Universidad del País Vasco Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Computación

Trabajo de Fin de Grado

Detección de casos preclínicos de la enfermedad de Alzheimer en base a imágenes por resonancia magnética

Curso 2022

Autora

Cristina Oueghlani Rodríguez

Universidad del País Vasco Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Computación

Trabajo Fin de Grado

Detección de casos preclínicos de la enfermedad de Alzheimer en base a imágenes por resonancia magnética

Curso 2022

Autora
Cristina Oueghlani Rodríguez

Director
Ibai Gurrutxaga

Co-director
Javier Muguerza

Resumen

A pesar de que en la actualidad el ser humano disponga de múltiples herramientas con las que poder anticipar la enfermedad de Alzheimer, a día de hoy sigue habiendo técnicas desconocidas con las que poder predecir dicha muerte neuronal.

Una manera de lograr pronosticarla es mediante la realización de una punción lumbar, donde analizan principalmente diferentes tipos de proteínas. Hallar esta enfermedad a tiempo es fundamental puesto que de esta manera se podría ralentizar el proceso del envejecimiento cerebral.

El objetivo al que aspira este proyecto es evitar la realización de punciones lumbares y estimar el riesgo de Alzheimer a través del análisis de resonancias magnéticas, con el fin de anticipar la enfermedad de manera más sencilla y del modo menos invasivo posible.

Sin embargo, a pesar de haber trabajado tanto con modelos supervisados como semi-supervisados, no se ha logrado obtener una gran tasa de acierto. Por lo que se ha llegado a la conclusión de que a día de hoy poder predecir el Alzheimer mediante volúmenes cerebrales no es un método fiable.

Índice de contenidos

Índice de figuras

Índice de tablas

Índice de ecuaciones

Capítulo 1

Introducción

1.1 Contexto

Hoy en día la enfermedad de Alzheimer (EA) es una de las enfermedades neuropsiquiátricas que más se investigan, debido a que en las últimas décadas ha aumentado considerablemente los casos, en los que, a día de hoy afectan a más de 22 millones de personas en el mundo. De manera análoga, se prevé que en tres décadas se dupliquen los casos. No solo son los propios pacientes los perjudicados sino también su entorno que sufren diariamente las consecuencias de la misma.

Este trastorno neurodegenerativo fue hallado por el psiquiatra alemán Alois Alzheimer en 1907. El doctor tenía una paciente la cual padecía perdidas de memoria a corto plazo y alucinaciones auditivas. Tras su fallecimiento, pudo encontrar en la autopsia características atípicas respecto al cerebro de su paciente en comparación con uno saludable. Aparecieron distintos tipos de anomalías; en particular la corteza cerebral era más estrecha de lo normal. Conjuntamente observó entre las neuronas un incremento de depósitos de placas de la proteína amiloide y un exceso de fabricación de ovillos de la proteína tau.

Sorprende que este cuadro, descrito formalmente en 1906, haya permanecido en una hibernación científica por casi 80 años para luego despertar creando un explosivo interés por su esclarecimiento y curación. [1]

Cabe remarcar que el Alzheimer no es lo mismo que la demencia, sin embargo, sí es una de las principales causas de demencia. La demencia se define como el deterioro adquirido en las capacidades cognitivas que entorpece la realización satisfactoria de actividades de la vida diaria. [2]

El Alzheimer es una demencia progresiva que tiene el déficit de memoria como uno de sus síntomas más tempranos y pronunciados. Por lo general, el paciente empeora progresivamente, mostrando problemas perceptivos, del lenguaje y emocionales a medida que la enfermedad va avanzando. [3]

Su aparición suele tener lugar a partir de los 65 años de edad, sin embargo, la enfermedad suele comenzar 10 años antes de la aparición de los primeros síntomas. Comienza con una leve pérdida de memoria y otras dificultades cognitivas, a pesar de ello, en esta fase inicial son capaces de desempeñar por sí mismos las tareas cotidianas.

En función de la edad de aparición de los síntomas se clasifica en:

- Enfermedad de Alzheimer de inicio precoz, si el comienzo es antes de los 65 años.

- Enfermedad de Alzheimer de inicio tardío, si comienza después de los 65 años.

A su vez estas dos formas se clasifican en dos subtipos:

- Familiar, si hay historia familiar.
- Esporádica, si no hay antecedentes familiares. [2]

Algunos de los síntomas que componen el Alzheimer son, entre otros; pérdida de memoria, dificultad a la hora de resolver problemas, complejidad en completar tareas cotidianas cuando antes no le resultaban costosas, confusión con la hora y los lugares, problemas a la hora de comunicarse, pérdida constante de objetos, etcétera.

1.2 Motivación

A pesar de los progresos que ha habido en los últimos años en el campo de Alzheimer, a día de hoy sigue siendo desconocido el tratamiento para su total desaparición. En el tiempo actual se ha conseguido el método por el cual es posible controlarlo y permitir que la enfermedad no avance de manera precipitada. A pesar de ello, hoy por hoy no se ha conseguido la manera de solventar dicha enfermedad y todo el daño que causa en su transcurso.

No solo son las propias personas que sobrellevan la enfermedad de Alzheimer las que sufren, paralelamente también lo hacen sus familiares y seres queridos que ven como cada día el deterioro de la persona a la que aman y como su esencia se desvanece

Por tanto, este proyecto se enfocará en la búsqueda de unos buenos resultados de clasificación con el propósito de realizar una pequeña contribución a la búsqueda de la prevención para esta enfermedad, que deja tanto dolor y sufrimiento a su paso.

1.3 Objetivo general

El objetivo principal de este proyecto de investigación es; mediante el aprendizaje automático, detectar de manera sencilla y lo menos invasiva posible la enfermedad de Alzheimer, trabajando únicamente con los escáneres cerebrales y dejando a un lado las punciones lumbares, debido a que esta última técnica es más agresiva.

Para la realización de este trabajo se ha llevado a cabo una colaboración con la fundación CITA[1] Alzheimer, quien ha proporcionado a través de la realización de dos estudios diferentes, datos cerebrales de 537 personas.

A continuación, se resumen las tareas principales de cara a conseguir el objetivo de este trabajo. Una vez analizado todos los datos de los pacientes, y habiendo localizado los valores perdidos, erróneos o extremos por medio de histogramas y boxplot, sucesivamente se han analizado las correlaciones entre las variables. Para ello, se ha utilizado la matriz de correlación de Pearson.

El siguiente paso a realizar ha sido una selección de variables. Para ello, se ha usado el código del proyecto Zixiao Shen [4] con el fin de encontrar las variables que estén más correlacionas con la clase que se quiere predecir, debido a que cuanto mayor sea la correlación mayor probabilidad habrá de predecir correctamente la clase.

A continuación, se han normalizado los valores de la base de datos debido a lo pequeños que son, ya que cabe la posibilidad de dificultar el trabajo a los clasificadores.

Seguidamente se han aplicado distintos algoritmos de mediante el uso de la librería scikit-learn[2], como el método de los K vecinos más cercanos (KNN), árboles de clasificación, Random Forest, Naive bayes, Redes Neuronales y máquinas de soporte vectorial (SVM). Con los distintos

[1] Centro de Investigación y terapias Avanzadas.

[2] https://scikit-learn.org/stable/

resultados que han aportado estos algoritmos se ha analizado el rendimiento del modelo mediante distintas métricas de clasificación, como por ejemplo accuracy, recall, precition, f1…

Este proceso se ha realizado para procurar clasificar dos de las columnas de la base de datos: la primera columna denota si el paciente sufre de Alzheimer a través de alguno de los indicadores. La segunda, por el contrario, refleja si el paciente es positivo en un indicador en concreto, exactamente refleja si posee niveles altos de la proteína beta-amiloide.

Por último, se han analizado los casos sin etiquetar (valores *NaN* de la columna indicadora de si el paciente sufre Alzheimer) es decir, ahora el proyecto cambia el enfoque a un aprendizaje semi-supervisado. En el transcurso del proyecto el objetivo constante ha sido alcanzar los resultados más óptimos posibles.

1.4 Estructura del documento

A continuación, se describirán los distintos capítulos que conforman este documento:

- Capítulo 1:

  Breve **introducción** del tema del que trata el proyecto. Así mismo, se examina el contexto, la motivación, los objetivos principales y la estructura de este documento.

- Capítulo 2:

  Formado por la **planificación y gestión del proyecto**, muestra los paquetes de trabajo, la duración estimada y el análisis de riesgos que la componen.

- Capítulo 3:

  En este capítulo, **marco teórico**, se explica de manera detallada el funcionamiento del cerebro y las partes que lo conforman. De igual manera se distingue el fundamento, las manifestaciones clínicas, el desarrollo, la detección, el tratamiento, las consecuencias y la prevención de Alzheimer. Así mismo se exponen los distintos tipos de aprendizaje existentes y el funcionamiento de los clasificadores que se han empleado a lo largo de este proyecto. Por último, se tratan las diferentes métricas empleadas a la hora de evaluar el rendimiento del modelo.

- Capítulo 4:

  Conformado por el **origen de los datos** se presenta el escenario para tener conocimiento acerca de la realización de los 2 estudios neurológicos que elaboró la fundación *Cita-Alzheimer*. Del mismo modo se explican ciertas columnas relevantes de la base de datos.

- Capítulo 5:

  En este capítulo compuesto por el **desarrollo del trabajo y los resultados**, aparecen detallados los dos entornos de trabajo que se han empleado, así como la implementación y los resultados obtenidos.

- Capítulo 6:

  Integra las principales **conclusiones** del proyecto, del mismo modo que se tratan las posibles **líneas de trabajo abiertas** que puedan suponer la mejora.

Capítulo 2

Planificación y gestión del proyecto

2.1 Planificación

La planificación es una parte fundamental en la realización del proyecto, pues recaba la estructura de la descomposición que se ha creado, así como la duración estimada para cada tarea.

Con el fin de organizar el trabajo, se ha creado una estructura de descomposición en la que se aprecia de manera visual la jerarquía que se ha llevado a lo largo del proyecto, orientada a cumplir con los objetivos de este mismo.

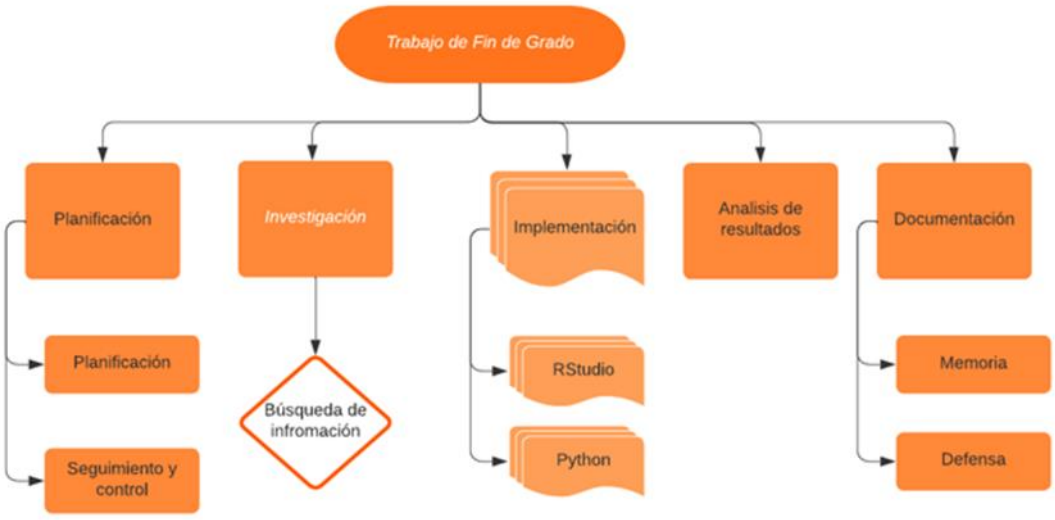

Figura 1 Diagrama de flujo

2.2 Gestión del proyecto

Para la óptima gestión del proyecto, se ha realizado periódicamente un seguimiento y control del mismo. Se ha redactado esta memoria que describe íntegramente lo que abarca el proyecto, se ha diseñado un programa que pretende dar solución a la propuesta de la fundación *cita*

Alzheimer, y por último se ha hecho una presentación en formato de diapositivas para presentar el trabajo ante un tribunal.

2.2.1 Seguimiento y control

El seguimiento y control del proyecto se ha realizado mediante reuniones periódicas por medio del correo electrónico y la plataforma de videoconferencias *webex*. El objetivo principal de las reuniones mantenidas entre el director y autora, ha consistido en la realización de un seguimiento del desarrollo del trabajo a lo largo de sus diferentes fases, de forma que se asegure la obtención de los objetivos inicialmente previstos y las actualizaciones que se han ido incorporando en el transcurso del tiempo al documento.

2.2.2 Tiempo de desarrollo de tareas

A fin de exponer el tiempo de dedicación previsto por cada tarea, se ha creado la siguiente herramienta gráfica que permite observar el alcance de las distintas fases expuestas en la planificación.

