Clasificación de la evolución de individuos consumidores de cannabis o sanos mediante *Machine Learning*

2 de enero de 2018

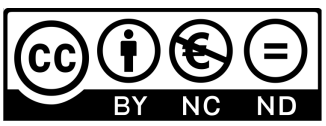
Miriam Mota Foix

Máster en Bioinformática y Bioestadística

Area: Estadística y Bioinformática

Dirigido por: Alex Sánchez Pla

.



FICHA DEL TRABAJO FINAL	
Título de trabajo:	Clasificación de la evolución de individuos consumidores de cannabis o sanos mediante *Machine Learning*
Nombre de la autora:	Miriam Mota Foix
Nombre del consultor:	Alex Sánchez Pla
Nombre del PRA:	Alex Sánchez Pla
Fecha de entrega:	02/01/2018
Titulación:	Máster universitario en Bioinformática y Bioestadística
Área del trabajo final:	Estadística y Bioinformática
Idioma del trabajo:	Castellano
Palabras clave	"machine learning", MRI, clasificación

Resumen del Trabajo:

El cannabis es la droga blanda más utilizada a nivel mundial y hay discusión acerca de que su consumo regular puede afectar la salud en concreto mediante cambios en el cerebro. Este estudio realiza i) una exploración de las distintas técnicas de preprocesado y métodos de análisis con imágenes de resonancia magnética y ii) una aplicación para clasificar los participantes de un estudio en que intentaba separar los individuos en consumidores o no de cannabis a partir este tipo de datos.

Abstract:

Cannabis is the most used soft drug globally, and there is discussion about how its regular consumption can affect health, particularly through changes in the brain. This study performs i) an exploration of the various preprocessing and analysis techniques from magnetic resonance images and ii)an application to classify a study participants that was attempting to assign individuals into cannabis consumers and non-consumers

Índice

Índice de figuras

Agradecimientos:

Me gustaría agradecer a mi tutor,
por haberme motivado y guiado en este trabajo.
A Google por estar 24h
disponible para todo tipo de dudas y
a compañeros y amigos por estar dispuestos
a ayudar en todo momento.

Comentario previo. Este trabajo partía de un estudio predictivo de desarrollo de esclerosis múltiple mediante una serie de biomarcadores genéticos (Nicolas Fissolo [11]). La idea era trabajar con las imágenes de resonancia magnética de dichos pacientes y aplicar técnicas Deep Learning (con la aceptación de los investigadores) para ver si mejoraban las predicciones. En el momento de iniciar el trabajo se nos negó el acceso a las imágenes (por parte del servicio) y se tuvo que redefinir el mismo. Se encontró un nuevo estudio disponible en la plataforma *openFMRI*. Para este estudio, finalmente, no se han aplicado técnicas Deep Learning para la clasificación debido a que una gran parte del trabajo ha consistido en el pre-procesado de los datos además de que el número de muestras disponible era muy bajo.

1. Introducción

El cannabis, también conocido como marihuana es la droga ilícita más utilizada a nivel mundial [8]. El cannabis en su estado fresco contiene ácido tetrahidrocannabinólico, el cual luego se convierte en THC. El compuesto químico psicoactivo predominante en el cannabis es el tetrahidrocannabinol (THC). El cannabis contiene más de 500 compuestos químicos diferentes, entre ellos al menos 113 cannabinoides.

La legalidad del uso del cannabis es muy distinta según el país. En la Figura 1 se muestran las regiones donde, actualmente, es legal el consumo de dicha droga, además de California que se ha legalizado su consumo a partir del 1 de enero de 2018. Prácticamente todos los países tienen leyes que conciernen al cultivo, posesión, venta y consumo de cannabis. El código penal español no prohíbe el consumo del cannabis. Sin embargo, el consumo está restringido a lugares privados, por lo tanto no está permitido hacerlo en público.

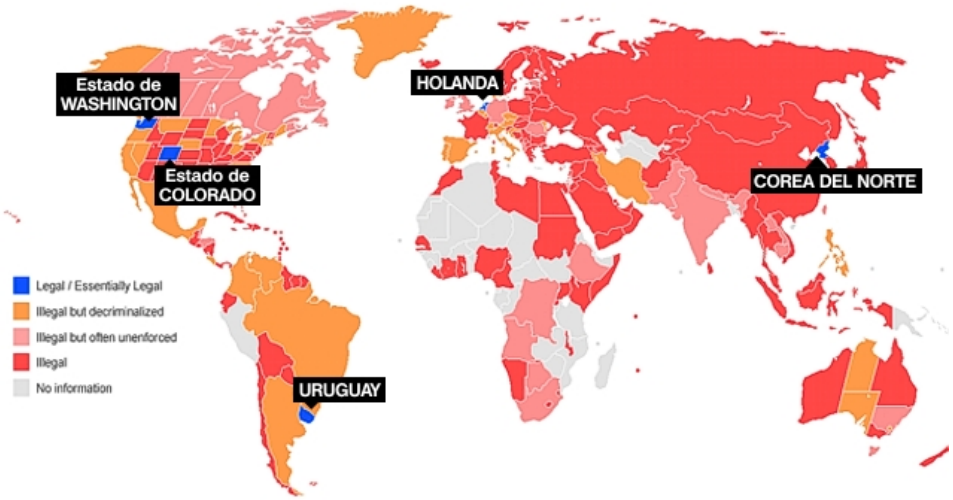


Figura 1: Regulación legal del cannabis

Existe un cierto consenso en que el consumo a largo plazo de la marihuana puede llevar a la adicción, es decir, al uso compulsivo de la sustancia, a pesar de conocerse sus efectos dañinos sobre el funcionamiento social, en el contexto familiar, escolar, laboral y recreativo. Según los expertos (Jacobus [16]), el sistema nervioso central en los adolescentes es más vulnerable a sufrir alteraciones en su estructura y conexiones neuronales por la introducción de sustancias exógenas como el cannabis.

1.1. Contexto y justificación del Trabajo

El uso de cannabis se ha asociado con un mayor riesgo de trastornos del estado de ánimo y ansiedad (Hans-Ulrich[33], Jan Copeland[7]), síntomas psicóticos y psicosis y con deterioro cognitivo (Valentina Lorenzetti [22]). Además, los estudios transversales de neuroimagen sugieren que la exposición crónica al cannabis y el desarrollo de trastornos por consumo de cannabis pueden afectar la morfología cerebral (Valentina Lorenzetti [21]),.

En este trabajo se busca explorar los distintos métodos disponibles para la lectura y procesado de imágenes para clasificar automáticamente los individuos mediante el uso de algoritmos basados en técnicas de *machine learning*. Estas técnicas permiten por ejemplo clasificar imágenes ([26]) de resonancia magnética (en este caso, de paciente fumadores de cannabis y controles) y de esta manera poder obtener una estimación de la probabilidad de pertenecer a los distintos grupos y en el caso de que se disponga y las técnicas lo permitan integrar el modelo final con datos clínicos.

1.2. Los datos para el análisis

Este estudio se basa en el trabajo *Grey Matter Changes Associated with Heavy Cannabis Use: A Longitudinal sMRI Study* de Laura Koenders y Janna Cousijn ([20]) donde se perseguian dos objetivos: i) búsqueda de diferencias en la materia gris entre los grupos de estudio y ii) cambios en el tiempo de la materia gris en consumidores de cannabis.

Se dispone para el presente trabajo de un total de 42 individuos agrupados en: alto fumadores de cannabis y controles sanos. Dichos pacientes fueron reclutados mediante anuncios de internet y en puntos de venta de cannabis (*coffee shop*) en Amsterdam en 2016.

Todos los participantes fueron sometidos a una exploración de MRI, a cada uno de ellos se le realizó un historial detallado de consumo de cannabis y una evaluación psicológica integral.

Los grupos de los que se dispone, mediante la clasificación realizada por los autores, son:

- **Alto consumo de cannabis (CB)**: se definió como el uso de cannabis durante al menos dos años, más de 10 días por mes y sin buscar tratamiento o tener un historial de tratamiento para el consumo de cannabis. Se dispone de un total de 20 individuos.
- **Controles sanos (HC)**: consumieron cannabis menos de 30 veces en su vida y no lo consumieron el último año. Se dispone de un total de 22 individuos.

1.3. Objetivos del Trabajo

Este trabajo consiste en un estudio transversal de imágenes estructurales de resonancia magnética. Los principales objetivos son:

1. **Revisar las técnicas de lectura y procesado de imagen y clasificación automática.**
2. **Clasificar los pacientes según grupo (cannabis / control).**

1.3.1. Objetivos específicos

1. Objetivo 1:
   - *a*) Explorar los distintos métodos de lectura, pre-procesado y análisis de resonancias magnéticas
   - *b*) Realizar un resumen de técnicas de clasificación apropiadas para este tipo de imágenes.
2. Objetivo 2:
   - *a*) Automatizar la extracción de características de la imagen
   - *b*) Aplicar algoritmos de clasificación seleccionados.
   - *c*) Evaluar las predicciones realizadas con las distintas técnicas.

Para alcanzar estos objetivos se propone i) revisar las distintas técnicas de procesado de imágenes de resonancia magnética y métodos de clasificación disponibles y ii) utilizar los distintos tipos de tejido (materia gris, blanca y líquido cefalorraquideo) con el fin de poder clasificar el grupo al que pertenece cada individuo.

