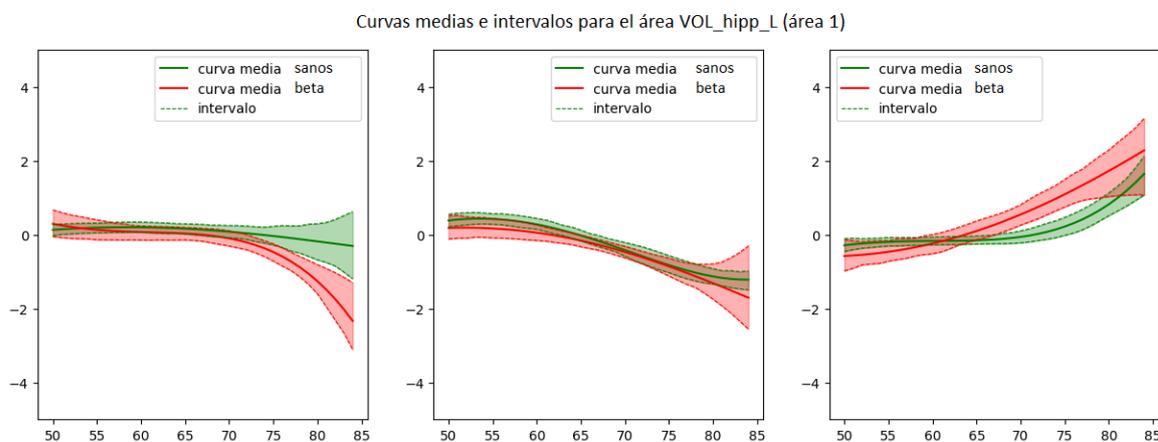


Figura 5.8: Ejemplo de curvas con sus intervalos para el área 1

## 5.7. Curvas a partir de muestras Bootstrap

Una vez concluido que el orden se ve muy afectado por el rango, y poco por el parámetro  $\lambda$ , se debe comprobar cómo afecta la muestra concreta a la generación de las curvas, y por consiguiente, a los ordenes de las áreas.

Para llegar a conclusiones más fiables, se ha generado un nuevo grupo de curvas mediante la técnica de Bootstrapping, de tal manera que cada curva resulte de la media de las curvas de cada una de las muestras.

El grupo 6 de curvas está generado a partir de 100 muestras bootstrap con un intervalo de confianza del 90 %. Estos intervalos de confianza permiten llevar a cabo nuevas aproximaciones con las que comparar resultados.

Se han llevado a cabo dos nuevas ordenaciones sobre las curvas medias, aprovechando los intervalos de confianza. La primera se ha realizado como las anteriores. En la segunda, se calculará la diferencia entre puntos de la curva sólo en las regiones donde no haya solapamientos de intervalo, es decir, cuando el intervalo inferior de la curva superior tenga un valor mayor que el intervalo superior de la curva inferior. Esta condición nos permite asegurar en qué tramos las curvas son realmente distintas, sin tener en cuenta tramos en los que la diferencia puede ser menos relevante.

En la figura 5.8 se muestra cómo quedarían las curvas medias y los intervalos de confianza para las materias GM, WM y CSF del área número 1, VOL\_hipp\_L. De ellas se puede concluir en qué tramos de las curvas se tomarán diferencias de distancias. Las curvas verdes representan el grupo sin patologías, mientras que las rojas el grupo con beta-amiloide. Para la primera gráfica (materia GM), la condición de *no solapamiento* se produce en el rango de edad [75, 85] aproximadamente. Por otra parte, en la segunda gráfica (materia WM) nunca se cumple esta condición, es decir, las curvas se encuentran solapadas a lo largo de todo el rango de edad, por lo que la diferencia en este caso sería de 0. Finalmente, para la tercera gráfica (materia CSF), se calcularía la diferencia en el rango [65, 80] aproximadamente, el único tramo donde no se solapan.

## 5. PROCESO DE IMPLEMENTACIÓN

Posición	Muestra total	Muestra bootstrapping			
	Grupo 3 Área	Diferencia total		Sin solapamiento	
	Área	Área	Diferencia	Área	Diferencia
1	VOL.18	VOL.18	22,112	VOL.50	13,164
2	VOL.17	VOL.17	19,684	VOL.20	13,040
3	VOL.46	VOL.46	19,176	VOL.9	12,300
4	VOL.55	VOL.45	17,977	VOL.55	12,036
5	VOL.60	VOL.55	17,919	VOL.19	11,519
6	VOL.45	VOL.50	17,588	VOL.18	11,491
7	VOL.50	VOL.60	17,145	VOL.42	11,131
8	VOL.9	VOL.9	16,350	VOL.3	10,440
9	VOL.20	VOL.20	16,182	VOL.37	10,092
10	VOL.53	VOL.53	15,943	VOL.45	8,330
11	VOL.42	VOL.19	15,479	VOL.60	7,703
12	VOL.13	VOL.42	15,244	VOL.27	7,151
13	VOL.19	VOL.54	14,831	VOL.46	5,940
14	VOL.54	VOL.13	14,817	VOL.17	5,604
15	VOL.31	VOL.3	14,363	VOL.41	5,391

**Tabla 5.7:** Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia GM

Teniendo en cuenta esta nueva forma de evaluación de diferencias, se ha llevado una última ordenación, cuyos resultados se muestran en las tablas 5.7 para materia gris (GM), 5.8 para materia blanca (WM) y 5.9 para líquido ceforraquídeo (CSF). En ella se visualizan la ordenación descendente de las 15 áreas con mayor diferencia para el grupo 6 de curvas (curvas medias usando bootstrapping), primero calculando la diferencia para todo el rango de edad, y después calculando la diferencia sólo en los rangos donde no existe solapamiento de curvas. Adicionalmente, se ha agregado una columna con la ordenación del grupo 3 de curvas, las generadas en el rango [50, 85] y con  $\lambda = 1$ , el grupo de curvas considerado como más representativo durante la segunda aproximación.

Adicionalmente, en las figuras 5.9 y 5.10 (para WM), 5.11 y 5.12 (para WM) y 5.13 5.14 (para CSF) se muestran los gráficos de todas las áreas ordenadas por diferencia descendente para las curvas bootstrap. Cada par de figuras muestra el orden calculado con la distancia total y con la distancia en los rangos sin solapamiento. Como puede verse, en el orden de distancias a partir de los rangos sin solapamiento hay áreas cuya diferencia es 0. Esto se debe a que las curvas de esas áreas se encuentran solapadas durante todo el rango de edad, y que por lo tanto pueden considerarse como áreas en las que la diferencia es tan pequeña que apenas es significativa.

Por último, en la figura 5.15 se muestran también las 5 curvas con mayor diferencia en los rangos sin solapamiento, divididos por materias. Se puede observar como esa diferencia se va reduciendo poco a poco desde el primer área hasta la quinta.