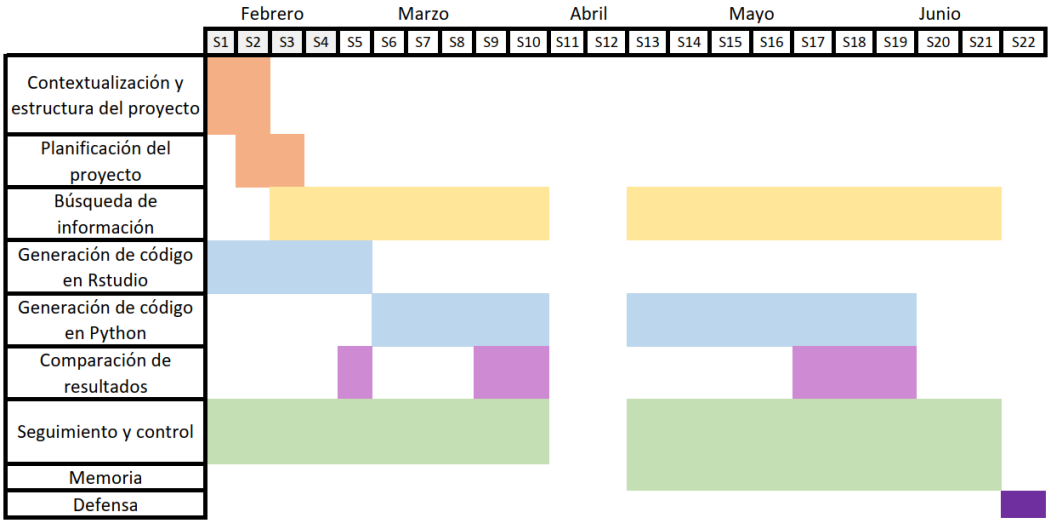

Figura 2 Tabla del tiempo de desarrollo de tareas

Contextualización y planificación del proyecto
Búsqueda de información
Generación de código
Seguimiento y documentación del trabajo
Defensa

2.2.3 Análisis de riesgos

Se ha realizado un análisis de los posibles eventos no deseados que puedan producir problemas o retardos en el proyecto. De la misma manera, se ha buscado una medida de prevención para cada amenaza identificada:

- La posible amenaza de la comprensión de los términos médicos crearía dificultad a la hora de dominar el área de la neurología. En dicho caso, se optaría por un incremento de tiempo a la hora la búsqueda de información sobre este campo.

- A pesar de que exista una baja probabilidad de perder la información recabada hasta el momento, se han ido creando periódicamente copias de seguridad en la plataforma Google Drive.

- No obtener unos buenos resultados a medida que se conforma el proyecto establecería la búsqueda de nuevos métodos con el propósito de intentar lograr resoluciones óptimas.

2.2.4 Desviaciones

A pesar de haber creado un plan inicial, un problema que ha habido a lo largo del proyecto ha sido la obtención de unos malos resultados, una preocupación incesante que truncaba por cada propósito que se había establecido. De modo que esta desviación ha obligado a buscar nuevos métodos y recursos con los que intentar obtener resultados óptimos.

Respecto al resto de tareas se han ceñido a los propósitos iniciales cumpliéndose periódicamente en el transcurso del proyecto.

2.2.5 Presentación

Tras la sinterización de la documentación se ha elaborado una presentación compuesta por diapositivas donde integra los contenidos principales para que permita una explicación clara de los temas abordados en el proyecto. A su vez cuenta con gráficos e imágenes para una explicación más ilustrativa del desarrollo.

Capítulo 3

Marco teórico

3.1 Entendimiento del cerebro

El cerebro humano pesa alrededor de kilo y medio, alberga cerca de diez mil millones de células llamadas neuronas, las cuales generan impulsos eléctricos para comunicarse entre sí, además de producir cambios químicos que permiten a dicho órgano llevar a cabo las funciones más sorprendentes, complejas y misteriosas del cuerpo humano, como son la generación de pensamientos o emociones, la imaginación, el lenguaje, el comportamiento, entre otras. [5]

3.1.1 Las neuronas

La neurona es un tipo de célula que representa la unidad estructural y funcional del sistema nervioso. Su función consiste en transmitir información a través de impulsos nerviosos, desde un lugar del cuerpo hacia otro. Estos impulsos nerviosos[3] son impulsos químicos y eléctricos. (…)

Las neuronas forman una extensa red en el cuerpo por donde circula el impulso nervioso en forma de mensaje químico y eléctrico. Este impulso viaja siempre en el mismo sentido, es decir, llega a la neurona a través de las dendritas, se procesa en el soma y posteriormente se transmite al axón, el cual se comunica con las dendritas de la contigua. [6]

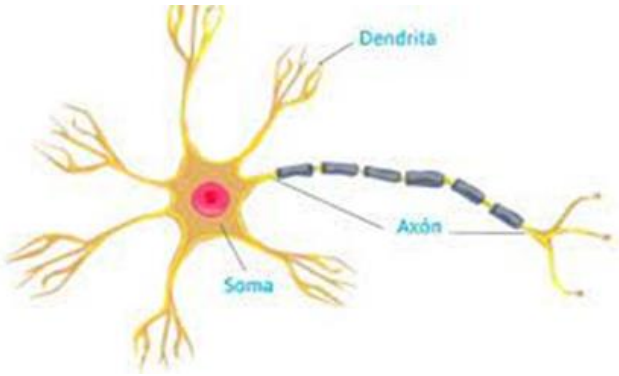

Figura 3 Partes que conforman una neurona

[3] Las neuronas transmiten señales eléctricas y químicas entre ellas para comunicar con precisión, rapidez y a larga distancia con otras células, ya sean nerviosas, musculares o glandulares.

3.1.2 El cerebro

El cerebro se encuentra conformado por dos hemisferios: el izquierdo, responsable de las funciones racionales, así como operaciones matemáticas, memoria verbal, capacidad de análisis… y el derecho; el encargado de expresiones no verbales como recuerdo de melodías, reconocimiento de personas, interpretación de imágenes…

Compuesto aproximadamente un 60% por materia blanca y un 40% de materia gris, la primera constituye la infraestructura del cerebro y la capa protectora que lo resguarda de posibles accidentes. Mientras que la segunda comprende las zonas del sistema nerviosos y se encuentra formada por la propia materia de las neuronas.

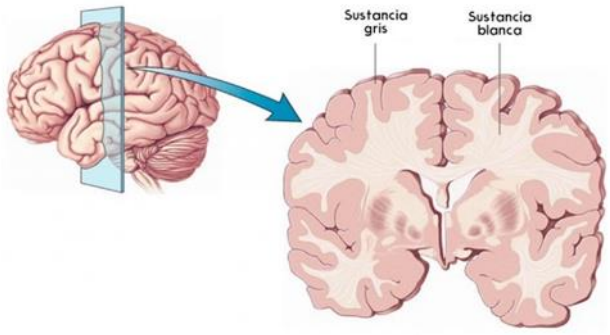

Figura 4 Materia blanca y gris del cerebro

La presencia de materia gris nos indica que la parte del encéfalo en la que se encuentra recibe información desde muchas áreas de sustancia blanca y que, de algún modo, funcionan como *clusters* de procesamiento de información y en los que los impulsos nerviosos que viajan por los axones se encuentran con un relevo que los dirige hacia otro destino.

Esto implica, entre otras cosas, que la materia gris y la materia blanca se necesitan para trabajar tal y como deben; no en vano son dos tipos de tejido cerebral diferenciados por la concentración

de la parte de las neuronas que predominan más en ellas (axones o somas), y estas pequeñas células nerviosas forman una unidad orgánica que no se puede separar sin destruirla. [7]

3.2 Enfermedad de Alzheimer

Este apartado se encuentra orientado al entendimiento de esta enfermedad que, entre otras características, supone la pérdida de memoria. A su vez, se explicará detalladamente las manifestaciones clínicas, el desarrollo de la enfermedad, la detección de la misma, los tratamientos que existen, los diferentes tipos de consecuencias y las posibles formas de prevención.

3.2.1 Qué es

"El Alzheimer es un tipo de demencia que causa problemas de memoria, pensamiento y comportamiento. Los síntomas generalmente se desarrollan lentamente y empeoran con el tiempo, hasta que son tan graves que interfieren con las tareas cotidianas." [8]

Una manera más precisa de explicarlo es como lo hace el sitio web de la organización *Mayo Clinic* "es un trastorno neurológico progresivo que hace que el cerebro se encoja (atrofia) y que las neuronas cerebrales mueran. La Enfermedad de Alzheimer es la causa más común de demencia, un deterioro continuo en el pensamiento, el comportamiento y las habilidades sociales que afecta a la capacidad de una persona para vivir de forma independiente". [9]

3.2.2 Manifestaciones clínicas

El resultado probable es desalentador. El trastorno generalmente progresa en forma permanente. Es común que se presente incapacidad total y la muerte normalmente sucede en

un lapso de 15 años, por lo general, a causa de una infección (neumonía por aspiración) o una insuficiencia de otros sistemas corporales. [10]

El procedimiento por el cual se reconoce la enfermedad es mediante distintos signos y síntomas. Los signos hacen referencia a aquellos indicios que se observan visualmente y pueden ser cuantificados, como por ejemplo la fiebre, un sangrado… Mientras que los síntomas son aquellos que únicamente percibe la propia persona que lo está padeciendo, ya sea el dolor, el cansancio…

Las manifestaciones más frecuentes que sufren los pacientes de Alzheimer son:

El desvanecimiento de la memoria hasta el punto de verse afectados por el olvido de sus seres queridos. Otro aspecto que se ve influido son tareas fisiológicas o habilidades básicas como no poder controlar la incontinencia (urinaria y fecal).

Otro síntoma es la desorientación temporoespacial, abandonan la facultad de estar al corriente para saber el día, mes o año en el que viven. A su vez, pierden la capacidad de orientación, en otros términos, no son capaces de reconocer el espacio físico en el que se encuentran, como identificar su propio domicilio o el barrio en el que residen.

Se ven afectados por una alteración en el estado de ánimo, es decir, sufren un trastorno de la conducta. Llegado el punto suelen apreciarse por parte del paciente conductas depresivas, arrebatos violentos… influyendo notablemente en la calidad de vida del paciente y en la de sus familiares.

Al mismo tiempo se encuentran problemas a la hora de comunicarse. El lenguaje se ve afectado, perdiendo la capacidad para comunicarse con los demás. La pérdida constante de objetos es otra señal de demencia, por ejemplo, les resulta difícil recordar dónde han dejado objetos, como puede ser, el mando de la televisión. Otro síntoma que caracteriza esta

enfermedad es la complejidad en completar tareas cotidianas cuando antes no le resultaban costosas.

Dentro del campo de los síntomas psicóticos se puede encontrar distintos apartados como las alucinaciones: los pacientes perciben a través de cualquier vía sensorial algo inexistente, las más frecuentes son las visuales. En este apartado también se localizan los delirios: sufren una dificultad en la interpretación de la realidad. Pueden ir acompañados de agresividad tanto verbal como física. Por ejemplo, que le hablan personas inexistentes, que le hablan desde el interior de la televisión y que incluso haya extraños dentro del hogar.

Las alteraciones neuropsicológicas en la enfermedad de Alzheimer son:

- *Memoria*: deterioro en la memoria reciente, remota, inmediata, verbal, visual, episódica y semántica.
- *Afasia*: deterioro en funciones de comprensión, denominación, fluencia y lectoescritura.
- *Apraxia*: tipo constructiva, apraxia del vestirse, apraxia ideomotora e ideacional.
- *Agnosia*: alteración perceptiva y espacial. Este perfil neuropsicológico recibe el nombre de Triple A o Triada afasia-apraxiaagnosia. No todos los síntomas se dan desde el principio, sino que van apareciendo conforme avanza la enfermedad. [10]

3.2.3 Desarrollo de la enfermedad

En la mayoría de los pacientes que sufren esta enfermedad, se puede observar que empieza a hacerse notoria una vez alcanzados los 65 años de edad, sin embargo, se desarrolla mucho antes de que aparezcan los primeros síntomas.

Sucede progresivamente; primero se encuentra la fase pre clínica donde partes del cerebro empiezan a verse afectadas. Suele pasar desapercibida, ya que el inicio es paulatino. Los síntomas más comunes suele ser la tendencia al olvido, pérdida de la noción del tiempo y desubicación temporal. Esta etapa suele durar entre 10-15 años.

La segunda fase es la prodrómica o pre demencia donde el paciente sufre un deterioro cognitivo más notable debido a que los signos y síntomas se vuelven más evidentes. En esta etapa intermedia empiezan a olvidar acontecimientos recientes, incluso los nombres de las personas de su entorno. Sufren una desubicación en el propio hogar, se incrementa la dificultad a la hora de comunicarse, empiezan a necesitar ayuda en el cuidado personal y empiezan a hacerse notorios cambios en el comportamiento.

Por último, está la fase de demencia o etapa tardía, donde el paciente se vuelve totalmente dependiente y no es capaz de desenvolverse en las actividades de la vida cotidiana. Las alteraciones que sufre en la memoria son graves y los síntomas y signos resultan incuestionables como la dificultad en el reconocimiento de familiares, una creciente desubicación, tanto espacial como temporal. La ayuda en el cuidado personal se vuelve indispensable y se intensifican las alteraciones en el comportamiento, hasta llegar incluso a las agresiones físicas.

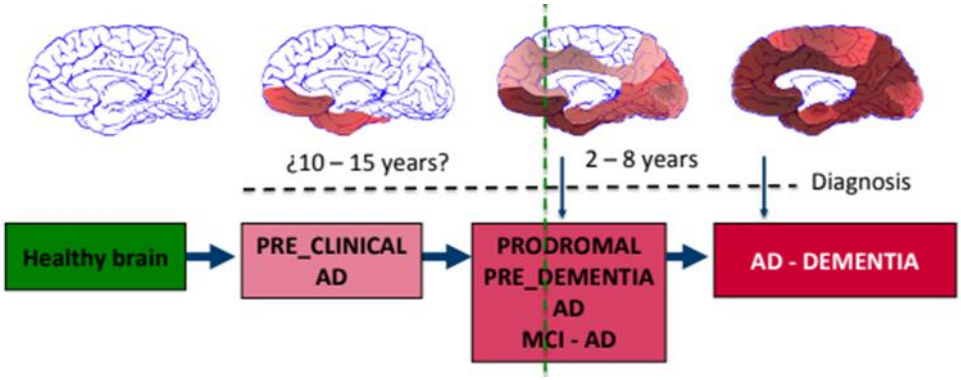

Figura 5 Desarrollo de la enfermedad de Alzheimer

Lo que ocurre en el cerebro es que hay una reducción notable de su masa. A pesar de que el cerebro pesa unos 1,5 kg, en las últimas fases de la enfermedad, cuando se encuentra totalmente contraída, disminuye su peso alrededor de 1kg. Tal y como se aprecia en la siguiente imagen, el volumen y la masa del cerebro de la derecha se ha reducido un tercio de tamaño con respecto al de la izquierda.