1.4. Tecnología: Resonancia magnética

Existen muchos tipos distintos de técnicas de adquisición de imagen. Los más comunes son la resonancia magnética estructural (sMRI) y la resonancia magnética funcional (fMRI). También existen, por ejemplo, las imágenes con tensor de difusión (DTI, Susumu Mori [24]). En nuestro caso se ha trabajado con imágenes de resonancia magnética estructural.

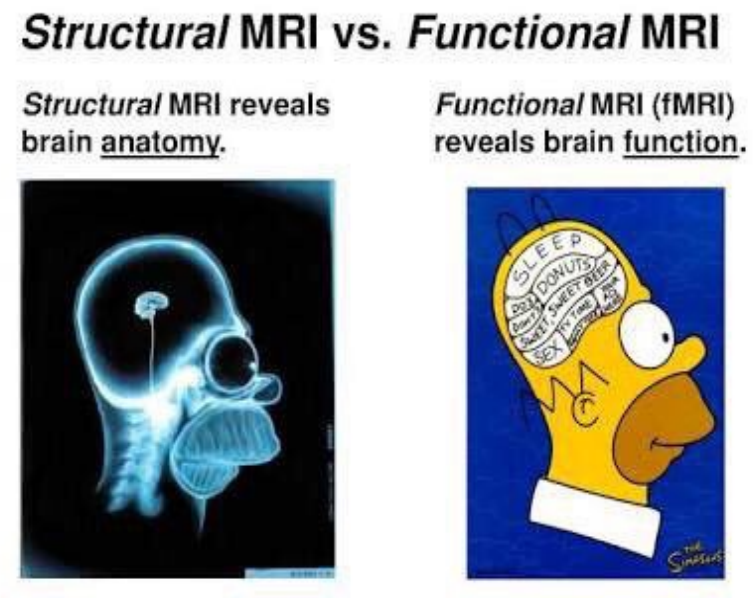


Figura 2: Diferencias entre sMRI y fMRI. Principalmente las sMRI detectan cambios en el volumen tisular de la sustancia gris y blanca, mientras que, las fMRI permiten evaluar los procesos de áreas y estructuras del cerebro en funcionamiento.

Que es un sMRI? La resonancia magnética estructural (sMRI) se trata de una técnica que mide la anatomía del cerebro. Al medir la cantidad de agua en un lugar determinado, esta técnica es capaz de adquirir una imagen anatómica detallada de nuestro cerebro. Esto nos permite distinguir con precisión entre

diferentes tipos de tejidos, como la materia gris y blanca. Las imágenes estructurales son imágenes de alta resolución del cerebro que se usan como imágenes de referencia para múltiples propósitos, tales como el registro (Ver Sección 2), la normalización, la segmentación y la reconstrucción de la superficie .

Como no hay presión de tiempo durante la adquisición de imágenes anatómicas (es decir, no se espera que la anatomía cambie mientras la persona está en el escáner), se puede usar una resolución más alta para registrar imágenes anatómicas, con una extensión de "vóxel" [1] de 0.2 a 1.5 mm, dependiendo en la fuerza del campo magnético en el escáner. Las estructuras de materia gris se ven en tonos más oscuros y las estructuras de materia blanca en colores claros.

Formato de los datos sMRI. Los escáneres de resonancia magnética generan sus datos de neuroimágenes en un formato de datos que depende de cada tecnología y con el que la mayoría de los softwares de análisis no pueden funcionar. Por ejemplo, DICOM (Krzysztof J. [12]) es un formato común, estandarizado y sin procesar de imágenes médicas. Los datos brutos se guardan por capas y deben convertirse a un formato que los paquetes de análisis puedan usar. El formato más frecuente se llama NIfTI.

1.5. Enfoque y método seguido

La elección de este trabajo ha venido motivada por la necesidad de trabajar con datos de imágenes de resonancia magnética. En el campo de la estadística clínica, habitualmente el formato de los datos viene definido por una matriz o tabla, y las técnicas usadas no suelen incluir métodos Machine Learning. Realizar el Trabajo de Fin de Máster en este campo nos pareció una oportunidad para explorar tanto las técnicas de procesado de imagen como investigar que

[1] Vóxel: elemento de volumen. La unidad básica de medida de una resonancia magnética.

técnicas son las más adecuadas para la clasificación de individuos en distintos grupos.

Para realizar una primera toma de contacto con el procesado y análisis de datos procedentes de resonancia magnética se ha completado satisfactoriamente el curso "Introduction to Neurohacking In R" de Coursera ([17]) impartido por la universidad John Hopkins.

Se ha llevado a cabo una revisión inicial de las técnicas de clasificación de imágenes mediante Machine Learning ([29]) con el objetivo de seleccionar las más adecuadas. Para ello se ha tenido en cuenta su idoneidad respecto al problema biológico a resolver, su complejidad de implementación y el tipo de resultados que genera (categorías, probabilidades).

1.6. Planificación del Trabajo

Para la realización del trabajo se ha seguido la calendarización que se muestra el diagrama de Gant de la Figura 3. Los principales hitos se enumeran a continuación:

Hasta el 25-nov-2017

- Preparación y elección del entorno de trabajo (software, paquetes y requerimientos del sistema)
- Extracción de los descriptores
- Automatización descriptores

Hasta el 25-dic-2017

- Aprendizaje mediante algoritmo de clasificación Machine Learning
- Terminar de escribir memoria.

Hasta el 05-ene-2017

- Entrega memoria final y presentación

Name	Duratic	Start	Finish
Obtención de las imágenes	26d?	09/20/2017	10/25/2017
Caracterización de las imágenes	8d?	10/26/2017	11/06/2017
Preprocesado de las imágenes	6d?	11/07/2017	11/14/2017
Entrenar algoritmo y extraer descriptores	6d?	11/14/2017	11/21/2017
Implementar algoritmo de clasificación deep learning	5d?	11/22/2017	11/28/2017
Inicio escritura de la memoria	2d?	11/29/2017	11/30/2017
Implementar algoritmo de clasificación tradicional	3d?	12/01/2017	12/05/2017
Comparación de clasificadores	4d?	12/06/2017	12/11/2017
Incluir variables clínicas al modelo deep learning	3d?	12/12/2017	12/14/2017
Final escritura memoria	11d?	12/11/2017	12/25/2017
Presentación	7d?	12/28/2017	01/05/2018
Correción	7d?	01/01/2018	01/09/2018

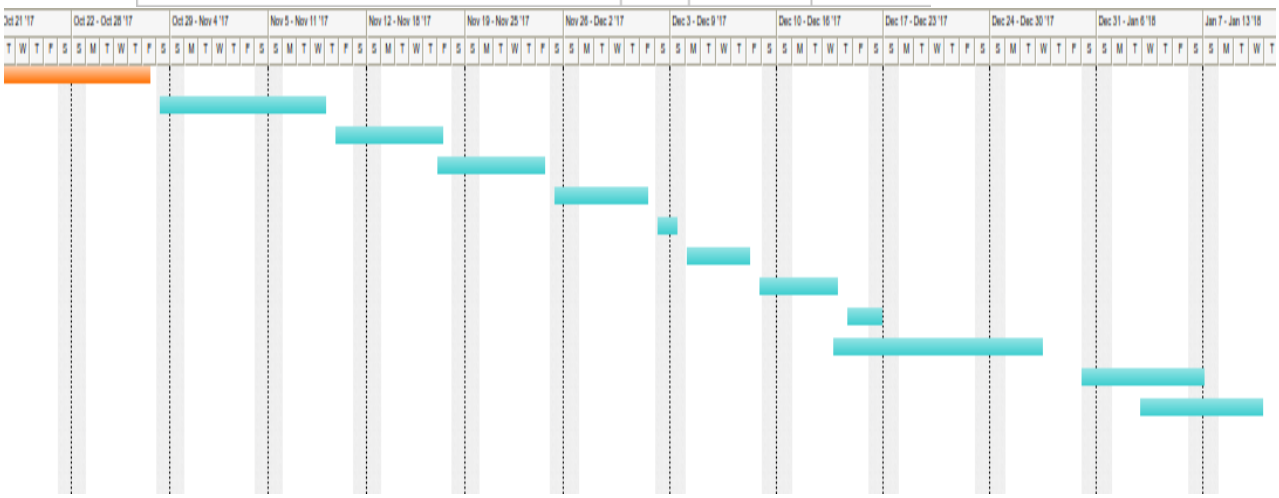


Figura 3: Calendarización TFM.

1.7. Breve sumario de productos obtenidos.

1. **Plan de trabajo.** Descripción del TFM y temporalización de objetivos y tareas

2. **Memoria.** Compendio y discusión del trabajo realizado y los objetivos conseguidos

3. **Producto.** i) Revisión de métodos disponibles para el procesado y análisis de datos de sMRI y ii) clasificación de los individuos en distintos grupos según materia gris.

4. **Presentación.** Audiovisual, con vídeo de presentación oral y diapositivas explicativas

5. **Auto-evaluación.**

1.8. Breve descripción de los otros capítulos de la memoria

El informe se ha realizado teniendo en cuenta la siguiente estructura:

1. **Introducción:** Contextualización del problema y descripción de los datos.

2. **Métodos:** Descripción de los métodos estadísticos y bioinformáticos usados para los análisis.

3. **Resultados:** Explicación de los resultados de los análisis realizados.

4. **Discusión y conclusiones**

1.9. Análisis

En general, un análisis de este tipo se realiza siguiendo el esquema que se muestra en la Figura 4. En este estudio se ha empezado a partir de los análisis de datos de imágenes, ya que, la parte previa ha sido realizada por los investigadores.