5.7. Curvas a partir de muestras Bootstrap

Posición	Muestra total	Muestra bootstrapping			
	Grupo 3	Diferencia total		Sin solapamiento	
	Área	Área	Diferencia	Área	Diferencia
1	VOL.36	VOL.36	21,188	VOL.36	13,017
2	VOL.35	VOL.35	16,136	VOL.66	11,731
3	VOL.66	VOL.66	14,386	VOL.50	9,171
4	VOL.68	VOL.68	13,973	VOL.68	9,024
5	VOL.50	VOL.50	13,313	VOL.35	8,429
6	VOL.19	VOL.65	12,761	VOL.65	7,832
7	VOL.65	VOL.19	12,032	VOL.17	7,466
8	VOL.51	VOL.17	11,938	VOL.67	7,450
9	VOL.20	VOL.67	11,878	VOL.18	5,799
10	VOL.43	VOL.43	11,615	VOL.34	4,441
11	VOL.17	VOL.20	11,312	VOL.9	3,849
12	VOL.67	VOL.51	11,179	VOL.43	3,609
13	VOL.61	VOL.49	10,988	VOL.53	2,063
14	VOL.9	VOL.9	10,543	VOL.7	1,652
15	VOL.49	VOL.7	10,490	hipp_L	1,299

**Tabla 5.8:** Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia WM

Posición	Muestra total	Muestra bootstrapping			
	Grupo 3	Diferencia total		Sin solapamiento	
	Área	Área	Diferencia	Área	Diferencia
1	VOL.48	VOL.48	18,910	VOL.55	14,675
2	VOL.47	VOL.47	17,641	VOL.56	12,758
3	VOL.36	VOL.36	17,397	VOL.52	9,803
4	VOL.55	VOL.55	17,242	VOL.60	9,413
5	hipp_R	VOL.3	17,038	VOL.48	9,401
6	VOL.54	hipp_R	16,731	VOL.3	9,193
7	VOL.56	hipp_L	16,300	VOL.13	9,160
8	VOL.3	VOL.35	16,107	VOL.47	8,896
9	VOL.35	VOL.56	16,005	VOL.36	8,440
10	hipp_L	VOL.54	15,895	VOL.46	8,355
11	VOL.13	VOL.46	15,338	hipp_L	8,095
12	VOL.50	VOL.52	15,299	VOL.59	8,059
13	VOL.25	VOL.13	15,235	VOL.14	7,371
14	VOL.46	VOL.50	15,227	VOL.45	7,313
15	VOL.45	VOL.25	15,192	VOL.25	7,059

**Tabla 5.9:** Ordenaciones para las curvas generadas con muestras bootstrap, con diferencia total y diferencia sin solapamiento. Materia CSF