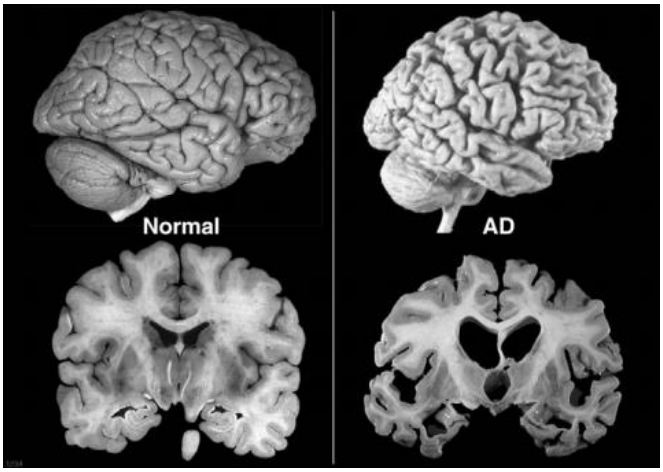

Figura 6 Comparación de un cerebro sano y uno enfermo

3.2.4 Detección

Las proteínas juegan un papel fundamental en el cerebro, ya que, gracias a la fabricación de estas, las neuronas consiguen comunicarse entre ellas. Las proteínas más significativas en el desarrollo de Alzheimer son las proteínas β-amiloide, tau y fosfo tau.

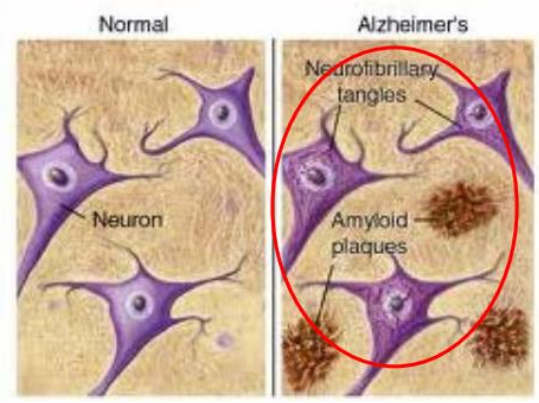

Figura 7 placas de amiloides en un cerebro que sufre Alzheimer

La proteína β-amiloide o amiloide es esencial en la transmisión de información entre neuronas. Sin embargo, la presencia de estas placas de proteína en el líquido cefalorraquídeo genera una muerte neuronal debido a los efectos tóxicos originados en estas, que a su vez genera una lenta disminución del volumen cerebral.

De igual forma, dentro de las neuronas existe la proteína tau; un componente natural empleado para proporcionar estructura a las neuronas. A causa de la destrucción de la neurona, el contenido sale haciendo que en el líquido haya más cantidad TAU de lo habitual.

De forma similar ocurre con la proteína fosfo tau o tau fosforizada, en un cerebro sano la proteína tau se une al fósforo, el cual existe de manera natural en el líquido, produciendo uniones químicas y convirtiéndose en TAU fosforizado. Sin embargo, un exceso elevado indica una mayor probabilidad de padecer esta enfermedad neurodegenerativa.

Una manera de detección de Alzheimer antes de que aparezcan los primeros síntomas es mediante punciones lumbares. Mediante este método extraen una pequeña cantidad de líquido espinal y miden los niveles de proteínas. Una vez obtenidos los resultados, los médicos son capaces de determinar si el paciente está desarrollando la enfermedad midiendo las cantidades de las proteínas antes mencionadas.

A pesar de ser una manera efectiva para determinar el riesgo de sufrir Alzheimer, es una práctica invasiva no agresiva que genera que muchos pacientes rechacen llevarla a cabo. Debido a lo cual, se están estudiando otras técnicas para detectar la enfermedad, como pueden ser las resonancias magnéticas. En este procedimiento se visualiza la estructura interna del cerebro y se localiza la posible existencia de anomalías, con el fin de estimar el riesgo de sufrir esta enfermedad. A través de imágenes miden principalmente la posible disminución en el tamaño del cerebro asociada con el deterioro cognitivo.

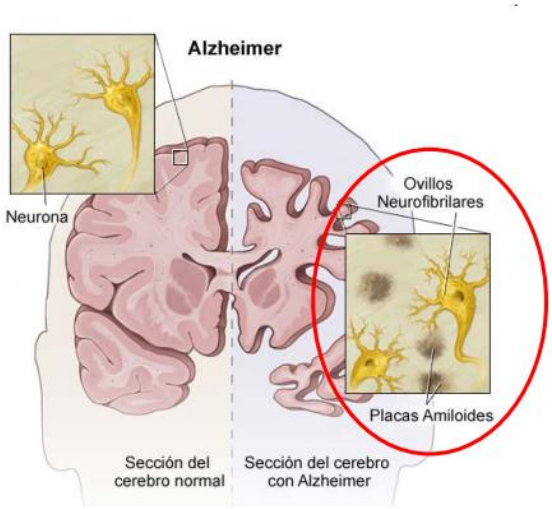

Figura 8 Lesiones neuropatológicas

3.2.5 Tratamiento

A día de hoy no se ha encontrado tratamiento alguno para la muerte neuronal que supone esta enfermedad. Sin embargo, se han encontrado distintas maneras de ralentizar la enfermedad mediante distintos tratamientos que frenen su progreso.

Existen tratamientos de tipo farmacológico y no farmacológico, ambos útiles y complementarios. El tratamiento farmacológico está destinado a modificar los síntomas

Se ha probado la eficacia de fármacos anticolinesterásicos que tienen una acción inhibidora de la colinesterasa, la enzima encargada de descomponer la acetilcolina, el neurotransmisor que falta en el Alzheimer y que incide sustancialmente en la memoria y otras funciones cognitivas. Con todo esto se ha mejorado el comportamiento del enfermo en cuanto a la apatía, la iniciativa y la capacidad funcional y las alucinaciones, mejorando su calidad de vida. [11]

A diferencia de los tratamientos farmacológicos, la eficacia de las intervenciones no farmacológicas resulta difícil de demostrar. No obstante, que no exista suficiente evidencia científica no significa que no funcionen o no puedan ser recomendados. En este campo se encuentran ciertas vitaminas que ayudan al mantenimiento de las funciones cognitivas en estos pacientes como vitaminas B12, B6, Ácido fólico.

No evitarán el avance de la enfermedad, ni mejorarán las capacidades cognitivas, aunque sí pueden contribuir a ralentizar el deterioro cognitivo y funcional. Favorecen la calidad de vida.

3.2.6 Consecuencias

Las consecuencias que produce esta enfermedad afectan a tres grupos diferentes, pues el impacto que sufre el enfermo, los cuidadores - familia y en la sociedad puede ser de carácter físico, psicológico, social y económico.

Para el enfermo, la discapacidad y dependencia pueden resultar abrumadoras. Las consecuencias físicas y psicológicas sobre el enfermo se han detallado en el apartado de manifestaciones clínicas.

Los cuidadores o familiares juegan un papel fundamental en el proceso de deterioro del paciente, pues son ellos los que le acompañan en el transcurso del largo camino de esta enfermedad. Sufren grandes cargas emocionales pues ven como su ser querido, en caso de ser familiar, va deteriorándose. Paralelamente la presión física que suele conllevar el estar al cuidado de otra persona suele variar según el estado en que se encuentre la enfermedad, pero suele ser duro. Adicionalmente, suele suponer un costo económico que puede causar estrés en las familias.

Por último, también hay importantes repercusiones sociales y económicas en lo que respecta a los costos médicos y sociales directos y a los costos referidos a la atención prestada fuera del ámbito institucional.

3.2.7 Formas de prevenirlo

Como bien se ha comentado con anterioridad, puede ocurrir que el motivo por el cual se diagnostica esta enfermedad sea por un historial familiar, o sea de manera esporádica, es decir, si no hay antecedentes familiares.

Tanto para una como para otra, existen métodos con los cuales se puede intentar reducir el riesgo de padecer demencia:

- Haciendo ejercicio con regularidad.
- No siendo fumador.
- Evitando el uso nocivo del alcohol.
- Controlando el peso.
- Tomando una alimentación saludable.
- Manteniendo una tensión arterial y unas concentraciones sanguíneas adecuadas de colesterol y glucosa.
- Otros factores de riesgo potencialmente modificables: la depresión, el bajo nivel educativo, el aislamiento social y la inactividad cognitiva.

3.3 Aprendizaje automático

El aprendizaje automático es un tipo de inteligencia artificial que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente. El aprendizaje

automático se centra en el desarrollo de programas informáticos que pueden cambiar cuando se exponen a nuevos datos. [12]

En este apartado se conocerán tres técnicas distintas de aprendizaje: supervisado, no supervisado y semi-supervisado.

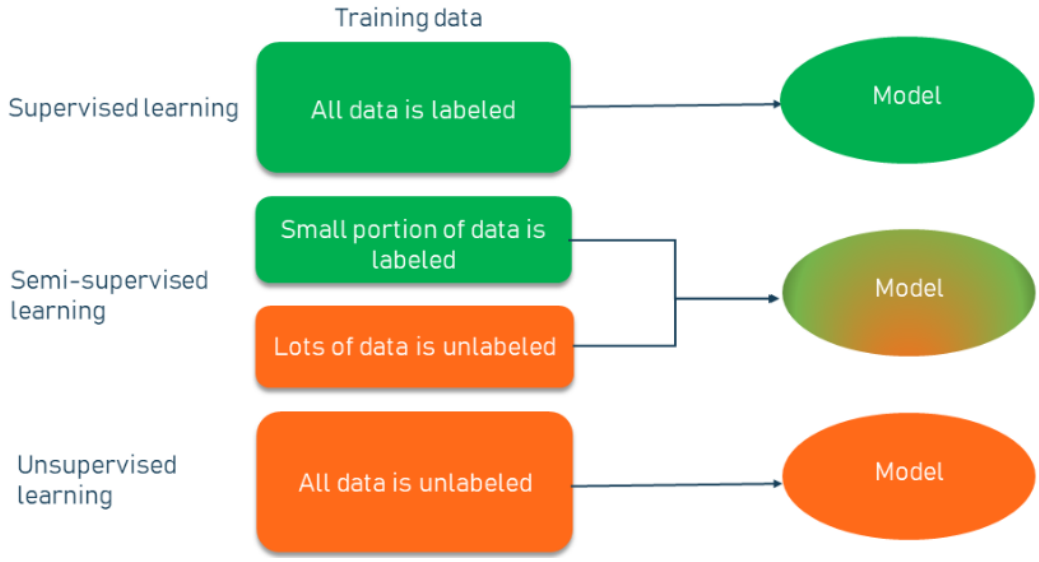

Figura 9 Aprendizaje automático supervisado, no supervisado y semi-supervisado

3.1 Aprendizaje supervisado

A partir de un conjunto de datos, previamente etiquetados manualmente, se entrena un modelo para que en el futuro pueda predecir la clase de manera automática a través de sus características.

La variable a predecir se puede categorizar de dos maneras: clasificación y regresión. El primer método, de acuerdo con los patrones que ha aprendido previamente, predice el grupo al que pertenece el dato que se está analizando. El segundo por el contrario pronostica un número, por lo que devuelve un valor determinado y no un grupo.

El funcionamiento es el siguiente: partiendo de la base de datos que se dispone, una proporción se emplea para entrenar el algoritmo con la finalidad de que más adelante pueda efectuar correctamente las predicciones. A esto se le conoce como fase de entrenamiento.

Seguidamente se realiza la fase de la prueba; los datos de los cuales no se han hecho uso se emplean para realizar pruebas. De esta manera se podrá saber si la clase obtenida coincide o no con la etiqueta y se podrá tener conocimiento de si el algoritmo realiza bien las pruebas.

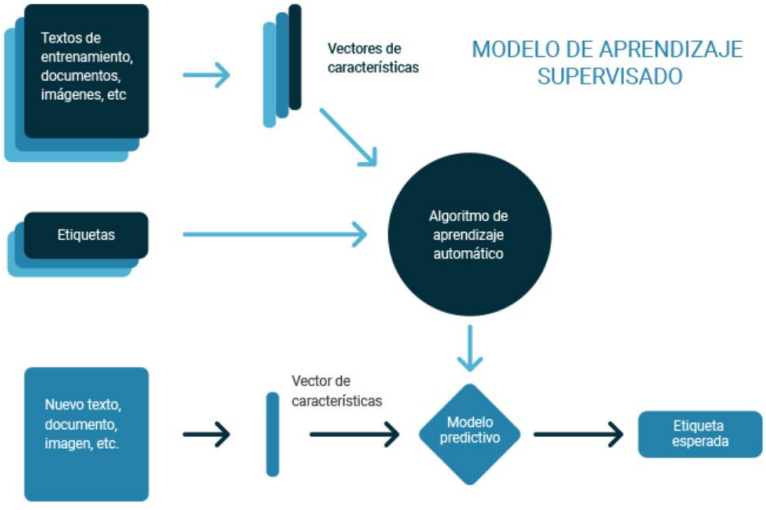

Figura 10 Modelo de aprendizaje supervisado

Algunos algoritmos que se pueden implementar en el aprendizaje supervisado serán los que se vean detalladamente en el apartado de Clasificadores.

3.2 Aprendizaje no supervisado

En el aprendizaje no supervisado únicamente se dispone de la muestra de los datos, por ende, se desconoce la etiqueta. En razón de lo cual, el modelo intenta encontrar patrones, similitudes, distancias ente las muestras o grados con concurrencia entre los datos según el tipo de características que haya.