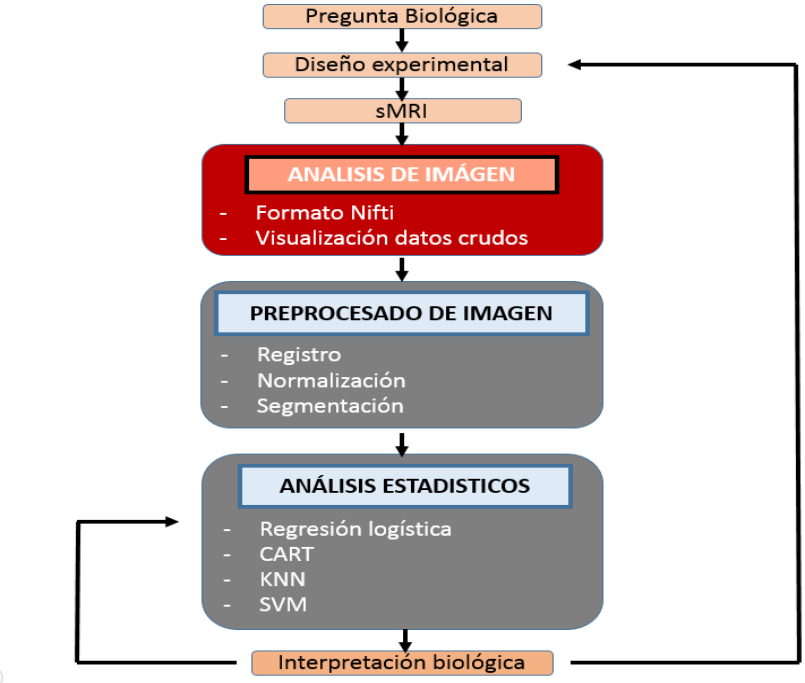


Figura 4: El análisis de MRI puede ser fácilmente visualizado como un proceso que empieza por una pregunta biológica y concluye con una interpretación de los resultados de los análisis que, de alguna forma, confiamos que nos acerque un poco a la respuesta de la pregunta inicial.

Los análisis estadísticos se han realizado usando el lenguaje de programación estadística R y las librerías desarrolladas para el análisis de MRI en el proyecto Neuroconductor [2] ([2]). Para más detalles sobre los métodos ver Sección 2.

[2] Neuroconductor (neuroconductor.org/)es un proyecto de código abierto para el análisis de datos en imágenes, basado en lenguaje de programación R.

2. Métodos

En esta sección se realiza una revisión de las técnicas disponibles. Se muestra una breve descripción de la idea intuitiva que persiguen las distintas técnicas junto a una breve explicación del marco teórico que las sustenta.

Se describen los distintos pasos involucrados en el análisis de neuroimágenes, dividido principalmente en tres bloques:

1. **Lectura y visualización** de las imágenes de resonancia magnética
2. **Pre-procesamiento** de los datos para posteriormente poder realizar el análisis inferencial.
3. **Análisis estadístico y clasificación:** Elección de las técnicas a utilizar.

2.1. Lectura y visualización de imágenes de resonancia magnética

El cerebro ocupa espacio, por lo tanto se ha trabajado con datos de volumen. Los datos se miden en vóxeles, que son como los píxeles utilizados para mostrar las imágenes en una pantalla, pero en 3D. Cada vóxel tiene una dimensión específica, en este caso, 1 mm x 1 mm x 1 mm: un cubo, por lo que este tiene la misma dimensión desde todos los lados. Cada vóxel contiene un valor que representa la señal medida en la ubicación determinada.

Un volumen anatómico estándar, con una resolución de vóxeles de 1 mm contiene casi 17 millones de vóxeles, que están dispuestos en una matriz 3D de 256 x 256 x 256 vóxeles.

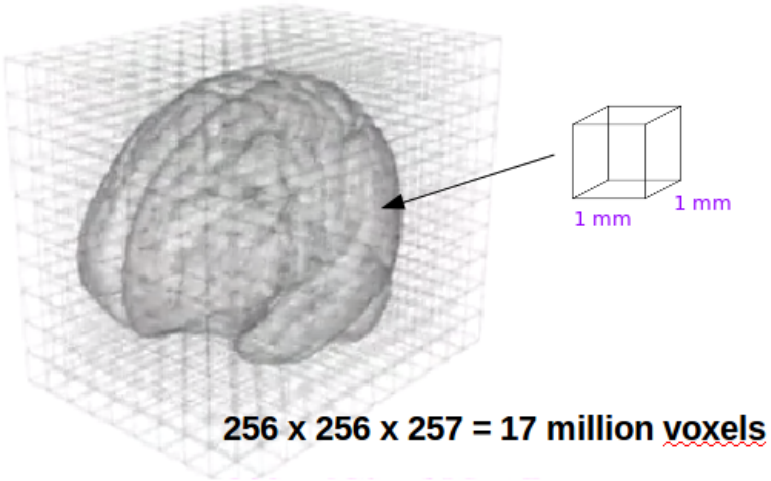


Figura 5: Estructura general resonancia magnética

Como el escáner no es capaz de medir todo el volumen de una vez, este tiene que medir partes del cerebro secuencialmente. Esto se hace midiendo cada uno de los planos del cerebro (habitualmente el horizontal). La resolución de los datos de volumen medidos, por lo tanto, depende de la resolución en el plano (el tamaño de los voxeles), el número de cortes y su grosor (cuántas capas) y el espacio entre las capas.

La calidad de los datos medidos (M. Symms [32]) depende de la resolución y los siguientes parámetros:

- **Tiempo de repetición:** tiempo requerido para escanear un volumen.
- **Tiempo de adquisición:** tiempo requerido para escanear una porción. $TiempoAdquisicion = TiempoRepeticion - (\frac{TiempoRepeticion}{numeroCapas})$.
- **Campo de visión:** define la extensión de un corte, por ejemplo, 256 mm x 256 mm.

De este proceso se obtienen los datos en formato DICOM ('Digital Imaging and Communication in Medicine') que es el formato estándar para el intercambio de imágenes médicas, aún así, este formato de datos no ha sido adoptado por la comunidad científica que trabaja en procesamiento de neuroimagen. Este tipo de datos habitualmente se transforman al formato Nifti (Neuroimaging Informatics Technology Initiative) ([3]), diseñado específicamente para el tratamiento de imágenes vinculadas a las neurociencias, y es el más empleado actualmente por los diversos paquetes de software de procesamiento de neuroimagen y con el que se ha trabajado para realizar los análisis.

2.2. Pre-procesado de imágenes

El análisis automatizado de imágenes de resonancia magnética es un reto debido a la inhomogeneidad de intensidad, a las características propias del individuo, la variabilidad de los rangos de intensidad y contraste, y el ruido. Por lo tanto, antes de la realización de los análisis, se requieren ciertos pasos para hacer que las imágenes sean comparables. Estos pasos se conocen comúnmente como pre-procesamiento. Los pasos típicos de pre-proceso para la resonancia magnética cerebral estructural incluyen los siguientes pasos: registro, normalización y segmentación.

2.2.1. Registro

El registro es la alineación espacial de las imágenes a un espacio anatómico común (Hiba A. Mohammed [15]). El registro de imagen interpaciente ayuda a estandarizar las imágenes de resonancia magnética en un espacio estándar. Existe también el registro intrapaciente que tiene como objetivo alinear las imágenes de diferentes secuencias, por ejemplo, un paciente en dos tiempos dis-

tintos, para obtener una representación multicanal para cada ubicación dentro del cerebro, en nuestro caso, dicho registro no se llevará a cabo ya que se usa únicamente una sMRI por paciente.

En resumen, el registro de imágenes se lleva a cabo con el objetivo de encontrar una correspondencia entre formas idénticas en dos imágenes distintas. Se necesita encontrar una transformación geométrica de uno respecto de la otra. Este tipo de pre-procesado es necesario debido a las distorsiones de movimiento entre la máquina y el objeto.

Se dispone de muchas técnicas que han sido desarrolladas para resolver dicho problema. Existen otros factores como, por ejemplo, ruido, degradación, distorsión, etc. que pueden afectar a la aplicación del registro. Es por ello que se dispone de una gran variedad de métodos de transformación. Dichos métodos, pero, siguen el esquema mostrado en la Figura 6 .

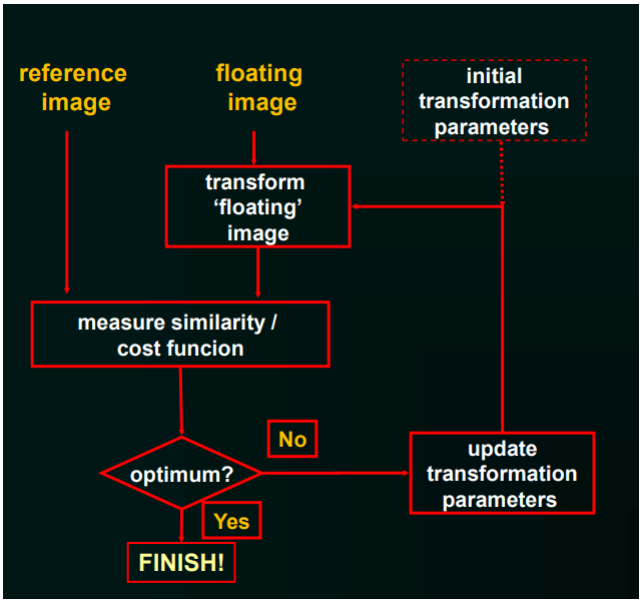


Figura 6: Funcionamiento general de un algoritmo de registro.