## 5. PROCESO DE IMPLEMENTACIÓN

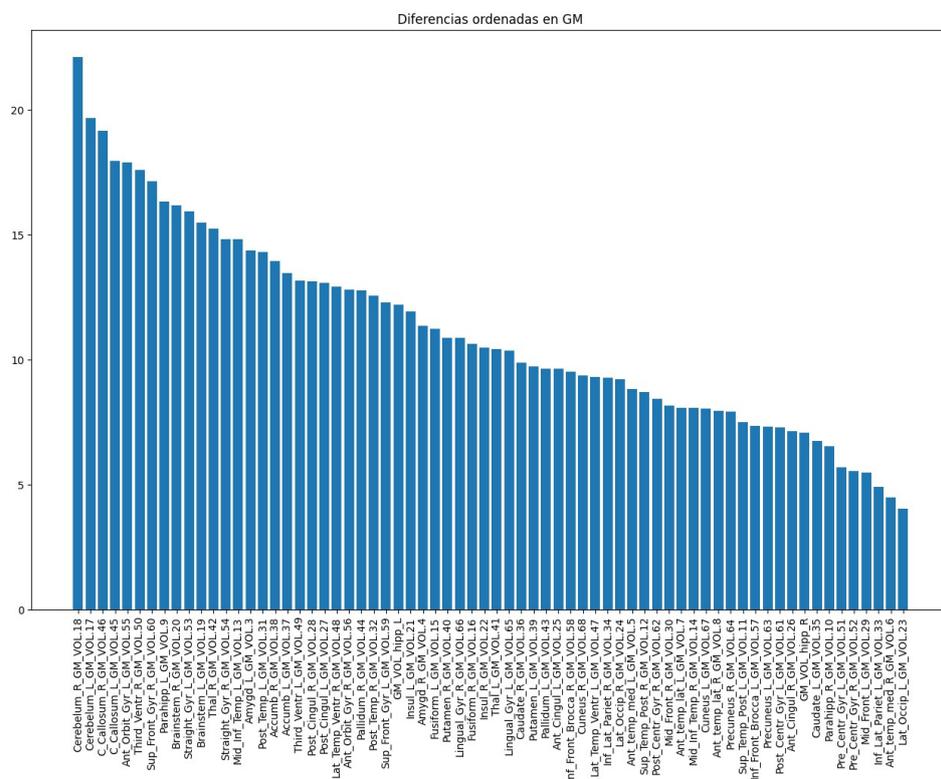


Figura 5.9: Áreas ordenadas por diferencia total en curvas bootstrap. Materia GM

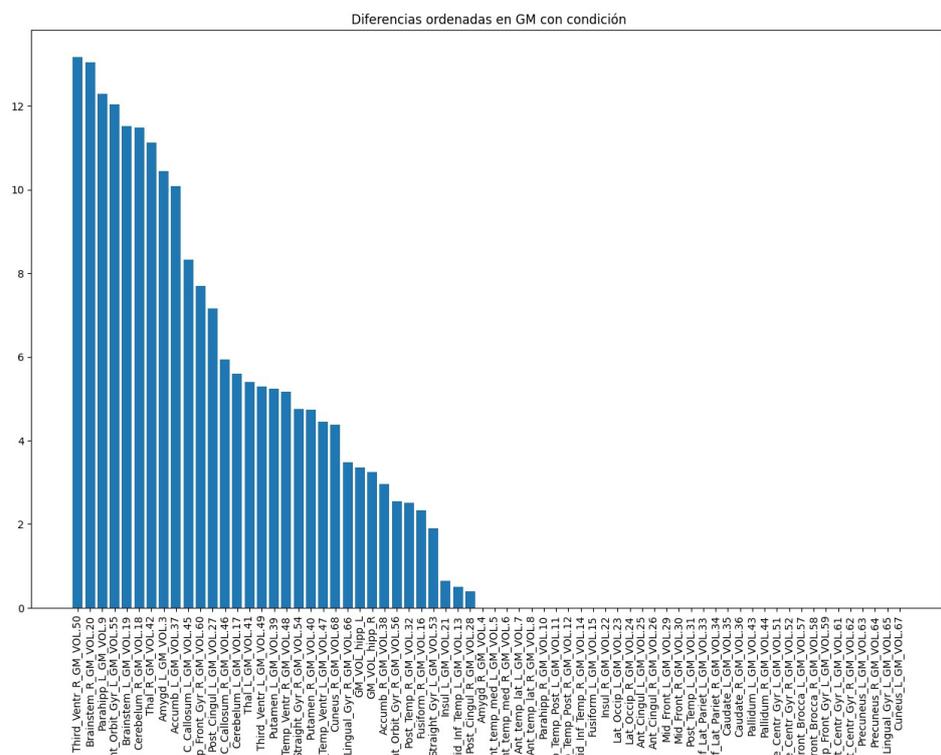


Figura 5.10: Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia GM

## 5.7. Curvas a partir de muestras Bootstrap

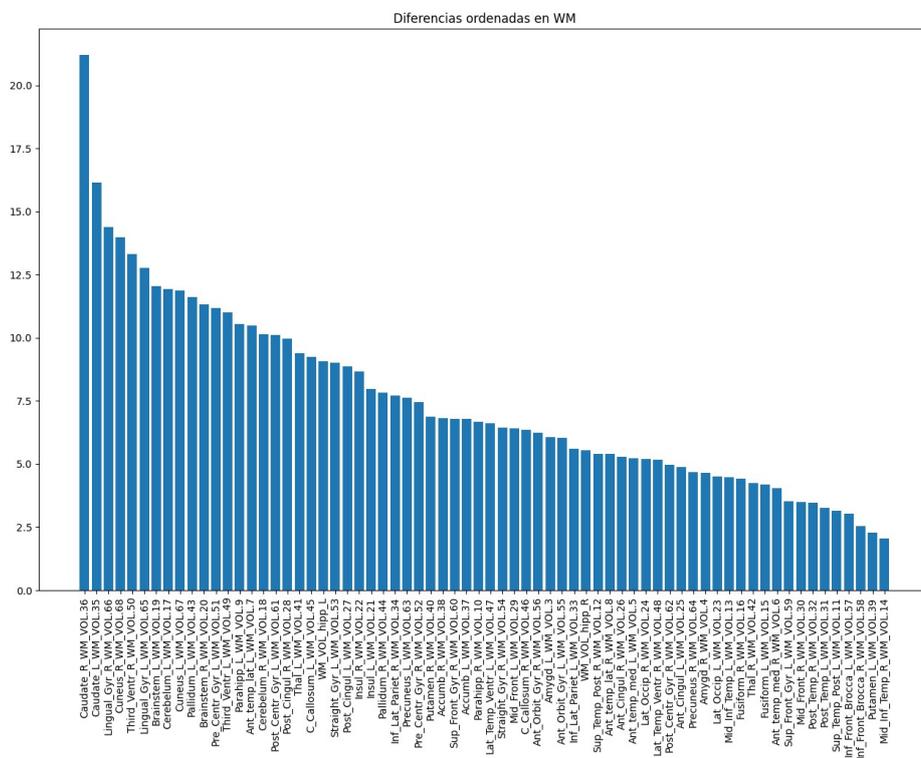


Figura 5.11: Áreas ordenadas por diferencia total en curvas bootstrap. Materia WM

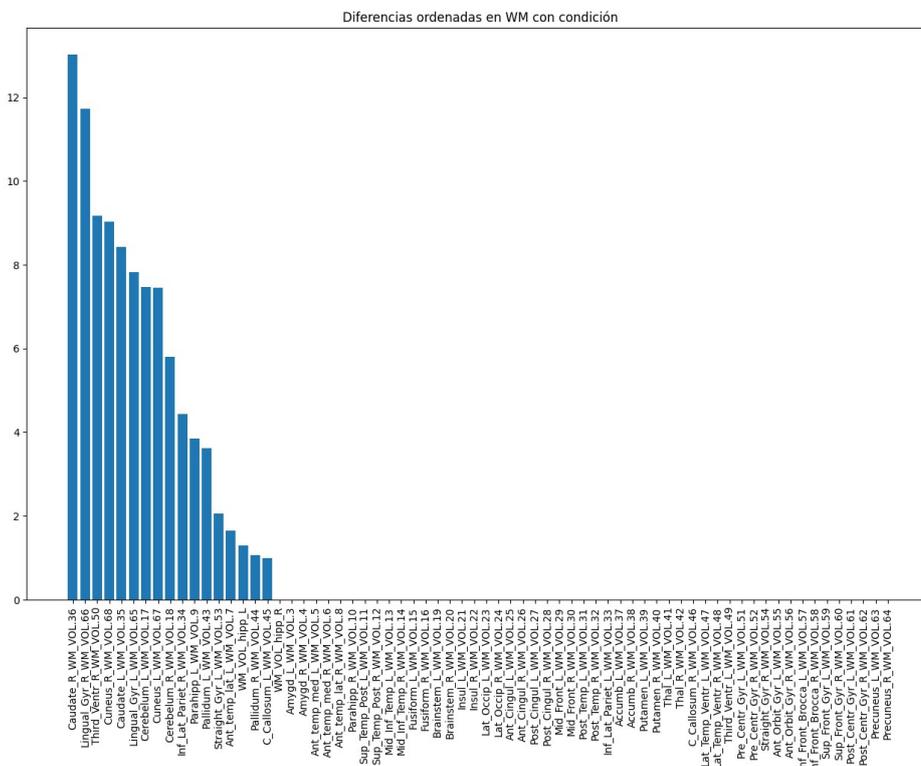


Figura 5.12: Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia WM

5. PROCESO DE IMPLEMENTACIÓN

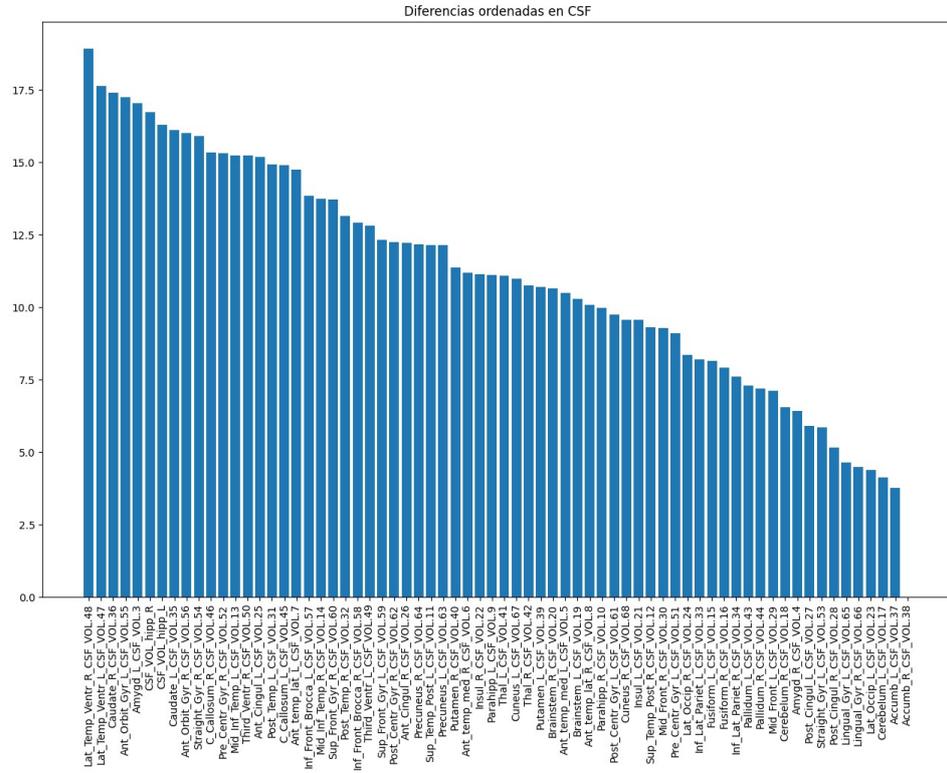


Figura 5.13: Áreas ordenadas por diferencia total en curvas bootstrap. Materia CSF

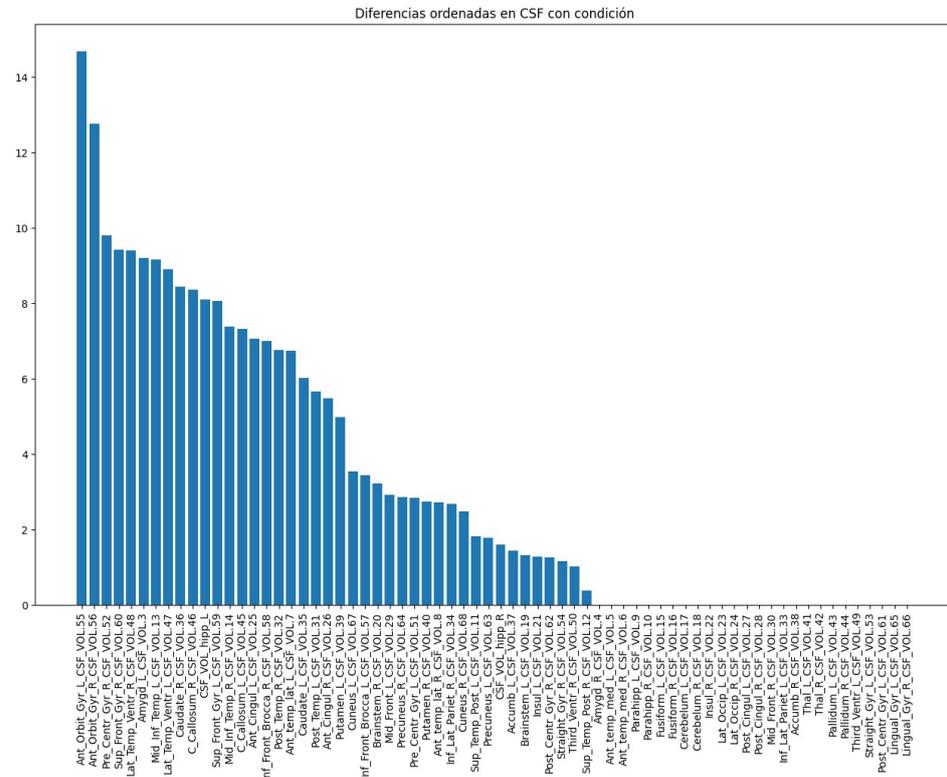


Figura 5.14: Áreas ordenadas por diferencia sin solapamiento en curvas bootstrap. Materia CSF