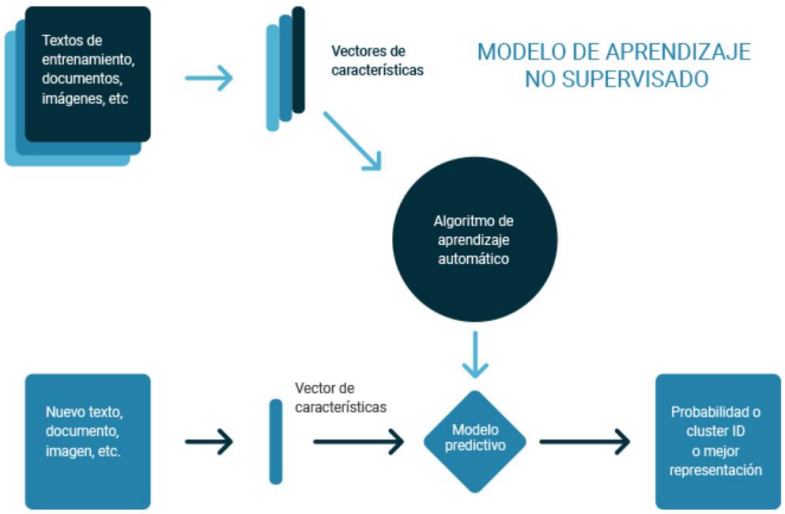

Figura 11 Modelo de aprendizaje no supervisado

Debido a que en el proyecto no se ha hecho uso de este método, no se profundizará en él.

3.3 Aprendizaje semi-supervisado

La combinación entre el aprendizaje supervisado y no supervisado es el resultado del semi-supervisado. Partiendo de un conjunto de datos, se tiene conocimiento sobre algunas de las etiquetas mientras que otras se desconocen. A partir de este punto, el modelo aprende sobre las muestras etiquetadas y posteriormente lo aplica al resto de datos sin etiquetar.

El desempeño del aprendizaje semi-supervisado comienza por el autoaprendizaje, a partir de una cantidad de datos etiquetados se entrena al modelo mediante métodos supervisados. En función de lo aprendido a partir de lo entrenado, realiza predicciones para el resto de modelo que se encuentra sin etiquetar.

A partir de este punto, conserva las predicciones más probables, es decir, las que superan el nivel de confianza impuesto por el programador. En caso de superar dicho grado de seguridad, se incorpora al conjunto de datos etiquetados con el propósito de entrenar el modelo mejorado a partir de dicha nueva entrada. Este procedimiento se realiza un número de veces considerables, dicho de otra manera, se efectúan varias iteraciones. Esta es una de las maneras de realizar aprendizaje semi-supervisado, ciertamente es la más habitual y la que se ha empleado en el proyecto.

El rendimiento varía sumamente de un conjunto de datos a otros, en ocasiones el método semi-supervisado disminuye la eficiencia frente a la vía supervisada.

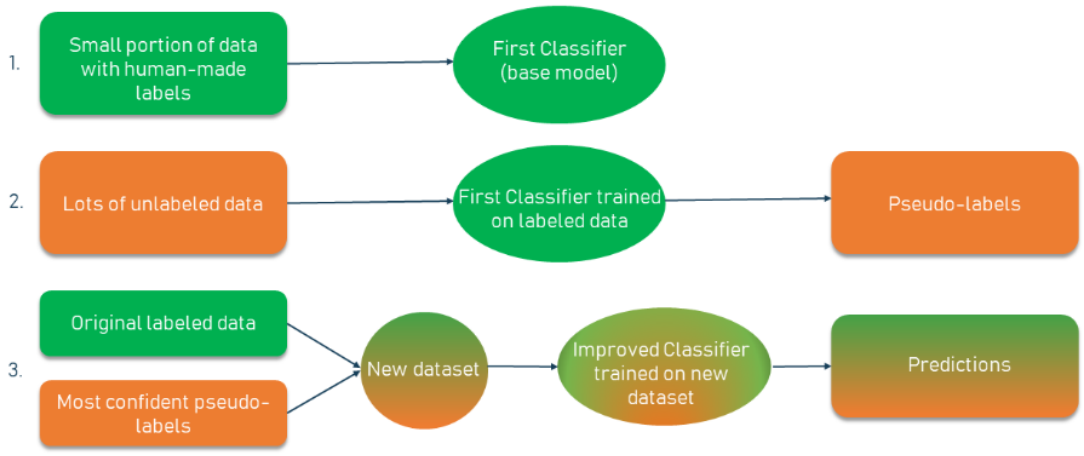

Figura 12 Modelo de aprendizaje semi-supervisado

3.4 Clasificadores

La finalidad es, mediante distintos algoritmos, entrenar el modelo con la intención de que sea capaz de determinar la categoría a la que pertenece un dato en particular. Las técnicas de aprendizaje automático utilizadas en el proyecto son los K vecinos más cercanos (KNN), los árboles de clasificación, Random Forest, Naive Bayes, las redes neuronales y las máquinas de soporte vectorial.

3.4.1 KNN

El método de los k vecinos más cercanos (KNN, K Nearest Neighbors, en inglés) estima directamente la probabilidad a posteriori de la clase, o sea, asigna la muestra x a la clase más frecuente de entre sus k vecinos más cercanos, según una cierta medida de similitud o distancia. La fase de entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de las clases de los ejemplos de entrenamiento. En la fase de clasificación, la evaluación de un ejemplo del que no se conoce su clase, es representada por un vector en el espacio de rasgos. Se calcula la distancia entre los vectores almacenados y del nuevo vector y se seleccionan los k ejemplos más cercanos, una distancia alta entre individuos nos indica que son muy diferentes y una baja que son muy similares. El nuevo ejemplo es clasificado con la clase que más se repite en los vectores seleccionados.

Dentro de las distancias que se han seleccionado para el cálculo se encuentra:

La métrica Euclidiana definida como:

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Ecuación 1 Métrica Euclidiana

Donde p y q son puntos del espacio n-dimensional. Y la métrica City Block, también conocida como métrica de Manhattan, es una función de distancia definida como:

$$dist(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

Ecuación 2 Métrica Manhattan

El KNN es uno de los clasificadores más utilizados por su simplicidad. La principal dificultad de este método consiste en determinar el valor de k, ya que si toma un valor grande se corre el riesgo de hacer la clasificación de acuerdo a la mayoría (y no al parecido), y si el valor es pequeño puede haber imprecisión en la clasificación a causa de los pocos datos seleccionados como instancias de comparación. [13]

3.4.2 Árboles de clasificación

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. [14]

Para la representación de cada caso, cada nodo no terminal determina una pregunta de alguna variable. Cada rama corresponde a un posible valor de la variable relacionada con el nodo. Por otro lado, cada nodo terminal indica la clase en la que se clasifica el caso.

La complejidad del árbol crece linealmente con el número de casos y exponencialmente con el número de variables. De lo que se concluye que, puede no ser una buena idea desarrollar el árbol hasta que se clasifiquen correctamente todos los casos.

Una manera para evitar el overfitting es realizando una pre-poda, también conocido como preprunning; se realiza una selección de variables previa basada en un test de independencia entre cada variable predictora X y la variable de la clase C. Otra manera de evitar el sobre ajuste es realizando una post-poda o postprunning donde se desarrolla el árbol completo y después se poda sustituyendo subárboles por hojas. Se basa en test de hipótesis que tratan de responder si merece la pena expandir o no una determinada rama.

3.4.3 Random Forest

Este algoritmo es una técnica de aprendizaje automático, el cual se encuentra compuesto por numerosos árboles de clasificación, con el fin de obtener el resultado más preciso posible.

El algoritmo Random Forest es un método de estimación combinado, donde el resultado de la estimación se construye a partir de los resultados obtenidos mediante el cálculo de n árboles donde los predictores son incluidos al azar. [15]

Gracias a la comparación de los múltiples árboles se evita el overfitting dando resultados más exactos y precisos, y haciendo que se convierta en uno de los algoritmos más precisos. Sin embargo, requiere un mayor tiempo computacional.

3.4.4 Naive Bayes (NB)

Algunas de las características de este algoritmo son su alta velocidad y su elevada precisión. Está basado en la siguiente ecuación:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Ecuación 3 Clasificador bayesiano

Donde:

- P(h|D): la probabilidad de la hipótesis h dados los datos D. Esto se conoce como probabilidad posterior de la hipótesis.
- P(D|h): la probabilidad de los datos d dado que la hipótesis h era verdadera. Esto se conoce como probabilidad posterior de los datos.
- P(h): la probabilidad de que la hipótesis h sea cierta (independientemente de los datos). Esto se conoce como la probabilidad previa de h.
- P(D): la probabilidad de los datos (independientemente de la hipótesis). Esto se conoce como probabilidad previa de los datos.

3.4.5 Redes neuronales (MLP)

Las redes neuronales están inspiradas en el cerebro humano, donde el funcionamiento es el siguiente: la capa de entrada está compuesta por distintas entradas, de manera que cada una está asociada a un peso. Posteriormente se encuentra la capa oculta, formada por varios nodos, llamados neuronas, de modo que para cada una de las entradas está asociada con cada neurona. Finalmente se localiza la capa de salida, obteniendo la predicción calculada por la red.

Se denomina arquitectura a la topología, estructura o patrón de conexionado de una red neuronal. En una red neuronal artificial los nodos se conectan por medio de sinapsis, estando el comportamiento de la red determinado por la estructura de conexiones sinápticas. Estas conexiones sinápticas son direccionales, es decir, la información solamente puede propagarse en un único sentido (desde la neurona presináptica a la postsináptica). En general las neuronas se suelen agrupar en unidades estructurales que denominaremos capas. El conjunto de una o más capas constituye la red neuronal.

Se distinguen tres tipos de capas: de entrada, de salida y ocultas. Una capa de entrada, también denominada sensorial, está compuesta por neuronas que reciben datos o señales procedentes del entorno. Una capa de salida se compone de neuronas que proporcionan la respuesta de la red neuronal. Una capa oculta no tiene una conexión directa con el entorno, es decir, no se conecta directamente ni a órganos sensores ni a efectores. Este tipo de capa oculta proporciona grados de libertad a la red neuronal gracias a los cuales es capaz de representar más fehacientemente determinadas características del entorno que trata de modelar. [16]

3.4.6 Máquinas de soporte vectorial (SVM)

Las máquinas de vectores soporte tienen una fundamentación matemática pura dentro de la teoría estadística de aprendizaje, a pesar de esto, la implementación básica cuenta con algunas limitaciones puesto que están diseñadas originalmente para problemas de clasificación binarios (dos clases), y además tienen como limitante que su algoritmo básico de entrenamiento genera gran cantidad de vectores soporte lo que ocasiona lentitud en la clasificación. Sin embargo, constituyen una poderosa y robusta herramienta destinada a labores de clasificación. La misma ha sido usada en el campo de la neuro informática como se puede ver en.

Este algoritmo se basa en mapear los puntos de entrenamiento a un espacio vectorial de una dimensión mayor, construir hiperplanos en un espacio multidimensional para separar los puntos en sus clases respectivas y clasificar un punto nuevo de acuerdo a su ubicación con respecto al hiperplano de separación. [13]

3.5 Cross-Validation

La validación cruzada es una técnica para evaluar modelos de Machine Learning (ML) mediante el entrenamiento de varios modelos de ML en subconjuntos de los datos de entrada disponibles y evaluarlos con el subconjunto complementario de los datos. La validación cruzada se utiliza para detectar el sobreajuste, es decir, en aquellos casos en los que no se logre generalizar un patrón. Se puede utilizar el método de la validación cruzada de K iteraciones para realizar la validación cruzada. En la validación cruzada de K iteraciones se dividen los datos de entrada en K subconjuntos de datos (también conocido como iteraciones). Puede entrenar un modelo de ML en todos menos uno (k-1) de los subconjuntos y, a continuación, evaluar el modelo en el subconjunto que no se ha utilizado para el entrenamiento. Este proceso se repite K veces, con un subconjunto diferente reservado para la evaluación (y excluido del entrenamiento) cada vez. [17]

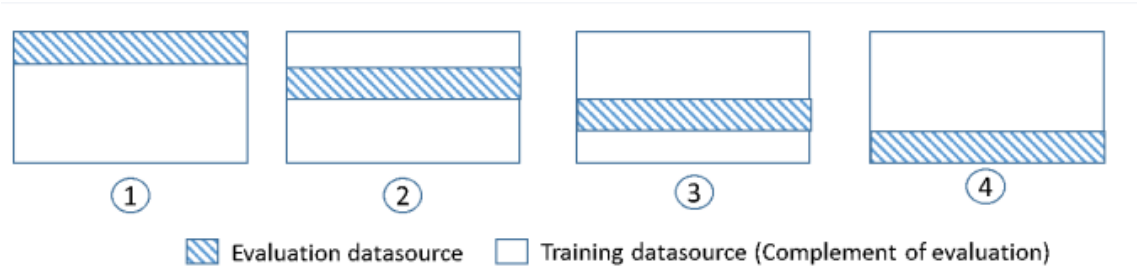

Figura 13 Modelo cross-validation

En la Figura 13 se muestra un ejemplo de los subconjuntos de entrenamiento y de los subconjuntos de evaluación complementarios generados para cada uno de los cuatro modelos que se crean y se entrenan durante una validación cruzada de 4 iteraciones. El modelo uno utiliza el primer 25% de los datos para la evaluación y el 75% restante para el entrenamiento. El modelo

dos utiliza el segundo subconjunto del 25% (del 25% al 50%) para la evaluación y los tres subconjuntos restantes de los datos para el entrenamiento y así sucesivamente. [17]

3.6 Selección de variables (FSS)

La selección de variables o Feature Subset Selecion (FSS) tiene como finalidad detectar el subconjunto óptimo de variables de un conjunto de datos instaurando un determinado método.

Su principal utilización reside en evitar el sobreajuste y en la mejora del rendimiento y la precisión del modelo. El hecho de emplear muchas variables no implica la mejora del clasificador, ya que algunas de ellas pueden llegar a ser irrelevantes y puede darse el caso de producirse redundancia entre las propias variables predictoras. Es por ello que el objetivo es identificar dichas variables.