Los métodos de registro se dividen principalmente en dos grupos:

- **Rígidos:** cuando la distancia entre dos puntos se preserva al mapearlos con otra imagen. Este tipo de transformación está limitada a rotación, transacción, escalado.
- **No rígidos:** Usadas principalmente cuando se necesita la alineación de estudios de múltiples pacientes con el fin de establecer un rango normal o para la realización de análisis poblacionales (intra-modalidad, inter-sujeto).

Las transformaciones rígidas (John Ashburner et. al [4]) son adecuadas cuando la diferencia entre imágenes implica solo rotación o escalado, dado que nuestros datos requiere un mapeo mas flexible para capturar las diferencias de forma y encontrar una buena alineación de características de la imagen se necesita utilizar el registro no rígido.

Se ha realizado una transformación denominada "elástica" que consiste en parametrización a través de un espacio vectorial regularizado. El algoritmo Demons (Xavier Pennec [25]) original proporciona un ejemplo clásico de usar un espacio vectorial regularizado para el registro no lineal. Los inconvenientes son que el espacio puede no conservar la topología subyacente y también puede resultar demasiado inflexible para capturar cambios posteriores (Nick Tustison [6]). Ambas deficiencias motivan el uso de difeomorfismos

Transformaciones Difeomórficas. El espacio de mapeo elástico, puede ser inadecuado para algunos escenarios de mapeo de gran deformación. El modelo difeomorfo que se ha llevado a cabo para los análisis es la *normalización simétrica* (SyN) ([5] y [19]) dicha transformación es invariante al orden de entrada de las imágenes. Una ventaja adicional del modelo difeomorfico sobre

el modelo elástico es que las transformaciones tanto hacia adelante como hacia atrás se calculan permitiendo así que el fijo modifique al móvil y el móvil al fijo, consiguiendo así una corrección más precisa. El modelo de transformación que se ha usado en este estudio es SyN con un tamaño de paso de 0.25. El tamaño de paso, 0.25, afecta la precisión. Cuanto más pequeño es, generalmente más preciso, pero lleva más tiempo calcular y puede no capturar tanta deformación si la optimización se queda atascada en un mínimo local.

2.2.2. Normalización

El cerebro de cada persona es diferente del del resto de personas. Los cerebros difieren en tamaño y forma. Para comparar las imágenes del cerebro de una persona con las de otra, las imágenes deben ser traducidas a una forma y tamaño comunes, lo que se denomina normalización.

La normalización (Shinohara [30]) consiste en ajustar la posición, la orientación y el tamaño de cada cerebro individual con un cerebro de referencia.Una vez que se completa este paso, se puede realizar un análisis grupal o una comparación entre los datos. Hay diferentes formas de normalizar los datos pero siempre incluye una plantilla y una imagen de origen.

La imagen de la plantilla es el cerebro estándar en el espacio de referencia en el que desea asignar sus datos. La imagen de origen se usa para calcular la matriz de transformación necesaria para asignar la imagen de origen a la imagen de la plantilla. Esta matriz de transformación se usa para mapear el resto de sus imágenes en el espacio de referencia.

Se ha realizado una normalización de intensidad del cerebro, conocida como normalización del cerebro completo (Ellingson [10]). Se ha calculado un Z-score

para cada vóxel usado la media μ_{WB} y la desviación estándar σ_{WB} calculada a partir de todos los vóxeles en la imagen de origen.

$$T1_{WB} = \frac{T1 - \mu_{WB}}{\sigma_{WB}}$$

Además es necesario el realizar la **extracción del cerebro** ('*skull stripping*') para centrarse en los tejidos intracraneales. Los métodos más comunes utilizados para este propósito han sido BET, *Brain Extraction Techniques* (Shaswati Roy [28]) y SPM.

En nuestro caso el método que se ha utilizado es el BET (Brain Extraction Tool) (Stephen M. Smith, [31]). Este método realiza los siguientes pasos:

1. Se calculan los percentiles 2 y 98 y se aplica la siguiente fórmula $(p98 - p2) * 10\,\% + p2$ usada para calcular el punto de corte que se utilizara para eliminar el ruido de fondo [1].
2. Calcula el radio del cerebro y la intensidad media de todos los puntos dentro del 'cerebro esférico' (usado en el último paso).
3. Realiza una pequeña región que va creciendo e iterando para obtener la superficie total del cerebro.
4. Suaviza la superficie
5. Usa la intensidad media para reducir la superficie a la superficie 'real'

2.2.3. Segmentación

La segmentación consiste en la identificación y extracción de los tejidos de cada volumen, mediante una identificación de la escala de grises con los llamados mapas de probabilidad de los tejidos (TPM).

Para realizar una segmentación automatizada es necesario disponer de distintas estructuras cerebrales, por ejemplo, materia blanca (WM), materia gris (GM) y líquido cefalorraquídeo (CSF), esto es utilizado para explorar en los desarrollos cerebrales la evaluación cuantitativa del tejido cerebral y el volumen intracraneal. Los enfoques basados en Atlas (Jimit Doshi [9]) que clasifican tejidos según un conjunto de características de intensidad locales, son los enfoques clásicos que se han utilizado para la segmentación del tejido cerebral.

Se distinguen tres tipos de tejido:

- **Sustancia gris** se asocia con la función del procesamiento de la información, es decir, a la función del razonamiento. Se localiza en la superficie del cerebro, formando la corteza cerebral, que corresponde a la organización más compleja de todo el sistema nervioso.
- **Sustancia blanca** conecta las diferentes áreas del cerebro, transportando los impulsos nerviosos entre neuronas. Se corresponde con la parte inferior del cerebro
- **Fluido Cerebro Espinal** es un líquido de color transparente que baña el encéfalo y la médula espinal, cuyas funciones vitales son proteger el encéfalo, transportar nutrientes al cerebro y eliminar los deshechos, compensar los cambios en el volumen de sangre intracraneal manteniendo una presión constante.

2.3. Análisis y clasificación de individuos

En este apartado se revisan brevemente [3] las técnicas estadísticas y de machine learning utilizadas en los análisis.

Para testar la hipotesis que los consumidores de cannabis tienen un menor volumen de la materia gris se ha realizado la prueba no paramétrica U de Mann Withney unilateral. Además del p-valor se ha tenido en cuenta el tamaño del efecto calculado de la siguiente forma:

$$r = \frac{Z}{\sqrt{N}}$$

donde N es el tamaño total de la muestra $N = (n1+n2)$. Se tiene en cuenta el valor absoluto del tamaño del efecto (RobertJ. Grissom[27]), considerándose un gran tamaño del efecto un valor de 0.5.

Para la **clasificación de los individuos** se ha abordado el problema desde dos prespectivas distintas.

En una primera aproximación se tiene en cuenta el volumen total de materia gris mediante regresión logística implementando posteriormente su correspondiente curva ROC (Hajian-Tilaki [14]) y un CART(*Classification and Regression Tree*, Marshal R J [23]).

En segundo lugar se ha tenido en cuenta la totalidad de los datos usado técnicas de machine learning (*k-nearest neighbor* y *Support Vector Machine*) para clasificar los individuos en ambos grupos según la materia gris. En este

[3] Así como las técnicas de imágen son nuevas, se les ha dedicado más extensión y detalle. Las técnicas estadísticas son la base del máster, es por ello que se ha realizado una revisión más superficial

caso, se ha tenido en cuenta cada uno de los vóxeles (después de realizar una reducción de la dimensión).

En ambos casos se ha creado **un conjunto de datos de entrenamiento y evaluación.** Antes de proceder a la ejecución de las técnicas estadísticas, se selecciona una muestra de entrenamiento para la construcción de los algoritmos , correspondiente a un 67 % de los datos (28 individuos), el 33 % restante (14 individuos), se usa para la validación (es decir, para testar el algoritmo).

Figura 7: Creación y uso de las muestras de entrenamiento y validación

2.3.1. Análisis estadístico (volumen total)

Se obtiene el volumen de materia gris (GM) para cada uno de los individuos a partir del mapa de probabilidad de dicho tejido. Un mapa de probabilidad consiste en una matriz de datos 3D del tamaño del cerebro (habiendo excluido previamente el cráneo). Cada vóxel indica una probabilidad de que ese punto pertenezca a materia gris. Se ha considerado que un punto es GM cuando dicha probabilidad es superior a 0.33 (recordemos que la segmentación extrae materia gris, blanca y CSF).

Para la clasificación de los individuos en grupos a partir del volumen total de materia gris obtenido de las resonancias magnéticas se han utilizado las técnicas que se muestran a continuación:

i) Se ha ajustado un modelo de regresión logística. Se ha tenido en cuenta como variable respuesta el grupo (cannabis o control) y como variable explicativa el volumen. Posteriormente, a efectos de visualización se ha calculado la curva ROC ([14]) a partir de dicho modelo. Una curva ROC se trata de una presentación gráfica de a sensibilidad vs (1- especificidad).

ii) En segundo lugar se ha realizado un árbol de decisión ([23]). Este método deriva de una metodología previa denominada *automatic interaction detection* (Kass, G. V. [18]). Un CART se basa en una estructura en forma de árbol, donde las ramas representan conjuntos de decisiones que generan reglas para la clasificación de un conjunto de datos en subgrupos de datos. Las ramificaciones se generan de forma recursiva hasta que se cumplen ciertos criterios de parada. Es decir, para cada una de las particiones se busca aquella variable y correspondiente punto de corte que mejor clasifica nuestros pacientes.

2.3.2. Machine Learning (información completa)

El aprendizaje automático (Machine Learning) se trata de una rama de la inteligencia artificial (ciencia que intenta imitar las características del sistema nervioso y neuronal humano).