## 5.7. Curvas a partir de muestras Bootstrap

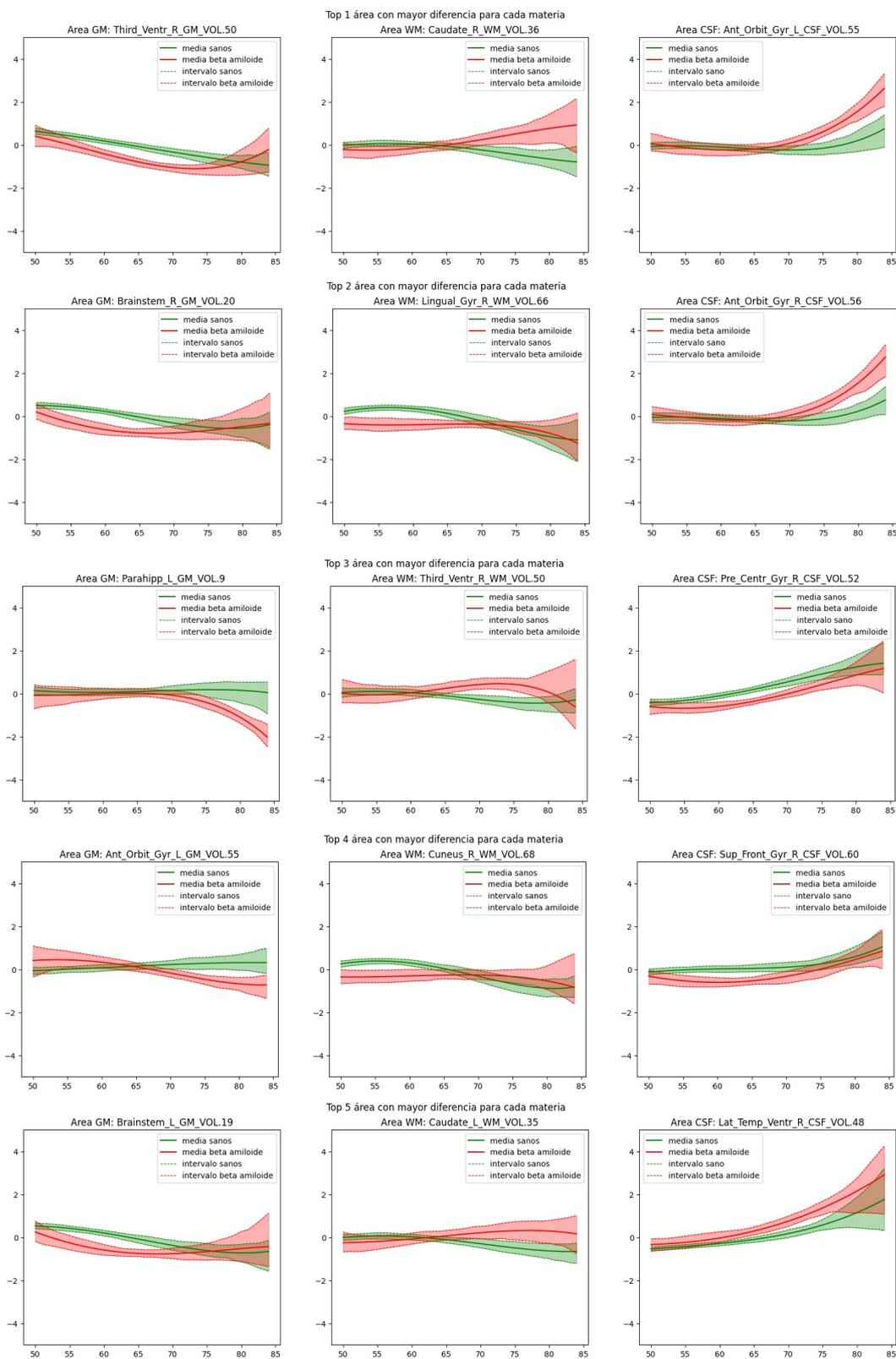


Figura 5.15: Las 5 áreas con mayor diferencia en los tramos sin solapamiento.

## 5.8. Resultados de la aproximación con Bootstrapping

Como se ha podido ver en las tablas, los órdenes de las curvas según las diferencias entre el grupo de curvas 3 y el grupo de curvas bootstrap (curvas medias) difiere ligeramente. Sin embargo, coinciden las cuatro o cinco áreas con mayor diferencia de ambos grupos, para todas las materias

También hay que destacar que aunque no coincidan en posición, la mayoría de las áreas del grupo 3 se encuentran en el grupo bootstrap, es decir, las áreas con mayor diferencia tienden a ser las mismas en ambas aproximaciones, pero se encuentran en un orden ligeramente distinto. Esto quiere decir que la muestra total no afecta demasiado en la generación de las curvas, o dicho de otra forma, las curvas, aunque sean aproximaciones certeras de los casos, no se sobreajustan demasiado a la muestra concreta. Sin embargo, es más conveniente dar por correctas las curvas bootstrap, puesto que este método de obtención aporta una mayor fiabilidad a los resultados.

Por otro lado, el orden de áreas teniendo en cuenta la diferencia sin solapamiento de intervalos es bastante más distinto que en los otros dos grupos anteriores. Si bien es verdad que algunas de las áreas son las mismas, el orden en el que se encuentran es muy diferente.

Si se comparan las diferencias sin solapamiento entre las 3 materias, pueden observarse ciertas tendencias. Para empezar, en WM se produce un descenso de las diferencias mucho más rápido que en las otras dos materias. Si se observa la gráfica 5.12, tan solo 17 de las 68 áreas cumplen en algún momento la condición de no solapamiento, frente a las de 33 GM y a las 42 de CSF. Además, los valores correspondientes a las 15 áreas con mayor diferencia de WM se encuentran en un rango de  $[1, 14]$ , mientras que los rangos de valores de GM y CSF son  $[5, 14]$  y  $[7, 15]$  respectivamente, lo que puede dar a entender que la cantidad de materia blanca presente en un paciente sea menos relevante en un hipotético diagnóstico.

## Clasificación de pacientes usando curvas

---

```

1 puntuacion_sanos = 0
2 puntuacion_beta = 0
3 forcada paciente
4   for area in num_areas
5     valor_cercania = calcular_cercania(paciente, curvas)
6     if paciente más cercano a curva sanos then
7       puntuacion_sanos += valor_cercania
8     else
9       puntuacion_beta += valor_cercania
10  rof
11 rof
12 if puntuacion_sanos > puntuacion_beta then
13   clasificar grupo de pacientes sanos
14 else if puntuacion_sanos < puntuacion_beta then
15   clasificar grupo de pacientes beta-amiloide
16 else
17   No clasificar
18 fi

```

---

Algoritmo 5.2: Clasificación de pacientes usando curvas

## 5.9. Clasificación de pacientes usando las curvas

Como objetivo final, se ha decidido llevar a cabo una pequeña clasificación aprovechando las curvas elaboradas mediante la técnica de bootstrapping, junto con sus intervalos de confianza.