No obstante, el uso del FSS se debe de realizar con prudencia debido a que existe el riesgo de perder información. Por consiguiente, encontrar los parámetros óptimos del modelo es fundamental. La técnica de selección de vaciables se divide en tres métodos: método *filter*, método *wrapper* y método *embedded*.

El proyecto se ha afrontado mediante el uso del método *filter*, concretamente por medio de la técnica de selección de características basadas en la correlación o Correlation based Feature Selection (*CFS*) explicado su funcionamiento en el apartado selección de variables CFS.

Esta técnica *filter* es computacionalmente rápida y fácil de aplicar para datos de alta dimensionalidad. Consigue que las precisiones obtenidas por el subconjunto de atributos sean mejores que otros métodos mencionados anteriormente.

3.7 Evaluadores en la clasificación

Son distintas las métricas que se emplean para predecir la eficiencia que dispone un modelo. En este apartado se contemplarán las diversas formas existentes.

3.7.1 Matriz de confusión (MC)

La matriz de confusión es un elemento que permite analizar la capacidad discriminante de un clasificador. Divide los casos clasificados en base a la clase observada y a la clase predicha por el clasificador, recogiendo los aciertos y errores de cada clase, e ignorando el desbalanceo entre las clases:

		Predicción	
		Positivos	**Negativos**
Observaciones	**Positivos**	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Tabla 1 Matriz de confusión (CM)

3.7.2 Estadísticos basados en la CM

Para poder evaluar el rendimiento del modelo supervisado es útil la utilización de las siguientes métricas, siendo el objetivo final la medición de la “bondad” o “capacidad discriminante” de los clasificadores, y teniendo en cuenta que el objetivo de estos es obtener valores altos en estas métricas.

Si el resultado se acerca a 0, denota una falta existente de relación. Sin embargo, en caso de que se aproxime más a 1 expresa que el clasificador clasifica correctamente más casos.

3.7.2.1 Accuracy (ACC)

Se utiliza para saber el porcentaje de veces que el modelo ha acertado, es decir, la tasa de acierto. La manera de calcularlo es:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ecuación 4 Accuracy

En caso de querer calcular el error sería tan fácil como:

$$errorAcc = 1 - accuracy$$

Ecuación 5 Error del accuracy

3.7.2.2 Precisión (PRE)

Mide la proporción de casos realmente positivos de entre los que el clasificador predice como positivos, obteniéndola de la siguiente manera:

$$precisión = \frac{TP}{TP + FP}$$

Ecuación 6 Precisión

3.7.2.3 Recall (REC)

Esta métrica sin embargo proporciona la cantidad de casos positivos que fueron correctamente clasificados, siendo la formula:

$$recall = \frac{TP}{TP + FN}$$

Ecuación 7 Recall

3.7.2.4 Specitifity (SPC)

Indica la cantidad de casos negativos que fueron correctamente clasificados. Se logra de la siguiente manera:

$$specitifity = \frac{TN}{TN + FP}$$

Ecuación 8 Specitifity

3.7.2.5 F1-score

Este tipo de métrica es útil en caso de buscar un balance entre la precisión (PRE) y el recall (REC) ya que fusiona las dos métricas:

$$F1 = \frac{2 * PRE * REC}{PR + REC}$$

Ecuación 9 F1

Capítulo 4

Origen de los datos

4.1 Introducción

La fundación *CITA-Alzheimer* realizó 2 estudios neurológicos a un total de 537 pacientes, donde el 55% eran mujeres y el tanto por ciento restante hombres. Teniendo como aspiración incrementar el conocimiento sobre esta enfermedad.

Los datos que aparecen son datos anonimizados, es decir, no se puede identificar la persona a la cual pertenece dicha información.

De los 537 sujetos, 177 no accedieron a realizarse la punción lumbar, por lo que este tipo de pacientes se considerarán como casos sin etiquetar. Se encuentra representado por el valor *NaN* en la columna indicadora la enfermedad. Así mismo, aquellos pacientes en los que aparezca un 1 en dicha columna denotará la posibilidad de estar desarrollando Alzheimer y si figura el valor 0 quiere decir que no se está experimentando la enfermedad.

4.2 Contenido de la base de datos

El estudio realizado por la fundación *cita Alzheimer* recaba un gran número de datos, no obstante, se han cedido únicamente una parte de ellos para poder llevar a cabo la siguiente investigación.

A continuación, se explican las columnas más significativas para la comprensión y utilización de los datos a lo largo del proyecto:

- Columna A: hace referencia al paciente (no se realiza uso de ella).
- Columna B: tipo de estudio realizado: DEBA o PGA (no se realiza uso de ella).
- Columna C: pertenece al género del paciente.
- Columna D: corresponde a la edad del paciente.

Las siguientes 4 columnas están representadas de manera binaria; 0 corresponde a negativo y 1 a positivo. En caso de dar positivo en alguno de los marcadores F, G o H, la columna E pasaría a definirse como positivo, ya que para los doctores representa la posibilidad de estar desarrollando la enfermedad, a causa de que el Alzheimer empieza antes de mostrar los primeros síntomas.

- Columna E: hace referencia al resultado de la punción lumbar.
- Columna F: representa la presencia de la proteína β-amiloide.
- Columna G: simboliza los niveles altos de proteína TAU.
- Columna H: denota la existencia elevada de proteínas fosfo tau.

Las columnas que vienen a continuación representan los datos de las 68 áreas por las que se compone el cerebro. Estos datos se han obtenido a partir de resonancias magnéticas, gracias al software de una empresa especializada con el que pudieron obtener el volumen de distintas zonas cerebrales (resultado que tenemos a partir de la columna I). En total se cuenta con la información de las 68 áreas, divididas en 2 hemisferios[4] simétricos, es decir 34 datos por cada hemisferio. Donde por cada área se dispone de su volumen, representado en cm^3 y desglosado en 3 apartados: materia blanca, materia gris y líquido cefalorraquídeo.

Sin embargo, a pesar de ser simétricos no implica que una enfermedad que afecta a un área implique que afecte a la otra. Es decir, se puede tener el hipocampo izquierdo dañado mientras que el derecho se encuentre en buenas condiciones.

[4] El cerebro consta de 2 hemisferios: el izquierdo, representado por un *L* y el derecho, determinado por una *R* en la base de datos.

De tal forma, un 67% de los pacientes de la base de datos aceptaron realizarse la punción lumbar mientras que el otro 33% no (ver figura 19). De aquellos pacientes que aceptaron someterse a la prueba de la punción lumbar, el 36,638% dieron positivo y el otro tanto por ciento negativo (ver figura 20)

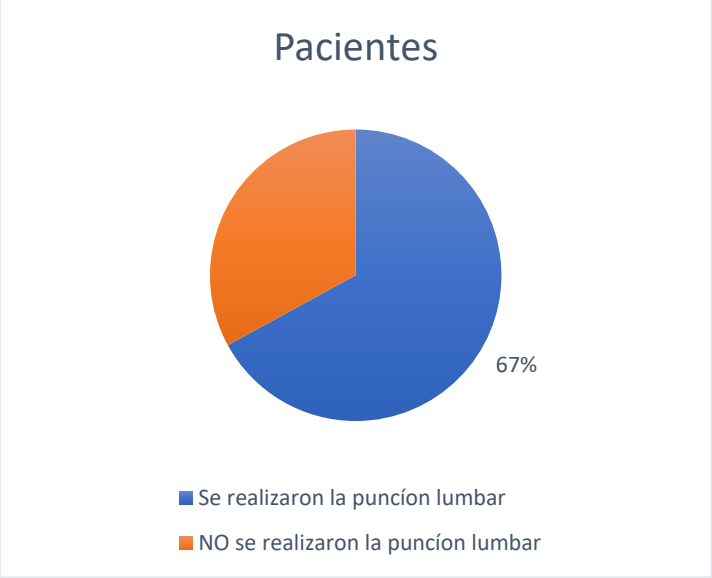

Figura 14 Gráfico que representa que el 67,039% de los pacientes de la base de datos se realizaron la punción lumbar y el 32'931% no.

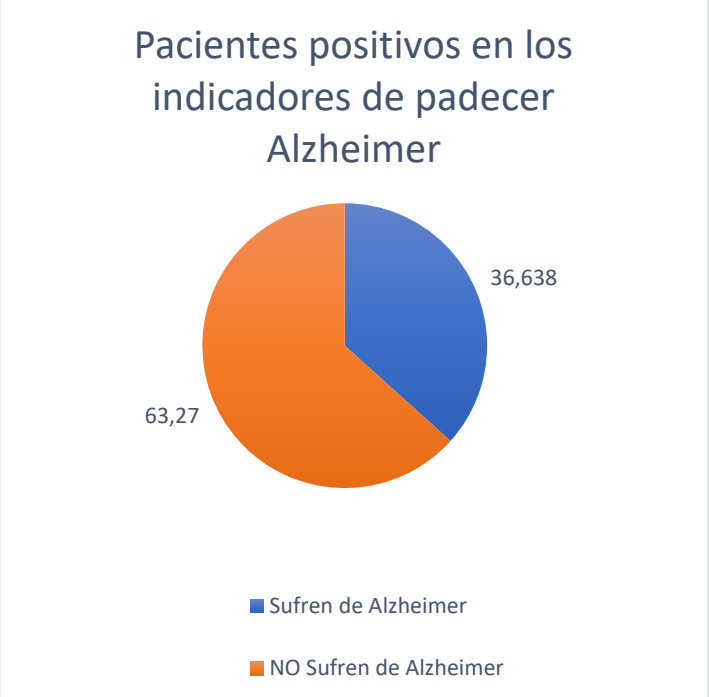

Figura 15 Gráfico que representa que el 36,638% de los pacientes de la base de son positivos en los indicadores de Alzheimer y el 63,27% no.

Por lo que, de los 537 pacientes que conforman la base de datos, 132 de ellos han dado positivo en Alzheimer, 288 han dado negativo y los restantes 177 pacientes no aceptaron realizarse la punción lumbar (ver figura 21).

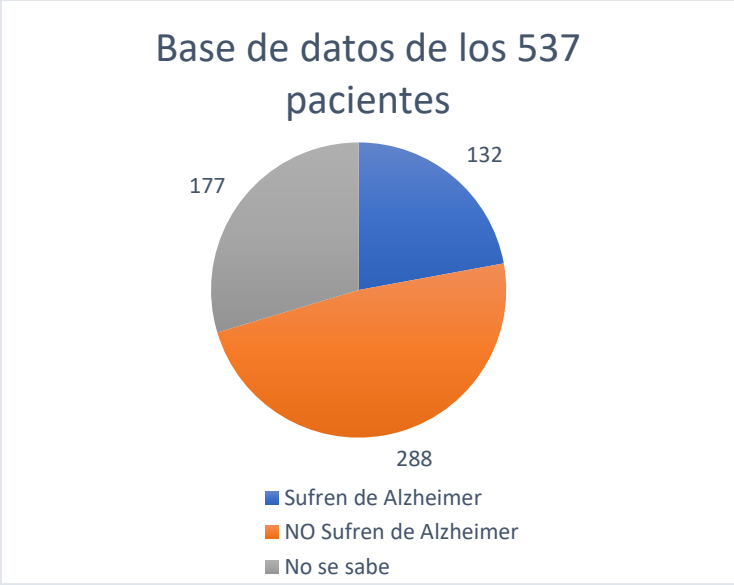

Figura 16 Gráfica representativa de los 537 pacientes de la base de datos, clasificados por pacientes que sufren Alzheimer, los que no y por los que se desconoce su clase.

4.3 Preproceso de los datos

Partiendo de la base de datos original, se ha llevado a cabo una modificación procediendo a la normalización de los valores respecto al volumen del área cerebral (V_a) así como también del volumen intracraneal (V_i, alude volumen del cerebro del sujeto).

De esta manera, se ha obtenido la proporción que ocupa cada área en el cerebro y se ha prevenido la existencia de valores faltantes y datos erróneos.

La operación a realizar es:

$$V_n = (V_a / V_i) \times 100$$

Ecuación 10 Ecuación que simboliza la proporción que ocupa cada área en el cerebro.

Capítulo 5

Desarrollo del trabajo y resultados

5.1 Entorno de trabajo

A lo largo de este proyecto han sido dos los lenguajes de programación los utilizados para la realización del trabajo: *R*[5] y *Python*[6].

5.1.1 R-Studio

Al comienzo del proyecto se optó por el uso del lenguaje de R-Studio por la fácil manipulación que conlleva trabajar con gran cantidad de datos y por lo útiles que son algunas de sus herramientas como los histogramas o boxplot.

En este entorno se ha trabajado en la creación de histogramas y boxplot, así como en una fácil visualización de estos respecto a las variables continuas que conforman el dataset.

Del mismo modo, se ha trabajado en la obtención de las frecuencias sobre las columnas género, indicadora de la enfermedad, β-amiloides, Proteína TAU y Proteína TAU Fosforizada.

En último lugar, con el fin de analizar la correlación entre variables se han creado dos matrices de correlación de Pearson. Con el fin de disponer e interpretar de manera sencilla los datos se han exportado a un documento de Excel y los RPlot a otro fichero.

5.1.2 Python

Una vez realizada la parte descriptiva se ha llevado a cabo un análisis de datos empleando distintos tipos de algoritmos mediante el aprendizaje supervisado y semi-supervisado.

---

[5] https://www.rstudio.com/

[6] https://www.python.org/downloads/

A lo largo del proyecto se ha hecho uso de *scikit-learn*[7], una biblioteca para aprendizaje automático de software libre para el lenguaje de programación *Python*. Incluye varios algoritmos de clasificación, regresión y análisis de grupos entre los cuales están máquinas de vectores de soporte, bosques aleatorios, Gradient boosting, K-means y DBSCAN. Está diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy. [18]

5.2 Implementación y resultados

5.2.1 Análisis descriptivo de los datos

Para este apartado se ha realizado la implementación en el entorno de *R Studio* de los dos primeros apartados, y en *Python* los que vienen a continuación. En él se trata la búsqueda de valores anómalos, el cálculo de frecuencias y el análisis de correlación entre variables.