Los algoritmos de machine learning pretenden identificar patrones complejos a partir de grande volúmenes de datos. Para ello es necesario una experiencia o conocimientos previos, en otras palabra, unos datos que sirvan para entrenarlos.

Habitualmente los algoritmos de aprendizaje automático se clasifican en dos subgrupos [13]:

- **Supervisados**: Aplican lo que se ha aprendido en el pasado a un nuevo conjunto de datos. Establece una correspondencia entre las entradas y salidas deseadas del sistema. Un ejemplo de este tipo de machine learning lo encontraríamos con problemas de clasificación.

- **No supervisados**: Realizan inferencias de conjuntos de datos, búsqueda de patrones.

En este estudio se ha trabajo con los algoritmos *K-nearest neighbor* y *Support Vector Machine* de aprendizaje supervisado a fin de poder clasificar los individuos en los dos grupos de interés. En la Figura 8 se muestra el esquema general que sigue este tipo de aprendizaje automático.

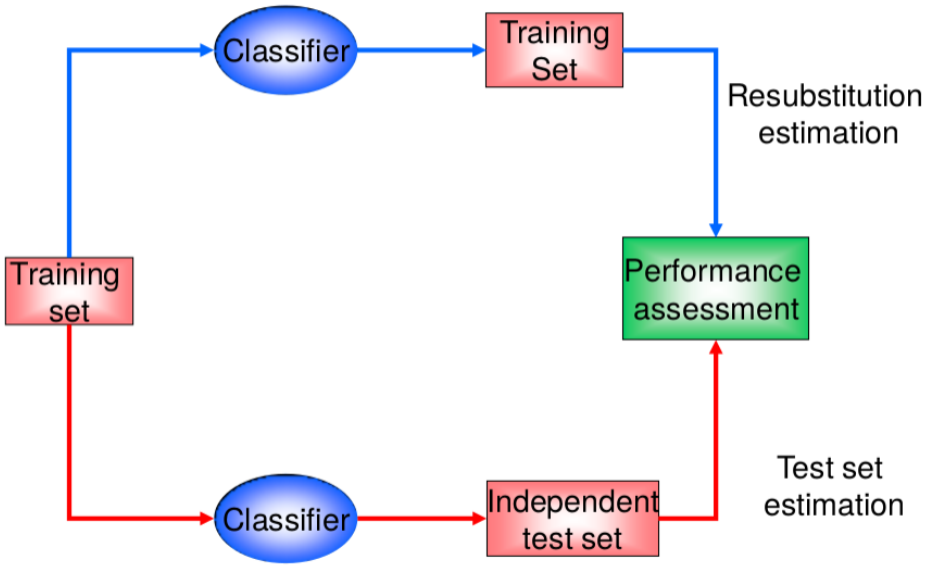


Figura 8: Funcionamiento algoritmos Machine Learning supervisados.

Support Vector Machine (SVM)

El método SVM se trata de un algoritmo de aprendizaje supervisado. Ha sido desarrollada como una técnica robusta para clasificación aplicado a grandes conjuntos de datos complejos con ruido; es decir, con variables inherentes al modelo que para otras técnicas aumentan la posibilidad de error en los resultados pues resultan difíciles de cuantificar y observar.

En resumen, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo pertenece a una categoría u otra, donde los datos de entrada son vistos como un vector p-dimensional.

En el funcionamiento de este algoritmo diferenciamos dos fases:

- Fase de entrenamiento: Busca un hiperplano que separe de forma óptima los puntos de una clase de la otra.
- Fase de uso para la resolución de problemas. En ella, las SVM se convierten en una "caja negra" que proporciona una respuesta (salida) a un problema dado (entrada).

En el Cuadro 1 se muestran las principales ventajas y desventajas de dicha técnica.

VENTAJAS	DESVENTAJAS
Aprendizaje automático	Elevado coste computacional
Eficiente en el caso no lineal	Ineficientes para entrenar (sobre ajuste)
El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente	Selección de función kernel (lineal, Gaussiano, radial,..)
Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada	No está diseñado para identificar los atributos importantes para construir la regla discriminante (Mala clasificación)
No hay óptimo local, como en las redes neuronales	

Cuadro 1: Pros y contras de los SVM

K vecinos más próximos

El método K-nn (K nearest neighbor) sirve para clasificar información y obtener predicciones de nuevos casos.

Su funcionamiento (Figura 2) es el que se enumera a continuación:

1. Calcula las distancias de todos los miembros de la base de datos y toma un número concreto de vecinos (k) que son los que tuvieron distancias más pequeñas.

2. Se visualizan los distintos grupos o clases.

3. Intentamos predecir un nuevo elemento calculando la distancia de Mahalanobis (o euclideana) al resto de individuos. Se clasificará en un grupo u otro dependiendo de los grupos de los k vecinos más cercanos.

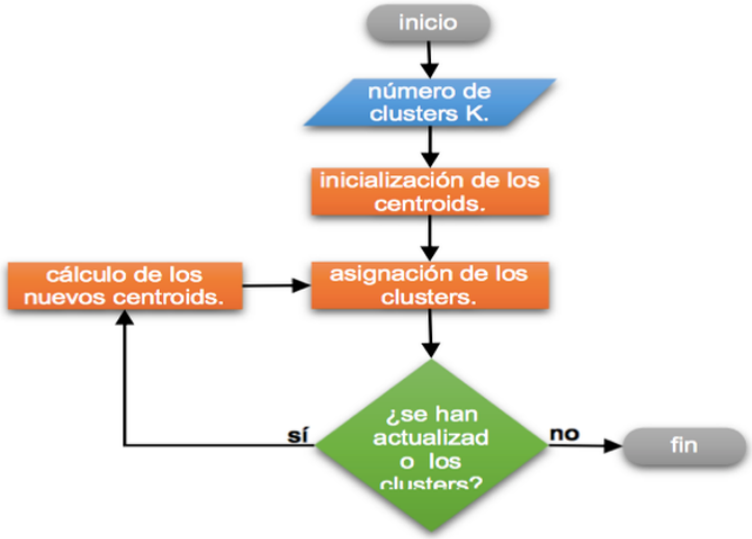


Figura 9: Gráfico de densidad de los datos normalizados

En el Cuadro 2 se muestran las principales ventajas y desventajas de dicha técnica.

VENTAJAS	DESVENTAJAS
No requiere la construcción de un modelo, intuitivo	La clasificación suele ser lenta
Trabaja con los datos originales	Solo funciona con datos numéricos
En general, mientras K sea mayor, la probabilidad de una predicción correcta aumenta.	Puede producir resultados equivocados si el cálculo de distancia no es el adecuado
Realiza su predicción en base a información local	

Cuadro 2: Pros y contras de K-nearest neighbor

2.3.3. Evaluación de la clasificación

La asignación de los individuos a los grupos realizadas con los algoritmos de clasificación, se evaluan con medidas de precisión a través de la matriz de confusión. Una matriz de confusión 3 es una tabla de frecuencias donde las columnas indican la clasificación real y las filas las categorías predichas.

		Realidad	
		Positivo	**Negativo**
Predicción	**Positivo**	VP	FP
	Negativo	FN	VN

Cuadro 3: Matriz de confusión. VP = Verdadero Positivo, VN = Verdadero Negativo, FP = Falso Positivo y FN = Falso Negativo

Se utilizan las siguientes medidas de precisión:

- **Accuracy:** Es el cociente entre los aciertos conseguidos respecto al total.

$$\frac{VP + VN}{VP + FP + FN + VN}$$

- **Sensibilidad:** mide la proporción de acertados (positivos) que hay entre los casos (consumidores de cannabis). Interesa una sensibilidad cercana al 100 %.

$$\frac{VP}{VP + FN}$$

- **Especificidad:** mide la proporción de acertados (negativos) que hay en el grupo de los casos. Interesa una especificidad cercana al 100 %.

$$\frac{VN}{VN + FP}$$

3. Resultados

En este apartado se muestran los resultados obtenidos en la ejecución de los análisis. El codigo utilizado esta disponible en Github [4] github.com/miriamMota.

3.1. Lectura y visualización de resonancia magnética

Cada una de nuestras imágenes contiene $1{,}1927552 \times 10^7$ vóxeles, dispuestos en una matriz 3D de 256 x 182 x 256 vóxeles.

Como exploración inicial de los datos se realizan una serie de gráficos para comprobar la validez de los datos. Estos gráficos incluyen visualización de las resonancias magnéticas de forma axial sagital y coronal, gráficos de intensidad y controles de calidad. En resumen, mediante gráficos se estudia la estructura de los datos con el fin de decidir si parecen correctos o presentan anomalías.

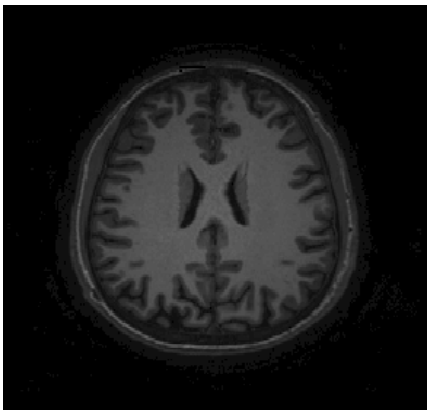


Figura 10: Visualización datos crudos de la capa 125 del individuo 101. En esta imagen se muestra un corte axial, es decir, una capa de la matriz 3D.

[4] Github es una plataforma de desarrolo colaborativo. Permite alojar proyectos utilizando un sistema de control de versiones.