La idea principal es utilizar las aproximaciones de las curvas y las distancias de cada paciente a éstas para construir un clasificador sencillo.

El procedimiento será el siguiente: calcular para cada paciente a clasificar dos puntuaciones: una que indique el nivel de cercanía a las curvas del grupo sin patologías y otra para el grupo beta-amiloide, para posteriormente clasificarlo en el grupo cuya puntuación sea mayor. El cálculo de las puntuaciones se llevará a cabo para cada una de las 68 áreas cerebrales, y 3 materias, se irá acumulando (ver algoritmo 5.2).

Antes de proceder con la clasificación, hay que redondear la edad del paciente. En la base de datos se dispone de la edad en formato continuo, pero la curva se define en valores enteros de edad. Por lo tanto, si un paciente tiene 60,25 años (la edad en la base de datos), para su clasificación se redondeará la edad al entero más cercano, es decir 60 años.

Hay que aclarar que las curvas aproximan la evolución del volumen cerebral de las distintas áreas, por lo que utilizarlas para clasificar pacientes midiendo la distancia que existe entre éstos y las curvas es una aproximación muy sencilla. Sin embargo esta clasificación se lleva a cabo para ver como se comportan las curvas en una hipotética clasificación de pacientes, la cuál se podría elaborar y mejorar más en futuros trabajos.

Para llevar a cabo la clasificación se proponen dos aproximaciones. En ambas se utilizará el grupo de curvas medias generado a partir de las muestras bootstrap.

### 5.9.1. Primera aproximación de clasificación

Con las curvas medias de todas las áreas y la edad redondeada del paciente, se calculará un valor de cercanía mediante la siguiente fórmula:

$$cercania = abs(X_{a,m} - \frac{S_{a,m,e} + B_{a,m,e}}{2}) \quad (5.1)$$

Dónde:

- $X_{a,m}$  es el volumen normalizado y estandarizado del área  $a$  materia  $m$  para el paciente  $X$ .
- $S_{a,m,e}$  es el valor de la curva del grupo de pacientes sanos para el área  $a$ , materia  $m$  y edad  $e$ .
- $B_{a,m,e}$  es el valor de la curva del grupo de pacientes beta-amiloide para el área  $a$ , materia  $m$  y edad  $e$
- el subíndice  $e$  es el valor de edad del paciente tras el redondeo.

Dicho valor será acumulado en la puntuación del grupo más cercano. Es decir, si el paciente se encuentra más cerca de la curva del grupo sano, se acumulará el valor de cercanía a la puntuación del grupo sano, y lo mismo en caso de estar más cerca del grupo beta-amiloide.

Tras haber calculado y acumulado el valor de cercanía del paciente para todas las áreas, se deberá clasificar. Se ha utilizado la siguiente fórmula:

$$prediccion = \log(puntuacionBeta/puntuacionSano) \quad (5.2)$$

En función del valor de predicción se llevará a cabo la clasificación del paciente:

- Si  $prediccion > 0$ , entonces clasificar grupo Beta-amiloide (1).
- Si  $prediccion < 0$ , entonces clasificar grupo Sanos (0).

### 5.9.2. Segunda aproximación de clasificación

La segunda aproximación es algo más compleja que la primera, ya que se aprovecha de los intervalos de confianza de las curvas.

Al igual que en la primera aproximación, se redondea la edad del paciente al entero más cercano. Después se procede a hacer el cálculo de las puntuaciones a los grupos, con ciertas restricciones:

1. Si existe un solapamiento de intervalos en los puntos de la curva de la edad del ejemplo, se descarta el cálculo de la cercanía en ese área y para ese ejemplo.
2. Si el ejemplo se encuentra entre las dos curvas, también se descarta el cálculo en ese área para ese ejemplo.
3. Si no se ha cumplido ninguna de las condiciones anteriores, acumulamos los siguientes valores en el grupo de pacientes más cercano al ejemplo:
  - Distancia entre el ejemplo y el intervalo más lejano de la curva más cercana. Por ejemplo, si el ejemplo se encuentra encima de la curva del grupo sanos, se acumulará la distancia entre el ejemplo y el intervalo inferior (el intervalo más lejano al punto).
  - Aumentar en 1 un contador de áreas. Dicho contador servirá para calcular distancias medias y para saber cuántas áreas han sido utilizadas y cuantas han sido descartadas.

Tras este proceso se dispone de la siguiente información para cada ejemplo a clasificar:

- Valor de cercanía del ejemplo al grupo de pacientes sanos (*distanciaAcumuladaSano*).
- Valor de cercanía del ejemplo al grupo de pacientes beta-amiloide (*distanciaAcumuladaBeta*).
- Número de áreas que han sido utilizadas para calcular la cercanía al grupo de pacientes sanos (*ContadorAreasSano*).
- Número de áreas que han sido utilizadas para calcular la cercanía al grupo de pacientes sanos (*ContadorAreasBeta*).

Gracias a esto, se podrá clasificar modificando el origen de los parámetros *puntuacionBeta* y *puntuacionSano* en la ecuación mencionada en el apartado anterior:

$$prediccion = \log(puntuacionBeta/puntuacionSano) \quad (5.3)$$

	Aproximación 1	Aproximación 2		
	$X - (S + B)/2$	Dist. acumulada	Dist. Media	Contador Áreas
Tasa aciertos	73,20 %	74,23 %	67,01 %	72,51 %

Tabla 5.10: Tasa de aciertos en la clasificación

		Predicción	
		Sanos	Beta
Real	Sanos	150	53
	Beta	25	63

Tabla 5.11: Matriz de confusión de la primera aproximación

		Predicción					
		Mét. acumulación		Método media		Método contador	
		Sanos	Beta	Sanos	Beta	Sanos	Beta
Real	Sanos	155	48	132	71	155	48
	Beta	27	61	25	63	32	56

Tabla 5.12: Matrices de confusión de la segunda aproximación

Donde  $puntuacionBeta$  y  $puntuacionSano$  pueden ser:

- Las distancias acumuladas del ejemplo al intervalo más lejano de la curva más cercana:
  - $-puntuacionBeta = distanciaAcumuladaBeta$
  - $-puntuacionSano = distanciaAcumuladaSano$
- Las distancias medias del ejemplo al intervalo más lejano de la curva más cercana, es decir:
  - $-puntuacionBeta = distanciaAcumuladaBeta/ContadorAreasBeta.$
  - $-puntuacionSano = distanciaAcumuladaSano/ContadorAreasSano.$
- El contador de áreas utilizadas en el cálculo de las distancias:
  - $-puntuacionBeta = ContadorAreasBeta$
  - $-puntuacionSano = ContadorAreasSanos$

Así se obtienen tres posibles clasificaciones para el ejemplo, una con cada uno de los criterios mencionados.

### 5.9.3. Resultados de la clasificación

La tabla 5.10 muestra los resultados de la clasificación de pacientes, usando las dos aproximaciones. También se muestran las matrices de confusión correspondientes a la primera (tabla 5.11) y la segunda aproximación (tabla 5.12).