5.2.2 Búsqueda de valores anómalos

Mediante el uso de histogramas y boxplot se ha llevado a cabo la búsqueda de valores perdidos, erróneos o extremos. En el *Apéndice 1* se puede visualizar de manera gráfica estos valores.

En él se crea por cada variable continua un histograma y boxplot de manera superpuesta, con la intención de realizar una comparación más sencilla de las variables continuas ya que permite ver ambas gráficas a la vez. Por ejemplo, en la siguiente imagen (ver Figura 14) se puede observar la representación gráfica de las edades de los pacientes, donde se aprecia que la edad más habitual oscila entre los 60 y 65 años. No obstante, se contempla que la edad de los pacientes sometidos a las pruebas fluctúan entre los 40 y 90 años.

[7] https://scikit-learn.org/stable/tutorial/index.html

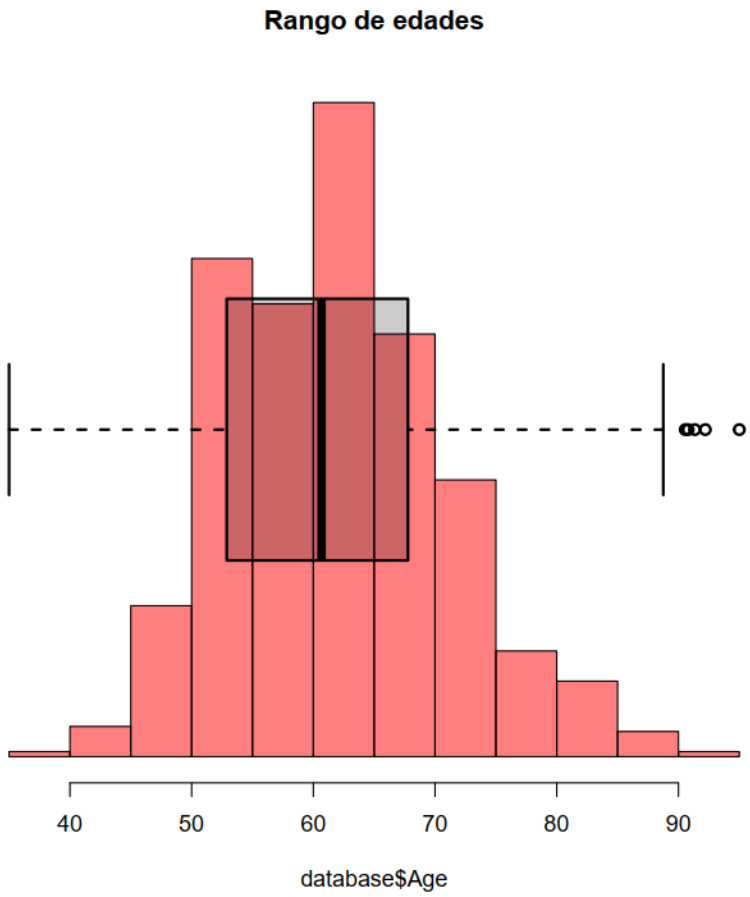

Figura 17 Histograma y boxplot de la edad de los pacientes

Referente a los valores, se puede apreciar que todas las variables siguen una distribución similar a la normal, aunque en algunos casos con colas muy largas. En todo caso, no se aprecia nada extraño y no se cree que sea necesario un preprocesamiento de los datos.

Lo único llamativo es la existencia de una variable cuyos valores son todo ceros (*Accumb_R_CSF*) la cual no aporta información, permitiendo prescindir de ella. La columna *Accumb_L_CSF* tiene

un único valor distinto a cero, así que también se ha decidido eliminarla debido a la escasa información que aporta.

5.2.1.2 Cálculo de las frecuencias

Tras la búsqueda de los valores anómalos, se ha llevado a cabo la obtención de las frecuencias, es decir, de los valores más repetidos de las siguientes columnas:

- Para la columna perteneciente al género, se ha calculado que un 43.20% de los pacientes son mujeres, mientras que el 56.78% restantes son hombres.

Las columnas que vienen a continuación hacen referencia a los resultados de la punción lumbar, donde cabe destacar que el 32.96% de los pacientes no se han realizado la punción lumbar.

- En la siguiente columna, la indicadora de si un paciente sufrirá Alzheimer, se puede observar que el 42.45% de los pacientes no estén desarrollando Alzheimer. Mientras tanto un 24.58% de los pacientes están desarrollando la enfermedad.

- A continuación, encontramos información sobre los β-amiloides, donde se localiza que el 47.11% no experimenta gran cantidad de esta proteína, mientras 19.92% sí.

- La próxima columna es la que aporta un análisis de los niveles de la proteína TAU, un 60.33% de los pacientes no presentan un gran número, mientras que el 6.70% manifiestan unos niveles altos.

- En último lugar, se encuentra la proteína TAU Fosforizada, por un lado, se localiza que el 60.33% de los pacientes no presentan gran cantidad de esta proteína, y por otra parte el 6.70% sí que los presentan.

En la siguiente tabla se puede observar de manera gráfica los resultados mencionados:

Columna	Referencia a	Cantidad de pacientes	Tanto por ciento
Género	Mujeres	232	43.20%
	Hombres	305	56.78%
Indicadora de si un paciente sufrirá Alzheimer	No se esté desarrollando Alzheimer	228	42.45%
	Sí se esté desarrollando Alzheimer	132	24.58%
	Incertidumbre de si se está o no desarrollando Alzheimer	177	32.96%
β-amiloides	No hay una gran presencia	253	47.11%
	Sí hay una gran presencia	107	19.92%
	Personas que no se realizaron la punción lumbar	177	32.96%
Proteína TAU	No hay una gran presencia	324	60.33%
	Sí hay una gran presencia	36	6.70 %
	Personas que no se realizaron la punción lumbar	177	32.96%
Proteína TAU Fosforizada	No hay una gran presencia	324	60.33%
	Sí hay una gran presencia	36	6.70%
	Personas que no se realizaron la punción lumbar	177	32.96%

Tabla 2 Resultado del cálculo de las frecuencias

Estos datos reflejan que de los 367 pacientes que se realizaron la punción lumbar, el 63.27% de los pacientes no padecen de Alzheimer. Por el contrario, el porcentaje de personas que sí sufren de Alzheimer y se realizaron la punción lumbar es del 36.638%.

5.2.1.3 Analizar la correlación entre variables

Con el fin de analizar la correlación entre variables se ha empleado *la Correlación de Pearson:* la matriz de correlación muestra los valores de correlación de Pearson, que miden el grado de relación lineal entre cada par de elementos o variables. Los valores de correlación se pueden ubicar entre -1 y +1. [19]

Se ha hallado la correlación entre las columnas género, edad y proteínas β-amiloides, TAU y fosfo TAU con respecto a estas mismas. Los resultados obtenidos muestran los datos y un Rplot con los mismos, con el fin de reflejar de manera más sencilla la resolución.

La matriz cuadrada que se visualiza a continuación, muestra una correlación significativa entre los marcadores de la enfermedad. La correlación obtenida era esperable pues son los indicadores de un mismo hecho: la enfermedad. No obstante, en un principio se consideraba que los marcadores del género y edad fuesen altos, a pesar de ello, se refleja que no es así y que la correlación más elevada es entre la proteína TAU y la fosfo TAU.

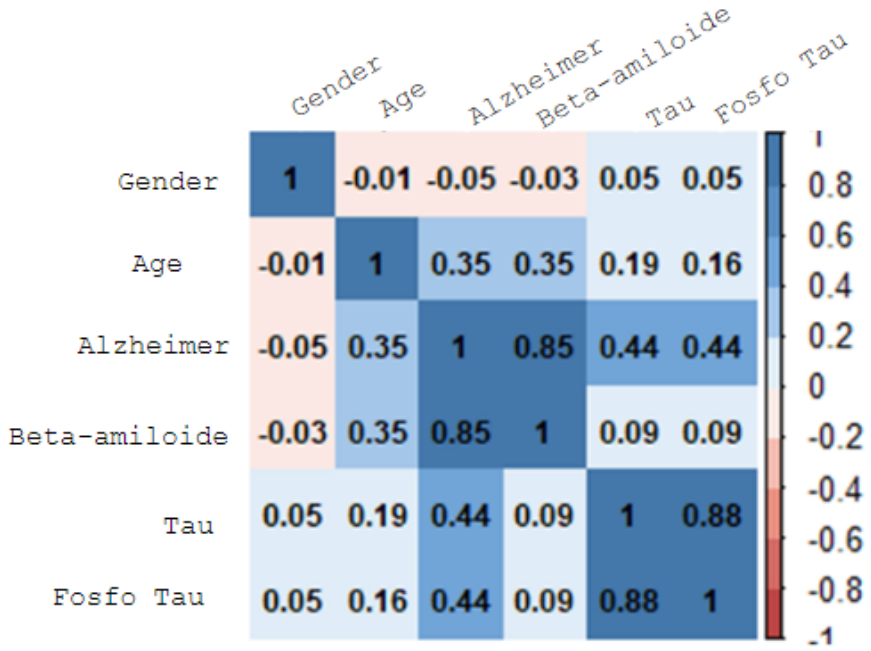

Figura 18 Matriz cuadrada de correlación de Pearson

En segundo lugar, se ha calculado la matriz de correlaciones de Pearson respecto a las columnas anteriores y las columnas restantes del *data frame*: las columnas indicadoras de la enfermedad con respecto a las 68 áreas por las que se compone el cerebro.

En este caso, también se ha creado un archivo Excel con dichos resultados y distintas matrices para su fácil visualización, se puede encontrar en el Apéndice 2. Debido a la gran cantidad de columnas, se han ido creando las matrices de correlación en intervalos de 10, por lo que en el fichero Apéndice 2 se ha obtenido un total 20 matrices.

A continuación, se muestra la correlación entre las columnas género, edad y proteínas β-amiloides, TAU y fosfo TAU con respecto al primer intervalo de 10 de las 208 columnas: materia gris, blanca y líquido cefalorraquídeo del hipocampo izquierdo, derecho, amígdala izquierda y la materia gris de la amígdala derecha.

Figura 19 Matriz de correlación de Pearson respecto a las columnas indicadoras de la enfermedad con respecto a las 68 áreas por las que se compone el cerebro.

A pesar de que haya cierta correlación entre las diferentes variables, ninguna supera el 0.62 de correlación. Como se suponía en un principio, el problema es más complejo y una única variable no resuelve el problema a abordar.

5.2.1.4 Normalizar y corregir valores perdidos

Debido a lo pequeños que son los valores que posee la base de datos, dificulta el trabajo de los clasificadores. Por lo que una solución ha sido normalizar los datos antes de usarlos para entrenar a los clasificadores ya que, como se ha comprobado, surgirían problemas en caso de no hacerlo.

La técnica de la que se ha dispuesto para ajustar los valores es la estandarización. Consiste en restar a cada valor la media de su columna y luego dividirla por la desviación estándar de su columna. Una manera más visual de percibirla es:

$$x = \frac{(x - x.mean)}{x.std}$$

Ecuación 11 Normalización

En la imagen que se muestra a continuación se presentan los datos antes y después de haberlos normalizado. En la gráfica de los datos originales se aprecia ciertamente la desigualdad entre los valores y tras su normalización, se transforma en una escala más uniforme.

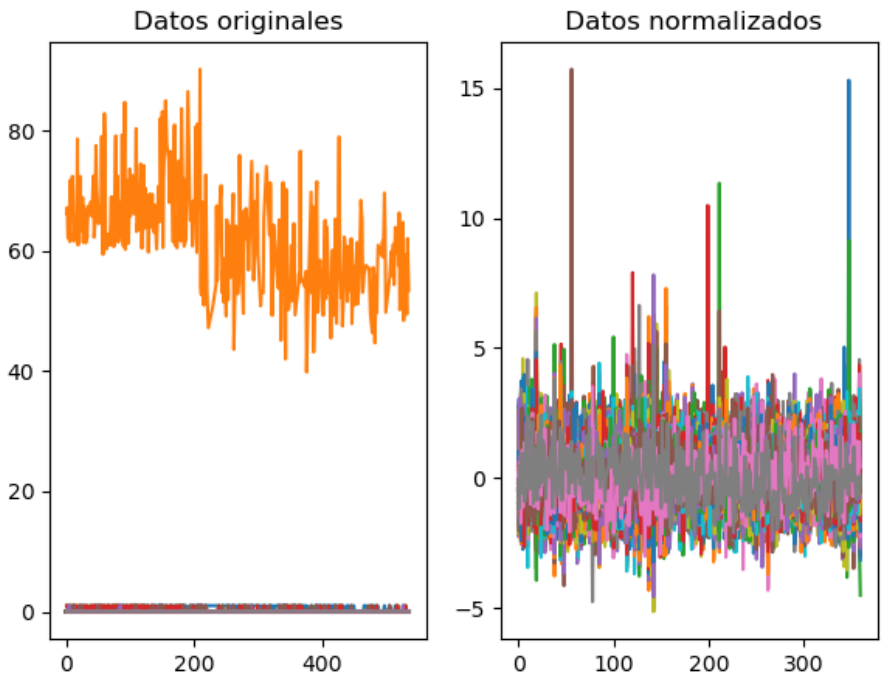

Figura 20 Gráfica comparativa entre los valores antes y después de haberlos normalizado.

Como se verá en el apartado Clase indicadora de si un paciente sufrirá Alzheimer también se ha realizado una normalización para las 12 variables seleccionadas por la selección de variables CFS cuando la clase a predecir es el resultado de la punción lumbar.

Para las 2 normalizaciones realizadas; todos los datos y la lista de las 12 variables, se han creado dos Excel en donde se pueden visualizar los nuevos datos normalizados.

A su vez, se han sustraído algunos detalles estadísticos de las 12 variables seleccionas. Por cada columna se ha calculado la media, la desviación estándar, el máximo y el mínimo.