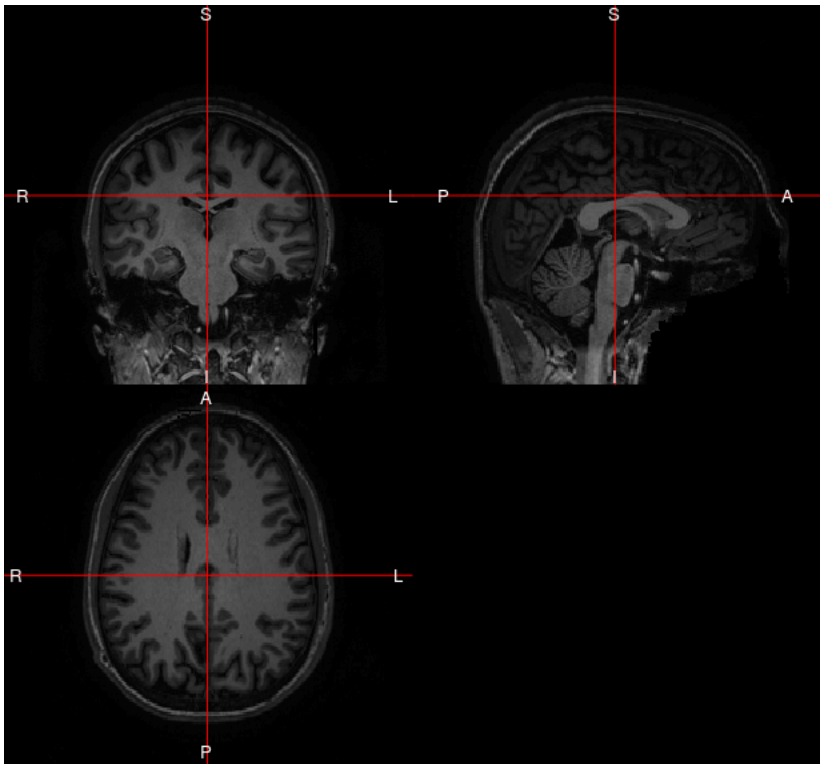


Figura 11: Visualización axial, sagital y coronal de los datos crudos del individuo 101

A continuación se muestra un gráfico de intensidades para cada una de las muestras. Se puede observar que existe gran variabilidad entre ellas sin establecerse un patrón común. Debido a dicha variabilidad es necesario realizar el pre-procesado de las sMRI.

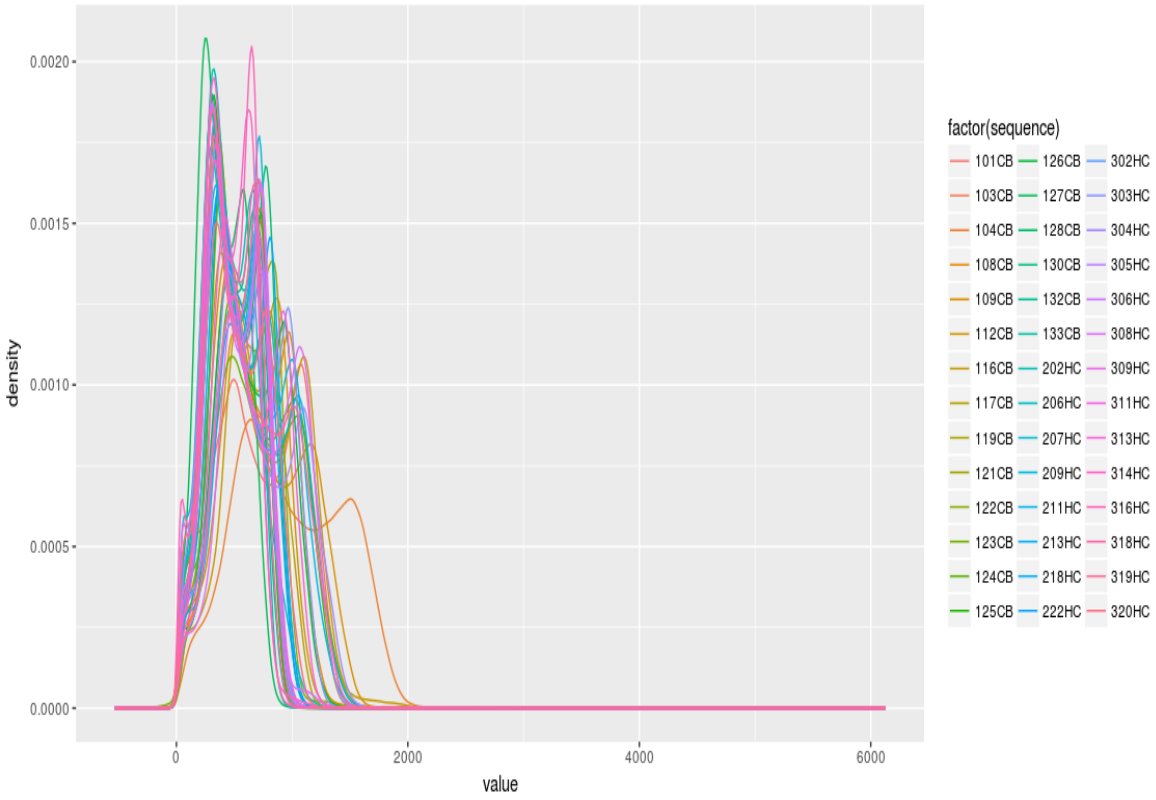


Figura 12: Gráfico intensidades datos crudos.

3.2. Pre-procesado de imágenes

Los datos procedentes de resonancias mágneticas se denominan datos 'crudo' y deben ser preprocessados de diversas formas antes de analizarlos.

Como se ha visto en el Apartado 2, hay muchos pasos involucrados en el análisis de datos de neuroimágenes. Y aún hay más posibilidades para combinarlos. A esto, se añade la complejidad que da la diversidad de paquetes de software distintos para cada paso.

La etapa de registro y normalización podría tener una influencia significati-

va en la precisión de segmentación -extracción de los distintos tipos de tejido- resultantes. Por lo que es importante evaluar cual de las posibilidades se adecua más a cada caso concreto.El objetivo del pre-procesamiento es preparar los datos de manera adecuada para ser alimentados al clasificador. Un adecuado pre-procesamiento dará lugar a una escala, sesgo, tejido cerebral y coordenadas espaciales similares de todos los datos. En este trabajo, se han realizado los siguientes pasos de pre procesamiento previo a enviarlos a los clasificadores.

3.2.1. Registro

En este apartado se ha realizado el registro no rígido (alineación entre distintos pacientes). Se ha utilizado la función `preprocess_mri_within` del paquete `extrantsr` de R. El método de transformación no rígido aplicado es el 'Syn' (*Symmetric normalization*) (Ver 2.2.1).

En las Figuras 13 y 14 se muestra la misma capa y paciente que se ha mostrado en el apartado anterior una vez realizado el registro no rígido.

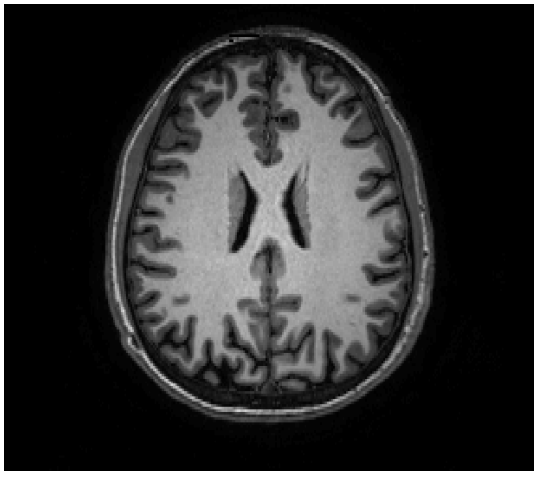


Figura 13: Visualización datos una vez realizado el registro de la capa 125 del individuo 101

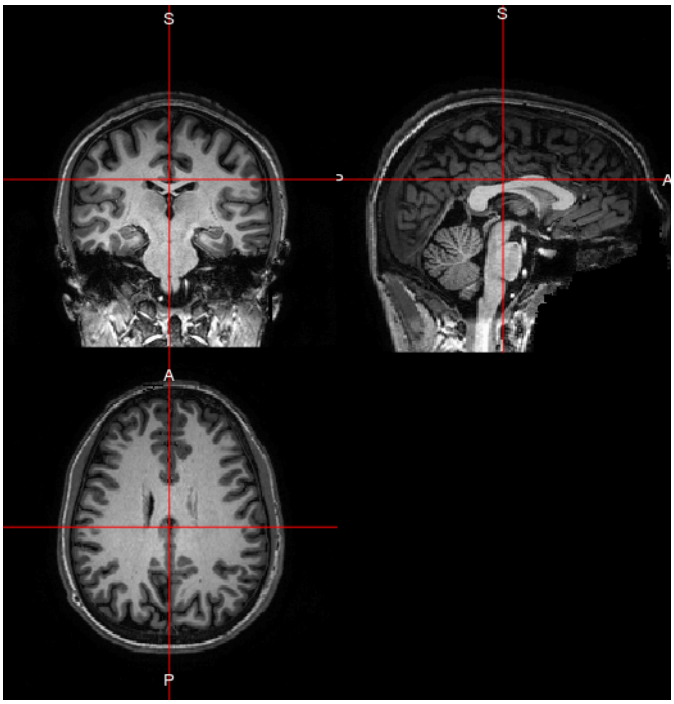


Figura 14: Visualización axial, sagital y coronal de los datos una vez realizado el registro del individuo 101

3.2.2. Normalización

Con el fin de poder comparar los datos, así como para eliminar sesgos técnicos se ha procedido ha normalizar con el método Z-score que se describe en la Sección 2.2.2.