Como era de esperar, los resultados obtenidos no son demasiado buenos. De las 4 clasificaciones. Las tasas de acierto más altas son la primera aproximación, que obtiene un 73,20 % y el método del contador de la segunda aproximación, que logra un 72,51 %. Como bien se adelantó, utilizar las curvas generadas para una clasificación con aproximaciones

## 5.9. Clasificación de pacientes usando las curvas

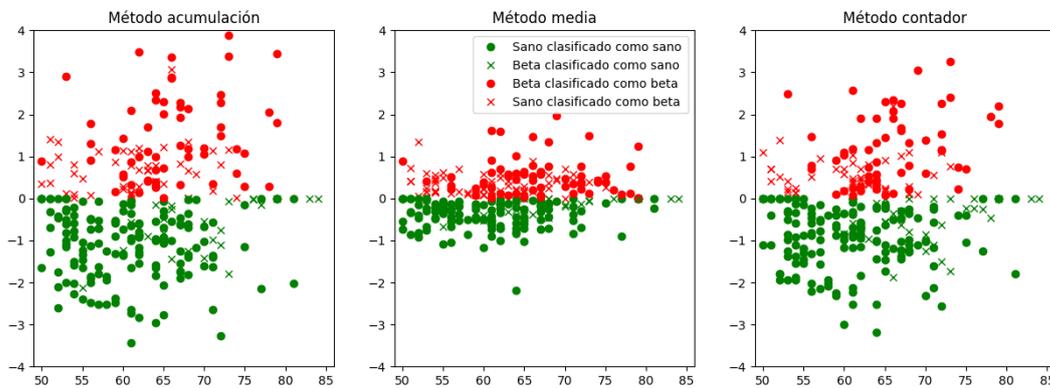


Figura 5.16: Resultados de la clasificación en los 3 métodos de la segunda aproximación

básicas como son el cálculo de distancias no aporta grandes resultados. Quizás se podrían refinar más las aproximaciones para obtener tasas de acierto más altas.

Si se observan las matrices de confusión de todas las aproximaciones, puede comprobarse como por lo general, clasifica mejor los de casos negativos, mientras que para clasificar los casos positivos presenta mayor dificultad.

Es decir, si se utilizasen las curvas como modelo clasificador, sería capaz de clasificar con cierto nivel de precisión los pacientes sanos, pero no lograría dar un diagnóstico correcto a pacientes que presenten beta-amiloide.

Con el objetivo de comprobar de una forma más visual los resultados de la clasificación, se han dibujado en 3 gráficos los resultados obtenidos para la segunda aproximación, pueden verse en la figura 5.16.

El cada gráfico representa la clasificación con los 3 métodos de la segunda aproximación: acumulación de distancias, distancia media a las curvas y conteo de áreas. El eje  $x$  representa la edad del paciente, mientras que el eje  $y$  representa el resultado de la operación  $\log(\text{puntuacion.Beta}/\text{puntuacion.Sano})$ .

El color verde indica que ha sido clasificado como paciente sano y el rojo clasificado como paciente con beta-amiloide. Si el ejemplo está dibujado con un punto significa que ha sido clasificado correctamente, en cambio, si ha sido dibujado con una cruz quiere decir que ha sido erróneamente clasificado.

En los método de acumulación de distancias y de contador de áreas (primer y tercer gráfico), se observa como por lo general, los valores de clasificación del logaritmo están más alejados de 0 que en el método de la distancia media (segundo gráfico). Unos valores más alejados de 0 implican una mayor seguridad en la clasificación, mientras que si el valor es 0 o muy cercano a él, el ejemplo resulta dudoso de clasificar con el método aplicado.

También puede verse como la mayor cantidad de ejemplos mal clasificados se encuentran en el rango  $[0,1]$  del logaritmo (eje  $y$ ). Esta información se complementa perfectamente con las matrices de confusión de la tabla 5.12, dónde puede comprobarse que a la hora de clasificar los casos como positivos es dónde más problemas se presentan.

	Aproximación 2		
	Dist. acumulada	Dist. Media	Contador Áreas
Todas las áreas	74,23 %	67,01 %	72,51 %
Top 15 áreas (dist. total)	70,79 %	64,94 %	71,47 %
Top 15 áreas (sin solapamiento)	70,79 %	65,29 %	71,82 %

**Tabla 5.13:** Tasa de aciertos en la clasificación con las 15 áreas de mayor diferencia

		Predicción			
		Dist. total		No solapamiento	
		Sanos	Beta	Sanos	Beta
Real	Sanos	153	50	150	53
	Beta	35	53	32	56

**Tabla 5.14:** Matriz de confusión con método de acumulación para las 15 áreas con mayor diferencia

		Predicción			
		Dist. total		No solapamiento	
		Sanos	Beta	Sanos	Beta
Real	Sanos	129	74	129	74
	Beta	28	60	27	61

**Tabla 5.15:** Matriz de confusión con método de media para las 15 áreas con mayor diferencia

## 5.10. Clasificación con las mejores áreas y uso del umbral

Con el objetivo de intentar mejorar los resultados en la clasificación, se han llevado a cabo dos pequeñas modificaciones:

- En vez de considerar la distancia entre el paciente y las 68 áreas cerebrales, utilizar sólo las 15 áreas con mayor diferencia entre el conjunto de ejemplos sanos y el conjunto de beta-amiloide.
- Aplicar un umbral de clasificación, de tal manera que si  $\log(\text{puntuacionBeta}/\text{puntuacionSano})$  no supera dicho umbral, el ejemplo no se clasifica.

Ambas modificaciones se aplicarán en los tres métodos de la segunda aproximación. Primeramente se han seleccionado las 15 áreas con mayor diferencia en cada una de las 3 materias usando las curvas medias generadas con las muestras Bootstrap (ver tablas 5.7, 5.8, y 5.9). Tanto con el criterio de distancia total, como con el criterio de distancia en los rangos de edad sin solapamiento.

Las tasas de aciertos pueden verse en la tabla 5.13, y las matrices de confusión en las tablas 5.14 para el método de acumulación, 5.15 para distancia media y 5.16 para contador de áreas.

5.10. Clasificación con las mejores áreas y uso del umbral

		Predicción			
		Dist. total		No solapamiento	
		Sanos	Beta	Sanos	Beta
Real	Sanos	161	42	158	45
	Beta	41	47	37	51

**Tabla 5.16:** Matriz de confusión con método de contador para las 15 áreas con mayor diferencia

	Aproximación 2		
	Dist. acumulada	Dist. Media	Contador Áreas
Umbral = 0	74,23 %	67,01 %	72,51 %
Umbral = 0.25	77,17 %	75,00 %	75,00 %
Umbral = 0.5	78,99 %	80,80 %	78,80 %
Umbral = 0.75	80,48 %	84,39 %	81,19 %

**Tabla 5.17:** Tasa de aciertos en la clasificación con las 15 áreas de mayor diferencia

		Predicción	
		Sanos	Beta
Real	Sanos	132	2
	Beta	25	14

**Tabla 5.18:** Matriz de confusión del método distancia media con umbral 0.75

Del mismo modo, se ha probado a establecer un umbral de clasificación para el mismo grupo de curvas. Los valores de los umbrales probados han sido 0,25, 0,5 y 0,75. El nuevo criterio de clasificación se aplica de la siguiente manera:

- Si  $\log(\text{puntuacionSanos} / \text{puntuacionBeta}) > \text{umbral}$ , entonces clasificar como positivo.
- Si  $\log(\text{puntuacionSanos} / \text{puntuacionBeta}) < -\text{umbral}$ , entonces clasificar como negativo.
- Si no, no clasificar el ejemplo.

Las tasas de acierto pueden verse en la tabla 5.17. En un primer momento puede determinarse que los resultados mejoran, puesto que las tasas de acierto son más altas, sin embargo para comprobarlo se muestra la matriz de confusión para el método distancia media y umbral 0.75, la prueba que ha obtenido la tasa de aciertos más alta (tabla 5.18). En ella puede verse el motivo por el que la tasa de aciertos ha aumentado: se han descartado muchos ejemplos en la clasificación. En total solo se han clasificado 16 casos del grupo beta-amiloide, cuando con los otros métodos se clasificaban alrededor de 110. Descartar una cantidad tan elevada de casos no es un buen resultado, a pesar de tener una tasa de acierto mejor que sin utilizar el umbral.