5.3 Selección de variables CFS

El siguiente paso es utilizar una técnica de selección de variables, para ver qué variables, es decir, volúmenes cerebrales, son los más adecuados para construir clasificadores. Para ello, se ha hecho uso del código de la librería *Correlation-based-Feature-Selection* [8] .

En un modelo puede darse el caso de que existan muchos datos innecesarios, es decir, que no proporcionen información y consecuentemente hace que el modelo se ralentice. A su vez puede ocurrir que, al aprender de estos datos irrelevantes, sea más inexacto. Es por esto, por lo que hay que seleccionar las variables relevantes y deshacerse de las que no lo son.

Usando la selección de variables se puede optimizar el modelo de varias formas:

1. Evitar el aprendizaje del ruido y el sobre ajuste: evitando así una predicción abstracta.
2. Mayor precisión (accuracy): aparte de querer obtener una predicción concreta queremos que esté lo más cerca de la respuesta correcta
3. Reducir el tiempo de entrenamiento: en la gran mayoría de modelos supone un crecimiento exponencial. [20]

La finalidad es la siguiente, siendo un proceso incremental primero se ha de buscar la variable que esté más correlada con la clase que se quiere predecir, ya que cuanto más correlada esté, más probabilidad hay de predecir correctamente la clase. En consecuencia, se selecciona y se prueba una a una con el resto de columnas; cuál es el mejor resultado combinando las dos variables (en términos de correlación). Dicho de otra manera, el objetivo es intentar maximizar

[8] https://github.com/ZixiaoShen/Correlation-based-Feature-Selection

la correlación que tienen las variables con la clase, pero luego intentar minimizar la correlación entre las que se comportan exactamente igual o de manera muy parecida.

Para eso, el algoritmo estima el mérito de un subconjunto scon k características con la siguiente ecuación [21]:

$$Merito_s = \frac{k\ r_{cf}}{\sqrt{k + k\ (k-1)\ r_{ff}}}$$

Ecuación 12 Fórmula para la selección de variables CFS

Donde *k* indica el número de características de ese subconjunto, r_{ff} la correlación media entre características promedio y r_{cf} la correlación media entre las características de la clase.

Han sido 2 las ocasiones en las que se ha hecho uso de esta librería. En la primera, como bien se ha mencionado hasta ahora, ha sido para predecir la clase cuando la columna es la que muestra si el paciente sufre de Alzheimer, para este caso las variables que están más correlacionadas entre ellas han sido 12. El siguiente caso donde se ha hecho uso de esta implementación ha sido para predecir la clase siendo la columna la indicadora de que el paciente padezca niveles altos de la proteína beta-amiloide, en el cual se ha obtenido una lista de 10 variables seleccionadas.

Como se aprecia, la reducción de las variables es drástica pues de 219 columnas únicamente se hacen uso de 12 o 10. A pesar de lo cual, no implica que el resultado se deteriore, al contrario, se optimiza.

5.4 Algoritmos de aprendizaje supervisado

A continuación, se tratarán los códigos que se han usado para su resolución mediante aprendizaje automático, es decir, únicamente se han usado aquellos datos de los que se conocen la clase. Para ello se ha hecho diferentes pruebas con distintas clases.

5.4.1 Clase indicadora de si un paciente sufrirá Alzheimer

Inmediatamente a continuación de haber importado el dataset, se prescinden de las filas donde aparezca el valor *NaN* en la columna indicadora de si un paciente sufrirá Alzheimer ya que el valor de *NaN* representa el desconocimiento de la etiqueta.

Resaltar el uso de la técnica de 10-fold cross-validation donde se han dividido los datos en 10 pliegues, los cuales 9 se han utilizado para entrenar mientras que el restante se ha usado para realizar la prueba. Generada la estimación se ha calculado la puntuación media y la desviación estándar para las métricas de clasificación accuracy, precisión, recall y f1.

A la hora de evaluar los resultados se ha obtenido los siguientes:

	Precisión en Train	**Accuracy**		**Precision**		**Recall**		**F1**	
		Media	σ	Media	σ	Media	σ	Media	σ
Naive Bayes	0.64	0.61	0.10	0.46	0.19	0.33	0.10	0.53	0.07
KNN	0.73	0.62	0.11	0.46	0.18	0.35	0.09	0.54	0.07
SVM	0.75	**0.67**	0.13	0.72	0.26	0.22	0.07	0.55	0.08
Multi-layer Perceptron	1.0	0.56	0.09	0.40	0.17	0.45	0.11	0.51	0.07
Random Forest	1.0	0.66	0.10	0.58	0.25	0.35	0.12	0.58	0.08
Decision Tree	1.0	0.60	0.11	0.47	0.18	0.49	0.16	0.55	0.09

Tabla 3 Resultados aplicando el aprendizaje automático (siendo la clase indicadora de si un paciente sufrirá Alzheimer y las 12 variables seleccionadas por CFS)

Los resultados obtenidos no son favorables, pues hay que tener presente que un clasificador que clasifique a todos como negativos, esto es no enfermos, obtendrá una tasa de acierto de 63,3% (la proporción de negativos en la base de datos) y según muestran los resultados en el mejor de los casos se obtiene un 67,5%, que no es mucho más. Se tiene en cuenta de que la tasa de acierto no es el único criterio, sin embargo, deja claro que el resultado no es bueno.

En vista de que el procedimiento anterior no ha tenido éxito, se ha decidido examinar qué tal lo hacen los clasificadores usando todas las variables y no solo las seleccionadas por el feature selection.

La mayoría de los algoritmos se comportan mal con tantas variables, no obstante, en determinadas ocasiones otros logran buenos resultados. De este modo se ha procurado afrontar el problema desde otra perspectiva.

	Precisión en Train	**Accuracy**		**Precision**		**Recall**		**F1**	
		Media	σ	Media	σ	Media	σ	Media	σ
Naive Bayes	0.66	0.64	0.11	0.48	0.15	0.39	0.14	0.56	0.08
KNN	0.76	0.64	0.11	0.43	0.19	0.26	0.15	0.51	0.05
SVM	0.88	0.66	0.12	0.59	0.38	0.22	0.14	0.52	0.06
Multi-layer Perceptron	1.0	0.62	0.08	0.39	0.22	0.38	0.24	0.52	0.08
Random Forest	1.0	0.66	0.14	0.63	0.28	0.12	0.19	0.53	0.06
Decision Tree	1.0	0.57	0.12	0.39	0.17	0.40	0.14	0.50	0.09

Tabla 4 Resultados aplicando el aprendizaje automático
(siendo la clase indicadora de si un paciente sufrirá Alzheimer y todas las variables)

A pesar de haber no haber obtenido una mejora en los resultados, puesto que se obtiene un 66,11% en el mejor de los casos, en algunos algoritmos se llega a apreciar un leve incremento de la precisión, en particular de Naive Bayes, KNN y Multi-layer perceptrón.

5.4.2 Clase indicadora de si el paciente es positivo la proteína beta-amiloide

En este punto se ha intentado detectar aquellos sujetos asociados al indicador de niveles altos de la proteína beta-amiloide. Existe cierta similitud entre esta columna y entre la indicadora de si el sujeto parece Alzheimer, sin embargo, no son análogas por lo que podría ser una manera de adquirir mejores resultados.

Tras realizar de nuevo la selección de variables se ha obtenido una nueva lista con las variables más correlacionadas entre ellas: el líquido cefalorraquídeo del hemisferio derecho, materia gris del hemisferio izquierdo y derecho, materia blanca del hemisferio izquierdo y derecho, líquido cefalorraquídeo de la parte anterior del giro temporal medio del hemisferio derecho, líquido cefalorraquídeo del Palladium del hemisferio derecho del cerebro, materia blanca del tercer ventrículo del hemisferio derecho del cerebro, líquido cefalorraquídeo del giro recto del hemisferio derecho y líquido cefalorraquídeo del Palladium del hemisferio izquierdo.

Tras ajustar el modelo en el conjunto de entrenamiento y aplicar cross-validation con 10 splits, para cada uno de los clasificadores se obtienen los siguientes resultados:

	Precisión en Train	**Accuracy**		**Precision**		**Recall**		**F1**	
		Media	σ	Media	σ	Media	σ	Media	σ
Naive Bayes	0.72	0.68	0.15	0.44	0.14	0.35	0.16	0.56	0.10
KNN	0.79	0.72	0.14	0.53	0.16	0.33	0.11	0.59	0.08
SVM	0.79	0.73	0.15	0.37	0.22	0.23	0.15	0.57	0.11
Multi-layer Perceptron	1.0	0.63	0.07	0.19	0.17	0.40	0.14	0.54	0.05
Random Forest	0.99	0.70	0.13	0.48	0.19	0.29	0.19	0.55	0.08
Decision Tree	1.0	0.64	0.12	0.41	0.22	0.40	0.20	0.54	0.09

Tabla 5 Resultados aplicando el aprendizaje automático (siendo la clase indicadora de si el paciente es positivo niveles altos de la proteína beta-amiloide. y las 10 variables seleccionadas por CFS)

Al parecer los resultados obtenidos son similares, ha habido un incremento escaso del rendimiento hasta alcanzar el valor de 0.73. Mas haber hecho uso de esta clase sirve para cerciorarse de que este camino tampoco llegará a ningún destino, pues el 63,27% de los pacientes que aceptaron realizarse la punción lumbar dio negativo.

A pesar de obtener una leve mejora del resultado, a lo largo del proyecto se ha mantenido la clase indicadora de si un paciente sufrirá Alzheimer, ya que es el verdadero indicador de la existencia de la enfermedad.

Es por ello que puede darse el caso en el que la columna indicadora si el paciente es positivo en la proteína beta-amiloide sea negativa, mientras que cualquier otro de los indicadores; tau o fosfo-tau sea positivo. En cuyo caso, el indicador que hace referencia a si un paciente sufrirá Alzheimer daría positivo, mientras que la columna indicadora de si el paciente es positivo en la proteína beta-amiloide negativo.

5.5 Algoritmos de aprendizaje semi-supervisado

Al haber estado trabajando hasta ahora en un escenario supervisado se han estado ignorando los casos en los que no se disponían de las etiquetas. Como bien se ha explicado anteriormente, en el aprendizaje semi-supervisado cuenta con casos etiquetados y no etiquetados. Es por ello que en este momento se contará con los casos no etiquetados que antes se omitían. Estos serán los que dispongan del valor *NaN*, por lo que ahora se trabajará con todos los datos de esta columna: los etiquetados, aquellos que sean 0 o 1, y los no etiquetados, los *NaN*.

Conforme a la resolución, ha variado respecto al aprendizaje supervisado pues ya no se emplea de 10-fold cross validation ya que, al disponer de casos sin etiquetar, estos no se pueden emplear para la fase de test. Por este motivo, se ha cambiado el sistema de validación y se ha

usado el método *hold-out* con 30 iteraciones. En cada iteración se han usado el 30% de los casos etiquetados para test, y el restante, esto es los etiquetados y los no etiquetados, para entrenar.

Para este apartado se han utilizado los algoritmos de la librería scikit-learn: *LabelPropagation* y *SelfTrainingClassifier*.

5.5.1 Algoritmo LabelPropagation

Consecutivamente, en vez de usar los algoritmos NB, SVM, Random Forest… el algoritmo a usar ha sido *LabelPropagation*.

Inicialmente se ha dividido la base de datos en dos grupos: el etiquetado; formado por las etiquetas 0 y 1, y el no etiquetado; compuesto por la etiqueta -1, anteriormente representado por el valor *NaN*.

Del grupo etiquetado por las clases 0 y 1 de la anterior fragmentación se ha destinado el 70% de los casos para el entrenamiento del algoritmo, simultáneamente el 30% restante se ha empleado para realizar el testeo. Este proceso se itera un número de 30 veces, en el que por cada una de ellas se juntan los datos del 70% de los casos que se acaban de etiquetar junto con el dataset no etiquetado. Se dispone de un total de 429 casos.

El procedimiento que se ha llevado a cabo se detalla gráficamente mediante el siguiente esquema. Para ello se ha hecho uso de la aplicación en línea de *MindMeister*[9]:

[9] https://www.mindmeister.com

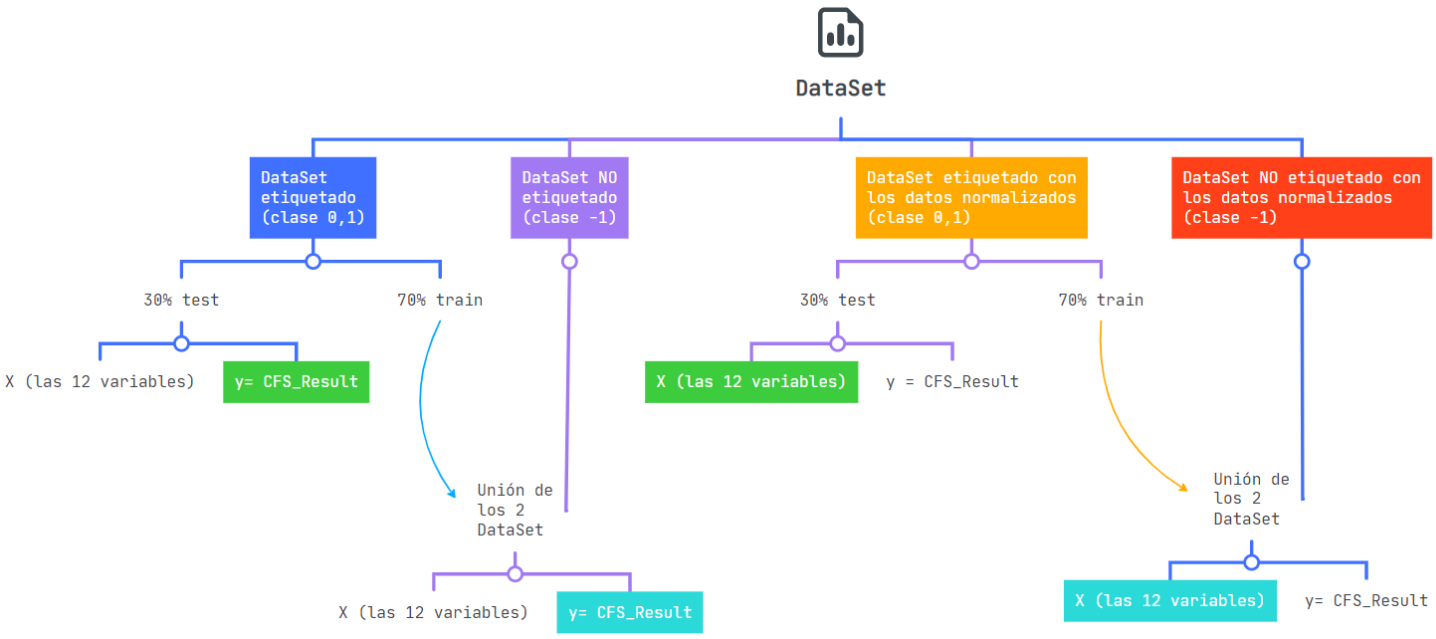

Figura 21 Esquema de la implementación de LabelPropagation

Tras la implementación descrita anteriormente, se ha obtenido en una hoja de Excel las métricas de clasificación y a su vez el cálculo de la media y la desviación estándar de los resultados obtenidos en las 30 repeticiones anteriores. Conjuntamente se ha calculado la media y la desviación estándar por cada una de las filas. En consecuencia, se ha obtenido:

	Media	**Desviación Estándar**
ACC mean on test	0,57	0,04
Precision mean on test	0,20	0,08
Recall mean on test	0,36	0,09
F1 mean on test	0,25	0,076

Tabla 6 Resultados de la media y la desviación estándar de la implementación del algoritmo LabelPropagation

5.5.2 Algoritmo SelfTrainingClassifier

Otro método para trabajar en un escenario semi-supervisado es mediante SelfTrainingClassifier, algo más complejo que el empleado en el anterior apartado.