En la Figura 15 se muestra el gráfico de densidad para cada uno de los individuos una vez normalizados los datos. Se puede apreciar que los valores normalizados han quedado claramente en una escala común donde se pueden comparar.

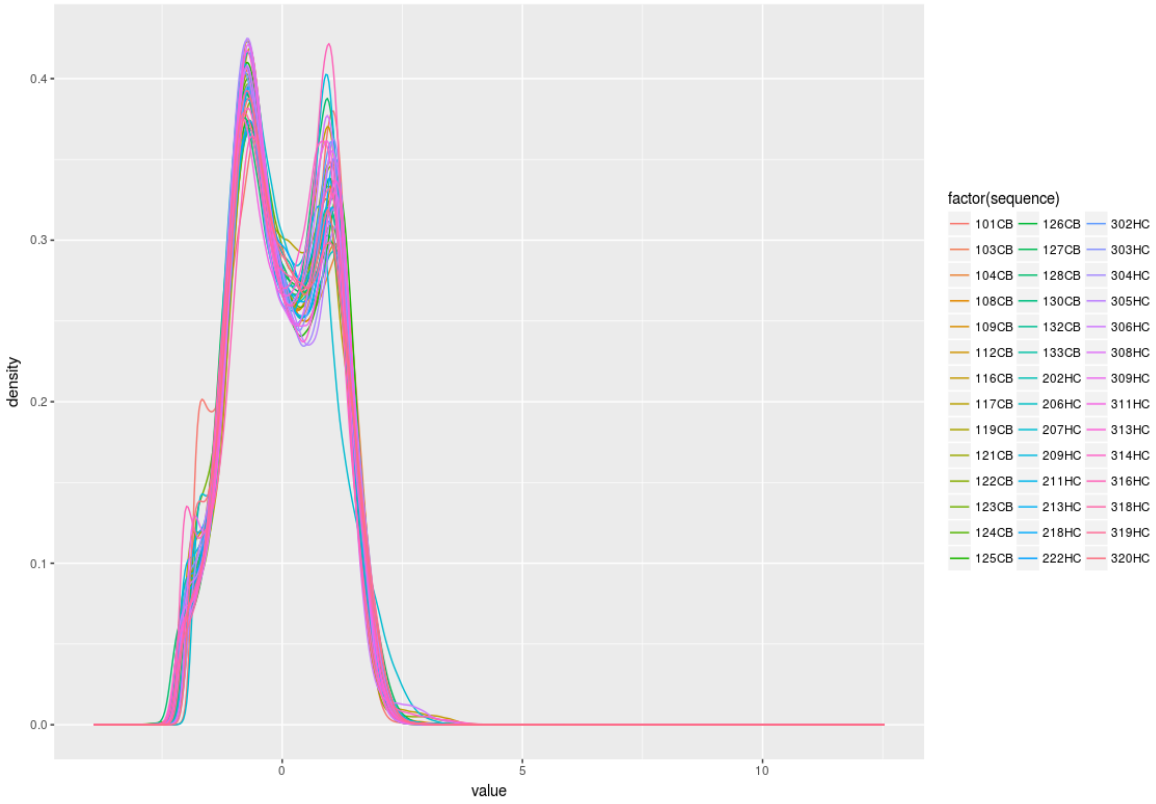


Figura 15: Gráfico de densidad de los datos normalizados

Se ha realizado la extracción del cerebro ('skull stripping') consistente en la eliminación del cráneo, ojos, etc.

En las Figuras 16 y 17 se muestra la misma capa y paciente que se ha ido mostrando en los apartados anteriores una vez realizada la normalización y extracción del cerebro. Se puede observar que la superficie total ha disminuido, teniendo por capa ahora una superficie de 157 x 133, es decir, un total de $2{,}0881 \times 10^4$ vóxeles.

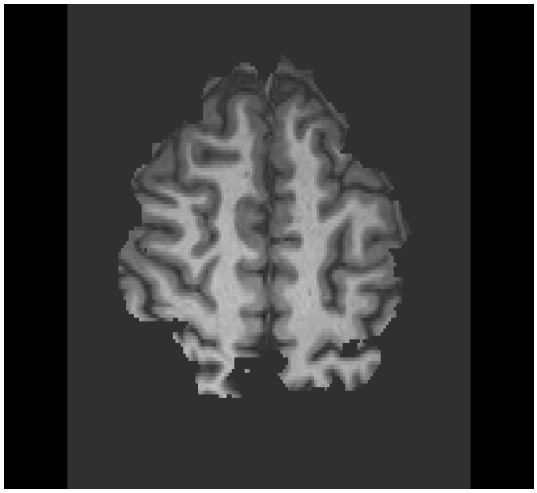


Figura 16: Visualización datos normalizados de la capa 125 del individuo 101

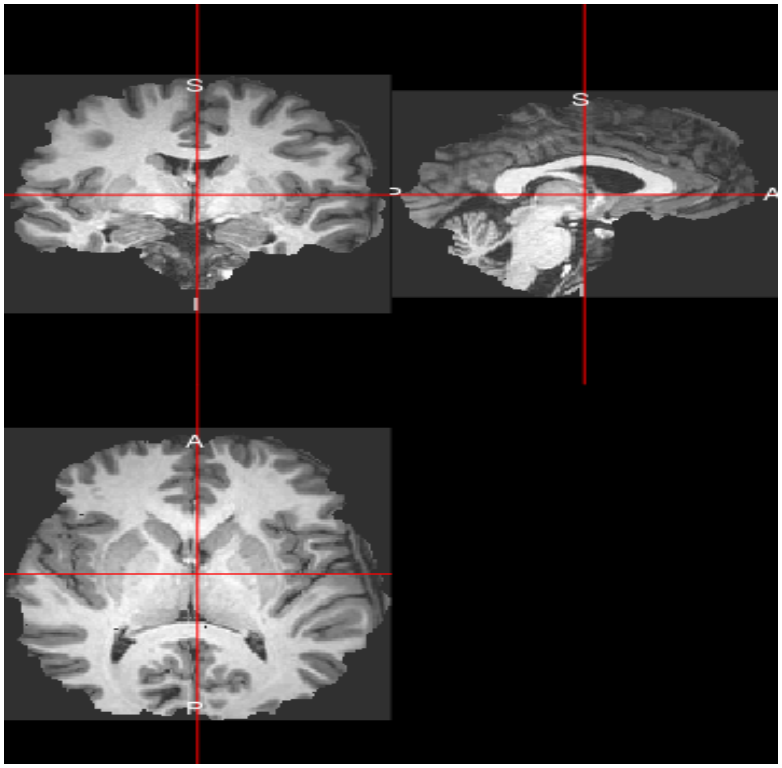


Figura 17: Visualización axial, sagital y coronal de los datos normalizados del individuo 101

3.2.3. Segmentación

La segmentación se trata de un proceso para extraer los distintos tipos de tejidos. Para llevar a cabo la segmentación del cerebro se han utilizado las funciones `bet y fast` del paquete `fslr` de R.

Para cada individuo se obtienen tres archivos: pve0 (CSF), pve1 (materia gris) y pve2 (materia blanca) [5] que contienen un objeto de las mismas características que los datos normalizados. Cada vóxel contiene la probabilidad de que dicho punto sea realmente materias gris.

La siguiente Figura 18 nos muestra en rojo que partes del cerebro son superiores a una probabilidad de 0.33 (punto de corte establecido para que dicho punto sea considerado materia gris.)

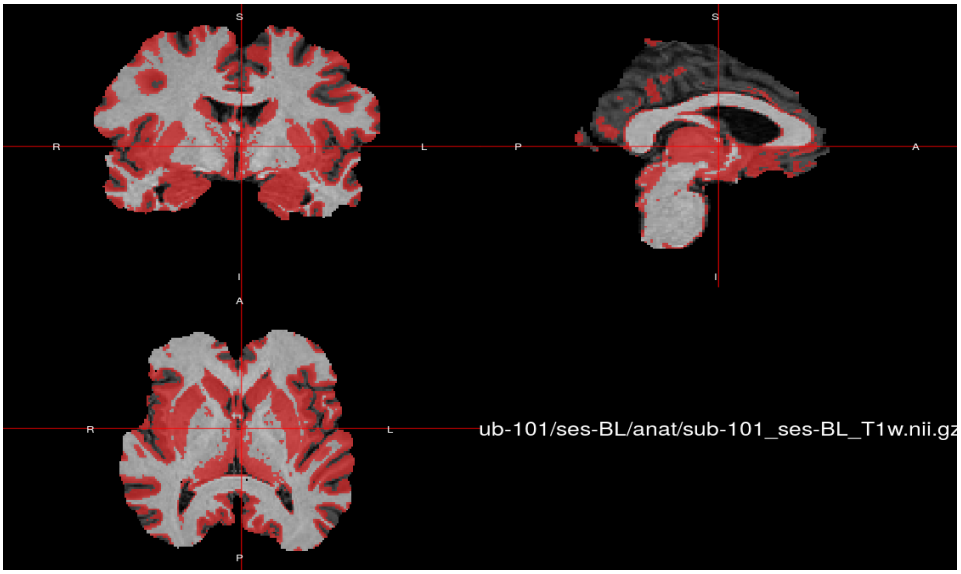


Figura 18: Visualización axial, sagital y coronal de la materia gris individuo 101

[5] PVE: Partial-Volume Effect

3.3. Análisis y clasificación de individuos

En este apartado se ha procedido a la clasificación de individuos según grupo (Cannabis o Control). Dicho problema se ha abordado principalmente desde dos perspectivas: volumen total de materia gris, información completa de todos los vóxeles.