Por lo tanto, con estas últimas pruebas se puede concluir que utilizar sólo las áreas con mayor diferencia en lugar de todas las áreas para la clasificación no mejora los resultados.

## 5. PROCESO DE IMPLEMENTACIÓN

---

Además, el uso del umbral de clasificación puede aumentar la tasa de acierto, a costa de no clasificar los ejemplos más dudosos, por lo que un umbral excesivo puede llegar a empeorar los resultados.

En cualquier caso, la clasificación de los ejemplos mediante los métodos propuestos no ha sido especialmente exitosa.

---

## Capítulo 6

# Conclusiones y trabajo futuro

---

### 6.1. Análisis de resultados y Conclusiones

El objetivo principal de este trabajo ha sido el de observar la evolución de las distintas áreas cerebrales en pacientes con la proteína beta-amiloide, y comparar dicha evolución con la de pacientes sin patologías de Alzheimer, para posteriormente diseñar una metodología de comparación de evolución de áreas cerebrales según la edad y aplicarla entre ambos grupos de pacientes.

Esta evolución se ha medido observando el volumen de las tres materias principales presentes en el cerebro: materia gris (GM), materia blanca (WM) y líquido cefalorraquídeo (CSF) en cada una de las 68 áreas cerebrales. Dicha información provenía de la base de datos proporcionada por CITA-Alzhéimer, que obtuvieron sometiendo a pruebas de resonancia magnética a 537 pacientes, de los cuales 360 llevaron a cabo también la prueba de punción lumbar, una prueba que permite comprobar la presencia de biomarcadores, como lo es la beta-amiloide.

Se llevaron a cabo varias aproximaciones de las tendencias de cada una de las áreas mediante la regresión Ridge, donde se comparaban los resultados variando algunos parámetros, como el rango de la edad de predicción o  $\lambda$ , que permite reducir la sensibilidad de la curva a los datos.

Tras varias pruebas, las áreas cerebrales que más diferencia presentaban entre pacientes sanos y pacientes con beta-amiloide son las siguientes (ver tabla 6.1):

En función de la aproximación que se llevaba a cabo, el orden de diferencia de estas áreas variaba ligeramente, pero siempre aparecían entre las más diferenciadas. Por lo tanto se puede concluir que un estudio en profundidad de esas áreas cerebrales puede dar alguna pista acerca cómo se origina la enfermedad.

## 6. CONCLUSIONES Y TRABAJO FUTURO

Áreas cerebrales más relevantes divididas por materias		
Materia Gris	Materia Blanca	Líquido Cefalorraquídeo
Cerebelum_R_GM_VOL.18	Caudate_R_WM_VOL.36	Lat_Temp_Ventr_R_CSF_VOL.48
Cerebelum_L_GM_VOL.17	Caudate_L_WM_VOL.35	Lat_Temp_Ventr_L_CSF_VOL.47
C_Callosum_R_GM_VOL.46	Lingual_Gyr_R_WM_VOL.66	Caudate_R_CSF_VOL.36
C_Callosum_L_GM_VOL.45	Cuneus_R_WM_VOL.68	Ant_Orbit_Gyr_L_CSF_VOL.55
Ant_Orbit_Gyr_L_GM_VOL.55	Third_Ventr_R_WM_VOL.50	Amygd_L_CSF_VOL.3
Third_Ventr_R_GM_VOL.50	Lingual_Gyr_L_WM_VOL.65	CSF_VOL_hipp_R
Sup_Front_Gyr_R_GM_VOL.60	Brainstem_L_WM_VOL.19	CSF_VOL_hipp_L
Parahipp_L_GM_VOL.9	Cerebelum_L_WM_VOL.17	Caudate_L_CSF_VOL.35
Brainstem_R_GM_VOL.20	Cuneus_L_WM_VOL.67	Ant_Orbit_Gyr_R_CSF_VOL.56
Straight_Gyr_L_GM_VOL.53	Pallidum_L_WM_VOL.43	Straight_Gyr_R_CSF_VOL.54

**Tabla 6.1:** Las áreas cerebrales con mayor diferencia según las pruebas de este trabajo

En la tabla 6.2, se puede ver a qué lóbulos cerebrales pertenece cada una de las áreas mencionadas, esto puede darnos una idea de en qué zonas se produce la mayor diferencia entre ambos tipos de pacientes. El lóbulo temporal es la región del cerebro en la que mayor número de áreas relevantes se han encontrado, seguido de los lóbulos temporal y occipital.

Como bien se ha comentado en el capítulo 3, el lóbulo frontal se encarga de controlar el razonamiento, las emociones y parte del habla. Por su parte, el lóbulo temporal controla la memoria y parte del habla, y el lóbulo occipital la visión. La enfermedad de Alzheimer causa problemas en estas cuestiones (especialmente aquellas que son controladas por los lóbulos frontal y temporal), lo que nos puede indicar en cierta manera que los resultados obtenidos en el trabajo no andan desencaminados.

Además, a la hora de diagnosticar un caso de Alzheimer, centrarse en el estudio de dichas áreas en vez de en la observación del cerebro en su totalidad puede ahorrar un tiempo muy valioso, lo que permitiría actuar con mayor rapidez en caso de detectarse un caso positivo de Alzheimer, evitando la prueba de punción lumbar.

Adicionalmente, se trató de utilizar las curvas obtenidas en el estudio como un clasificador de pacientes de Alzheimer, para que, en caso de obtener una alta tasa de acierto, llegar a utilizarlo como un método fiable de diagnóstico que sustituyera otras pruebas invasivas como lo es la punción lumbar. Sin embargo, los resultados de la clasificación no han sido especialmente buenos. Se utilizaron un total de 4 métodos de clasificación, todos ellos basados en el cálculo de la distancia entre el paciente y las curvas de los grupos sano y beta-amiloide. Las tasas de acierto de estos métodos de clasificación, recogidos en las tablas 5.10, 5.13 y 5.17 rondaban el 70 – 80 %; resultados insuficientes si hablamos de algo tan importante como el diagnóstico de la enfermedad de Alzheimer.

## 6.1. Análisis de resultados y Conclusiones

Nombre en la bd	Nombre completo	Materias relevantes	Región cerebral
Cerebelum_R Cerebelum_L	Cerebelo dcho Cerebelo izq	GM GM, WM	Cerebelo
C_Callosum_R C_Callosum_L Third_Ventr_R	Cuerpo calloso dcho Cuerpo calloso izq Tercer ventrículo dcho	GM GM GM, WM	Centro/núcleo
Ant_Orbit_Gyr_L Ant_Orbit_Gyr_R Sup_Front_Gyr_R Straight_Gyr_L Straight_Gyr_R Pallidum_L Caudate_R Caudate_L	Giro orbitofrontal anterior izq Giro orbitofrontal anterior dcho Giro frontal superior dcho Giro recto izq Giro recto dcho Globo pálido izq Núcleo caudado dcho Núcleo caudado izq	GM, CSF CSF GM GM CSF WM WM, CSF WM, CSF	Lóbulo frontal
Parahipp_L Amygd_L hipp_R hipp_L	Giro parahipocampal izq cuerpo amigdaliano izq Hipocampo dcho Hipocampo izq	GM CSF CSF CSF	Lóbulo temporal
Lingual_Gyr_R Lingual_Gyr_L Cuneus_R Cuneus_L	giro lingual dcho giro lingual izq cúneo dcho cúneo izq	WM WM WM WM	Lóbulo occipital
Lat_Temp_Ventr_R Lat_Temp_Ventr_L	ventriculo lateral dcho ventriculo lateral izq	CSF CSF	lóbulo parietal
Brainstem_R Brainstem_L	Tronco encefálico dcho Tronco encefálico izq	GM WM	Tronco encefálico

**Tabla 6.2:** Clasificación de las áreas en los distintos lóbulos cerebrales

## 6.2. Trabajo futuro

En vista a los resultados obtenidos, quedan varias líneas de trabajo abiertas:

La primera de ellas es contrastar con las personas expertas de CITA-Alzhéimer si efectivamente las áreas obtenidas como las que mayor diferencia presentan entre pacientes sanos y pacientes con beta-amiloide son verdaderamente útiles tanto para el estudio de la enfermedad como para su diagnóstico. Esto puede ser un punto de partida de cara a futuros trabajos.