La implementación que se ha llevado a cabo es muy similar con respecto a la vista en la *Figura 18*, con la diferencia de que en este momento se puede crear una instancia del clasificador supervisado para el proceso de aprendizaje, como SVM, KNN, RandomForest, Decision Tree... Pese a lo cual, no vale cualquier tipo de clasificador pues debe devolver la probabilidad de pertenencia a cada clase y se ha de pasar como el parámetro *base_estimator* al crear la instancia de *SelfTrainingClassifier*.

Tras la implementación llevada a cabo por el uso de *SelfTrainingClassifier*, se ha obtenido en una hoja de Excel las métricas de clasificación por cada instancia del clasificador. Del mismo modo se ha conseguido la media y la desviación estándar de los resultados obtenidos en las 30 repeticiones anteriores:

Naive Bayes		
	Media	**Desviación Estándar**
ACC mean on test	0,5367	0,0972
Precision mean on test	0,3657	0,3361
Recall mean on test	0,3895	0,1074
F1 mean on test	0,2920	0,1719

Tabla 7 Resultados del clasificador Gaussian de la media y la desviación estándar de la implementación del algoritmo SelfTrainingClassifier

KNN		
	Media	**Desviación Estándar**
ACC mean on test	0,5842	0,0478
Precision mean on test	0,1750	0,0861
Recall mean on test	0,3886	0,1378
F1 mean on test	0,2289	0,0864

Tabla 8 Resultados del clasificador KNN de la media y la desviación estándar de la implementación del algoritmo SelfTrainingClassifier

SVM		
	Media	**Desviación Estándar**
ACC mean on test	0,6246	0,0375
Precision mean on test	0,0177	0,0243
Recall mean on test	0,2094	0,3210
F1 mean on test	0,0316	0,0430

Tabla 9 Resultados del clasificador SVC de la media y la desviación estándar de la implementación del algoritmo SelfTrainingClassifier

Random Forest		
	Media	**Desviación Estándar**
ACC mean on test	0,5882	0,0374
Precision mean on test	0,1474	0,0812
Recall mean on test	0,3518	0,1282
F1 mean on test	0,2009	0,0923

Tabla 10 Resultados del clasificador Random Forest de la media y la desviación estándar de la implementación del algoritmo SelfTrainingClassifier

Decision Tree		
	Media	**Desviación Estándar**
ACC mean on test	0,5379	0,0452
Precision mean on test	0,3871	0,0806
Recall mean on test	0,3650	0,0711
F1 mean on test	0,3734	0,0682

Tabla 11 Resultados del clasificador Decision Tree de la media y la desviación estándar de la implementación del algoritmo SelfTrainingClassifier

Capítulo 6

Conclusiones y líneas de trabajo abiertas

6.1 Conclusiones

El objetivo principal de este proyecto es estimar de forma fiable la posibilidad de padecer Alzheimer mediante resonancias magnéticas, con el fin de evitar la práctica de la punción lumbar.

Partiendo de volúmenes cerebrales obtenidos por resonancia, se han utilizado diversos algoritmos de aprendizaje supervisado y semi-supervisado. Los mejores resultados se han obtenido mediante el aprendizaje supervisado, a partir del algoritmo SVM, alcanzando un nivel de accuracy del 67.50% en relación con la clase indicadora de si el sujeto es positivo en los indicadores de Alzheimer.

No obstante, no se considerada un buen resultado. Se sabe con certeza que el 63,27% de los pacientes de la base de datos no sufren de Alzheimer, y según el mejor resultado obtenido indica que en el mejor de los casos se obtiene un 67.50% de acierto.

Llevando a cabo el propósito fijado al comienzo del proyecto, se determina que no se han logrado los resultados esperados. Tras obtener porcentajes de clasificaciones bajas, usando tanto el modelo supervisado como el semi-supervisado, se llega a la conclusión de que poder predecir la enfermedad de Alzheimer a día de hoy mediante volúmenes obtenidos desde imágenes de resonancias magnéticas no es un método fiable, pues aún abarca mucho margen de mejora.

Se ha considerado que, con un método diferente de procesado de imágenes se podrían obtener resultados más favorables, pues se podrían lograr imágenes de mayor calidad facilitando de esta manera la búsqueda de información. Se ha llegado a la conclusión de que no es factible utilizar directamente los valores de los volúmenes cerebrales.

Una vez finalizado el proyecto se puede determinar que a pesar de que las punciones lumbares sean más invasivas, también son más fiables. En cuanto a esta enfermedad respecta, detectarla a tiempo es imprescindible, pues la muerte neuronal es un asunto de alta gravedad que hay que solventar y remediar antes de que vaya a más. Dado que a día de hoy no existe cura para la pérdida de memoria que produce y la única manera de ponerle solución es frenándola o moderando su aceleración, es imprescindible detectarla lo antes posible.

6.2. Líneas de trabajo abiertas

El trabajo realizado en este proyecto deja accesible diversas puertas, con la finalidad de continuar investigando en la detección de Alzheimer.

El intento de mejorar el clasificador para que en un futuro se pueda detectar el Alzheimer sin necesidad de realizar punciones lumbares, es una futura línea de trabajo abierta. Los algoritmos de clasificación disponen de numerosos parámetros, y como a lo largo del proyecto se han usado los parámetros por defecto, en un futuro se podrían configurar de diferentes maneras.

Se considera interesante la posibilidad de utilizar técnicas de extracción de características directamente de las imágenes de resonancia o realizar un preprocesado más inteligente a los datos. Como por ejemplo usando *PCA* que es un método de reducción de dimensionalidad que permite simplificar la complejidad de espacios con múltiples dimensiones a la vez que conserva su información. [22] O también haciendo uso de *autoencoder* donde la entrada es codificada por la red para centrarse solo en la característica más crítica. [23]

Por otro lado, la eficiencia ha dejado bastante que desear, cabe la posibilidad de que con una base de datos en la que haya más sujetos se puedan conseguir mejores resultados. En particular

una opción sería obtener una gran cantidad de datos no etiquetados de otras bases de datos y tratar de mejorar los resultados usando las técnicas semi-supervisadas.

Apéndices

Bibliografía

[1] P. Fuentes, «Alzheimer´s disease: A historical note,» *Revista chilena de neuro-psiquiatría,* vol. Suppl2, nº 9-12, p. 41, 2003.

[2] R. E. M. J. Flint Beal M, «Enfermedad de Alzheimer y demencias afines,» de *Harrison TR. Principios de medicina interna. 14ª edición*, México, Editorial Interamericana Mc Graw-Hill, 1998, pp. vol. II: 2613-2616.

[3] B. B. J., « Enfermedades degenerativas del sistema nervioso,» de *Demencias. Enfermedad de Alzheimer*, Madrid,España, Elseiver, 2006, pp. vol II:1486-1489.

[4] Z. Shen, «GitHub,» Correlation-based-Feature-Selection, 11 August 2019. [En línea]. Available: https://github.com/ZixiaoShen/Correlation-based-Feature-Selection. [Último acceso: 27 Mayo 2022].

[5] E. Dzul y I. d. J. Uscanga Uscanga, «El cerebro en el tiempo (Recorrido de la neurociencia),» 2 junio 2022. [En línea]. Available: https://www.uv.mx/cienciauv/blog/cerebroeneltiemponeurociencia/. [Último acceso: 7 junio 2022].

[6] «Ambientech: Ciencias, Salud y Medio ambiente. Educación Secundaria,» ¿Qué es la neurona? - Glosario de ciencias | Ambientech, [En línea]. Available: https://ambientech.org/la-neurona. [Último acceso: 16 Mayo 2022].

[7] A. Triglia, «Psicología y Mente,» Materia gris del cerebro: estructura y funciones, 4 diciembre 2016. [En línea]. Available: https://psicologiaymente.com/neurociencias/materia-gris-cerebro. [Último acceso: 16 mayo 2022].

[8] «ALZHEIMER'S ASSOCIATION,» ¿Qué es el Alzheimer?, [En línea]. Available: https://www.alz.org/alzheimer-demencia/que-es-la-enfermedad-de-alzheimer.

[9] M. Clinic, «Mayo Clinic,» Enfermedad del Alzheimer, 19 Feb 2022. [En línea]. Available: https://www.mayoclinic.org/es-es/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447.

[10] K. D, «Review provided,» Enfermedad de Alzheimer, mayo 2006. [En línea]. Available: http://www.nlm.nih.gov/medlineplus/spanish/ency/article.

[11] M. Fidel Romano, M. D. Nissen, N. M. Del Huerto Paredes y D. C. A. Parquet, «Enfermedad de alzheimer.,» *Revista de posgrado de la vía cátedra de medicina,* nº 9-12, p. 75, 2007.

[12] C. d. TechTarget, «ComputerWeekly.es,» ¿Qué es Aprendizaje automático (machine learning)? - Definición en WhatIs.com, 4 enero 2017. [En línea]. Available: https://www.computerweekly.com/es/definicion/Aprendizaje-automatico-machine-learning#:~:text=El%20aprendizaje%20autom%C3%A1tico%20es%20un,se%20exponen%20a%20nuevos%20datos.. [Último acceso: 9 junio 2022].

[13] M. P. H. P. Delgado CD, « Algoritmos de aprendizaje automático para la clasificación de neuronas piramidales afectadas por el envejecimiento.,» *Revista Cubana de Informática Médica,* vol. 8, nº 3, pp. 559-571, 2016.

[14] M. Abdelmalik , I. Inza y P. Larrañaga, «DocPlayer. Tema 8: Árboles de clasificación. Departamento de Ciencias de la Computacı ́on e Inteligencia ArtificialUniversidad del Pa ́ıs Vasco,» [En línea]. Available: https://docplayer.es/50101310-Tema-8-arboles-de-clasificacion.html.

[15] S. R. Sánchez, «UOC - Master BI - Business Analytics,» Algoritmos de clasificación, Octubre 2016. [En línea]. Available: https://rstudio-pubs-static.s3.amazonaws.com/237547_0171c04b6d2e4550aea58853c056d29d.html#proceso-de-clasificacion-mediante-arboles-de-decision-multimples-paquete-randomforest.

[16] M. Abdelmalik, I. Inza y P. Larrañaga, «Tema 8. redes neuronales. Redes Neuronales, U. del P. Vasco, 12, 17.,» 1997. [En línea]. Available: https://www.researchgate.net/profile/Pedro-Larranaga/publication/268291232_Tema_8_Redes_Neuronales/links/55b7b5c408ae9289a08c0c68/Tema-8-Redes-Neuronales.pdf.

[17] «Amazon Machine Learning: Guía para desarrolladores,» [En línea]. Available: https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf#cross-validation.

[18] Ivanbenetanco43, «Wikipedia,» Scikit-learn, 8 Noviembre 2020 . [En línea]. Available: https://es.wikipedia.org/wiki/Scikit-learn.

[19] «minitab18,» Interpretar todos los estadísticos y gráficas para Análisis de elementos, [En línea]. Available: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/item-analysis/interpret-the-results/all-statistics-and-

graphs/#:~:text=La%20matriz%20de%20correlaci%C3%B3n%20muestra,lo%20general%20tienen%20correlaciones.

[20] Simplelilearn, Dirección, *Feature Selection In Machine Learning.* [Película]. YouTube.

[21] J. S.Fischer, Correlation-based Feature Selection in Python from Scratch, 6 Aug 2021. [En línea]. Available: https://johfischer.com/2021/08/06/correlation-based-feature-selection-in-python-from-scratch/.

[22] O. Simeone, «A Very Brief Introduction to Machine Learning,» 5 Nov 2018. [En línea]. Available: https://arxiv.org/pdf/1808.02342.pdf. [Último acceso: 16 May 2022].

[23] R. B. A. T. M. &. C. J. D. González, «Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial,» *Revista Iberoamericana de Automática e Informática industrial,* vol. 14, nº 4, pp. 394-405, 2017.

[24] F. Lopera, «Enfermedad de Alzheimer,» *Perspectivas en Nutrición Humana,* pp. 29-32, 29 ene 2004.