Alternativamente, si se quisiera continuar con la idea del clasificador, la búsqueda de nuevos métodos o heurísticos aplicables a la clasificación en función de la distancia paciente-curva puede ser una gran idea. Se ha descubierto que con los cuatro métodos utilizados se obtiene una gran cantidad de casos *dudosos*. Un nuevo método que reduzca esta cantidad de ejemplos y permita clasificarlos con mayor seguridad podría ayudar a aumentar las tasas de acierto actuales.

Sin embargo, para ello se requeriría de una base de datos más sólida, con una mayor cantidad de ejemplos y con una distribución más homogénea, ya que la base de datos actual concentra la mayoría de sus ejemplos en el rango de edad [50-70] años, y dificulta el estudio en edades más tempranas y más avanzadas.

Gracias gracias a que se ha diseñado la metodología e implementado el motor, sería más fácil extender los análisis realizados e incluir otras variables de análisis como pueden ser otros biomarcadores neurodegenerativos (las proteínas tau y fosfo-tau por ejemplo), el género, etc.

Los resultados de este trabajo pueden ser un buen inicio para muchos nuevos estudios que ayuden tanto al estudio como al diagnóstico de la enfermedad e intentar lograr el ansiado objetivo de evitar siempre que sea posible, la prueba de punción lumbar.

---

# Apéndice

---

Para evitar la sobrecarga de la memoria con un número excesivo de gráficas, se han recogido algunas de ellas en el siguiente apéndice. En él se recogen las gráficas que muestran las áreas cerebrales ordenadas por diferencia entre los grupos de pacientes sanos y beta-amiloide, para cada uno de los grupos de curvas y materias.

**Apéndice:** Gráficas de las áreas ordenadas por edad para los distintos grupos de curvas.



---

# Bibliografía

---

- [1] JJ Libre Guerra, L. García Arjona, and JP. Diaz Marante. Demencias y enfermedad de alzheimer un recorrido por la historia. *GeroInfo - Medigraphic*, 9:5–7, 2014. Ver página 1.
- [2] M Myers. Single brain scan can diagnose alzheimer’s disease. *SImperial College London*, 2022. Ver página 2.
- [3] M Inglese, N. Patel, K Linton-Reid, et al. A predictive model using the mesoscopic architecture of the living brain to detect alzheimer’s disease. *Communications Medicine*, 2022. Ver página 3.
- [4] Asociación Internacional de Alzhéimer. Alz - ¿Qué es el Alzhéimer? [En línea] <https://www.alz.org/alzheimer-demencia/que-es-la-enfermedad-de-alzheimer>, 2021. Accedido el 21 de abril, 2023. Ver página 9.
- [5] G García-Ribas, JL López-Sendón Moreno, and J García-Caldentey. Biomarcadores en la enfermedad de alzheimer. *Rev Neurol*, 2014. 58: 308-17. Ver página 10.
- [6] Universidad Nacional de Córdoba. Proyecto integrador - scientific figure on researchgate. Disponible en: [https://www.researchgate.net/figure/Figura-22-Representacion-histologica-de-un-tejido-cerebral-normal-y-uno-con-placas\\_fig1\\_344495348](https://www.researchgate.net/figure/Figura-22-Representacion-histologica-de-un-tejido-cerebral-normal-y-uno-con-placas_fig1_344495348). [Accedido el 10 Junio, 2023]. Ver página 10.
- [7] MayoClinic. MayoClinic - ¿Qué es el Alzhéimer? [En línea] <https://www.mayoclinic.org/es-es/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447>, 2023. Accedido el 23 de abril, 2023. Ver página 10.
- [8] National Institute of Aging. NIA - Biomarkers for dementia detection and research. [En línea] <https://www.nia.nih.gov/health/biomarkers-dementia-detection-and-research>, 2021. Accedido el 23 de abril, 2023. Ver página 11.
- [9] Anónimo. El cerebro humano. <https://www.cognifit.com/es/cerebro>, 2018. Accedido el 15 de abril, 2023. Ver página 12.
- [10] MedlinePlus. MedlinePlus - Anatomía del cerebro. [En línea] <https://medlineplus.gov/spanish/ency/anatomyvideos/000016.htm>, 2021. Accedido el 15 de abril, 2023. Ver página 12.
- [11] MA. Arraez, I. Herruzo, T. Acha, and M Benavides. Lóbulos cerebrales. manifestaciones clínicas. In *Tumores del Sistema Nervioso Central en el adulto y en la infancia*, pages 71–83. Nova Sidonia, 2003. Ver página 12.
- [12] National Institute of Neurological Disorders and Stroke. NIH - Brain basics: Know your brain. [En línea] <https://www.ninds.nih.gov/health-information/>

- [public-education/brain-basics/brain-basics-know-your-brain](#), 2018. Accedido el 17 de abril, 2023. Ver página 12.
- [13] BIOMARCADORES DE INFLAMACIÓN EN LÍQUIDO CEFALORRAQUÍDEO Y SANGRE para el diagnóstico de meningitis infecciosas - Scientific Figure on ResearchGate. Disponible en: [https://www.researchgate.net/figure/Figura-15-Lobulos-y-cisuras-del-cerebro\\_fig5\\_353236280](https://www.researchgate.net/figure/Figura-15-Lobulos-y-cisuras-del-cerebro_fig5_353236280). [Accedido el 10 Junio, 2023]. Ver página 12.
- [14] PsychoLab. Psycholab - diferencia entre neurona y neuroglía. Disponible en: <https://www.psycholab.com/diferencia-entre-neurona-y-neuroglia/>. [Accedido el 10 Junio, 2023]. Ver página 13.
- [15] Microsoft. Microsoft - ¿qué es el aprendizaje automático? [En línea] <https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-machine-learning-platform>. [Accedido el 2 Mayo, 2023]. Ver página 14.
- [16] JA Sanchez. ¿cómo aprenden las máquinas? - machine learning y sus diferentes tipos. [En línea] <https://datos.gob.es/es/blog/como-aprenden-las-maquinas-machine-learning-y-sus-diferentes-tipos>. [Accedido el 2 Mayo, 2023]. Ver página 14.
- [17] E. Alpaydin. Examples of machine learning applications. In *Introduction to Machine Learning, third edition*, pages 5–11. MIT Press, 2014. Ver página 14.
- [18] J Amar Rodrigo. Regularización ridge, lasso y elastic net con python. [En línea]: <https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html>. (CC BY 4.0) [Accedido el 3 Mayo, 2023]. Ver página 15.
- [19] J Starmer. Regularization part 1 - ridge regression. Disponible en: <https://youtu.be/Q81RR3yKn30>. [Accedido el 2 Mayo, 2023]. Ver página 16.
- [20] S Morante. Telefontech - precauciones a la hora de normalizar los datos. [En línea]: <https://empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar>. [Accedido el 24 Mayo, 2023]. Ver página 17.
- [21] Minitab. ¿qué es un procedimiento bootstrap? [En línea]: <https://support.minitab.com/es-mx/minitab/20/help-and-how-to/probability-distributions-random-data-and-resampling-analyses/supporting-topics/resampling-analyses/what-is-bootstrapping/>. [Accedido el 25 Mayo, 2023]. Ver página 18.
- [22] T Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. Ver página 18.