Antes de proceder a la clasificación se ha realizado un análisis descriptivo gráfico y numerico. La Figura 19 muestra un diagrama de cajas agrupado en consumidores de cannabis y controles sanos. Podemos apreciar mas dispersión en los individuos consumidores de cannabis, tendiendo a tener menos volumen total de materia gris.

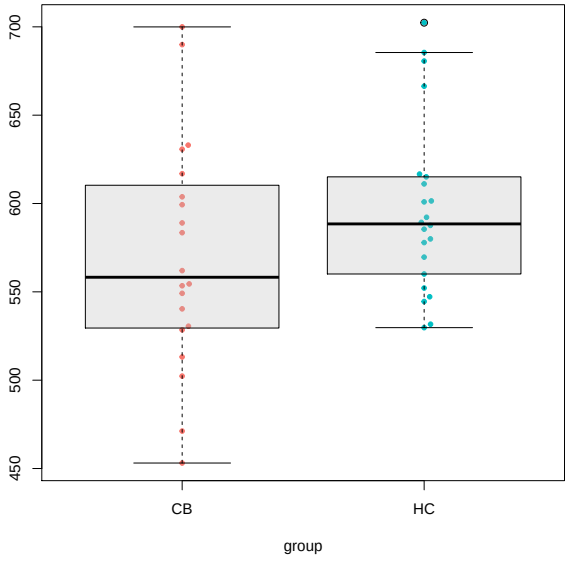


Figura 19: Diagrama de cajas

El (Cuadro 4) muestra un resumen exploratorio por grupo y el p-valor resultante de la prueba no paramétrica U de Mann-Withney unilateral para evaluar si existen diferencias entre grupos para el volumen total de materia gris.

	HC	CB
	N = 22	N = 20
Volume	588 [562;614]	558 [530;607]

Cuadro 4: Análisis exploratorio por grupo (mediana, IQR).
Effect size: 0.21.
P.valor prueba U de Mann Withney: 0.09

Los resultados del test no son concluyentes pero podemos decir que hay indicios de que nuestras hipotesis están bien encaminadas.

Previo a la ejecución de los algoritmos se ha seleccionado una muestra de entrenamiento para la construcción de los algoritmo , correspondiente a un 67 % de los datos (28 individuos), el 33 % restante (14 individuos), se ha usado para la validación (es decir, para testar el algoritmo). Los individuos quedan distribuidos de la siguiente manera:

	CB	HC
Entrenamiento	12	16
Evaluación	8	6

Cuadro 5: Distribución individuos según muestras

3.3.1. Análisis estadístico (volumen total)

En este apartado se ha buscado clasificar los individuos según el **volumen total de materia gris**. Para ello se ha utilizado el paquete `oro.nifti` de R, que nos permite mediante las funciones `prod y voxdim` - y estableciendo un punto de corte para que un vóxel sea considerado materia gris- calcular el volumen para cada uno de los individuos. El punto de corte establecido ha sido una probabilidad de 0.33.

Para clasificar los individuos en consumidores de cannabis o controles sanos se ha ajustado un modelo de regresión logística y posteriormente se ha calculado su correspondiente curva ROC.

El cuadro 6 muestra la matriz de contrastes con las predicciones realizadas usando regresión logística.

	HC	CB
HC	5	6
CB	1	2

Cuadro 6: Tabla de predicciones con regresión logística. Las columnas son los datos reales y las filas las predicciones.

En la Figura 20 se muestra con una cruz azul el punto de corte óptimo calculado con el Indice de Youden (punto que maximiza la especificidad y sensibilidad). Además, se indica el AUC (Area Under the Curve), se considerará un AUC adecuado a partir de 70 %.

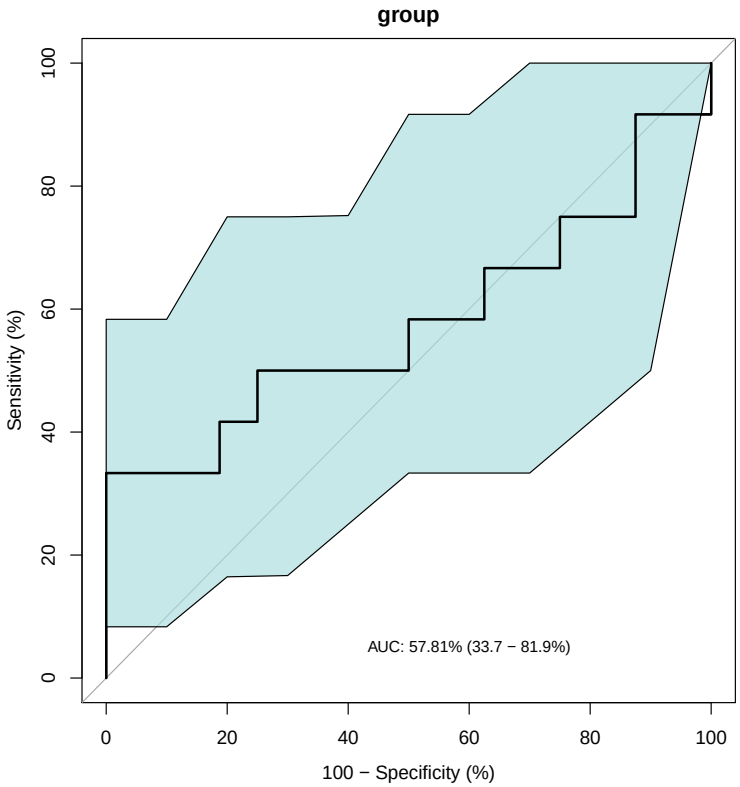
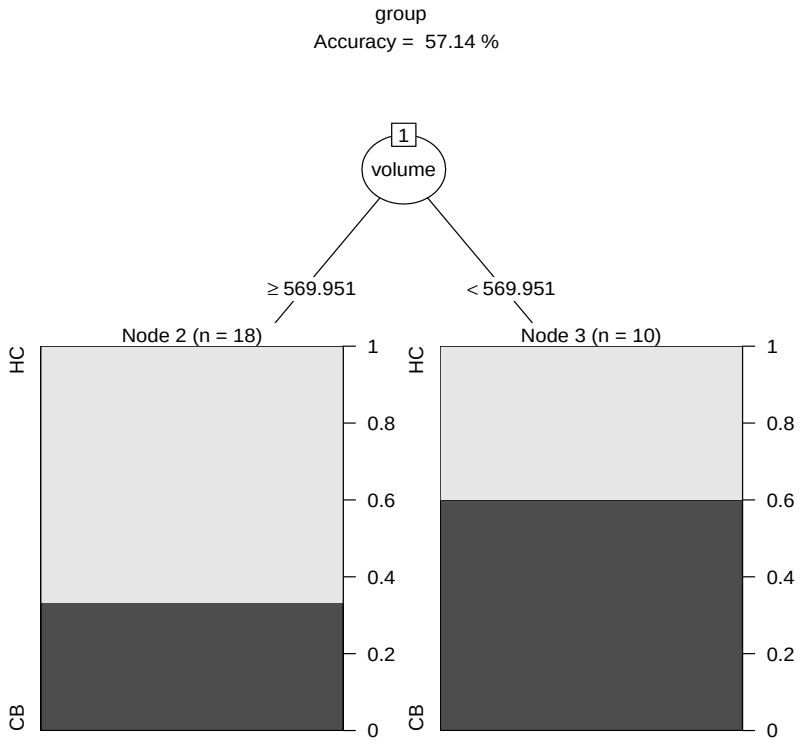


Figura 20: Curva ROC modelo de regresión logística

Se ha obtenido una 'accuracy' (clasificación total correcta) del 50 %. La especificidad, es decir, la correcta predicción de los controles ha sido de un 25 %. La sensibilidad -correcta predicción de los casos- con un total del 83.33 %

A continuación se ha realizado un árbol de clasificación (CART). Se han clasificado a los individuos según los grupos (Cannabis y Control) a partir del volumen total de materia gris.



En el gráfico resultante aparecen aquellos puntos de corte que mejor discriminan a los individuos. En el título se muestra el error de clasificación interna, es decir, usando el árbol obtenido sobre nuestros datos, no se clasificará correctamente dicho porcentaje de los individuos. La especificidad ha sido del 62.5 % y la sensibilidad ha sido de 50 %

3.3.2. Machine Learning (información completa)

En este apartado se explican los **resultados obtenidos para cada uno de los algoritmos utilizados**.

Previo a la ejecución de los algoritmos se ha realizado un reducción de la dimensión eliminando aquellos vóxeles que tenían una probabilidad de 0 - en materia gris- para todos los individuos, quedándonos así con un total de 2472367.

En primer lugar se ha procedido a la **construcción del modelo**. Una vez obtenido el algoritmo con el que trabajaremos lo **validaremos con los datos test**. Para evaluar los **resultados de clasificación** se ha usado la función `contrastMatrix`.

Suport Vector Machine

Para llevar a cabo la implementación del algoritmo SVM se ha utilizado la función `svm` del paquete `e1071` de R. Despues de varias iteraciones se ha establecido como función del kernel la 'lineal'.

Los resultados de clasificación que se han obtenido se muestran en el siguiente Cuadro:

	CB	HC
CB	5	1
HC	3	5

Cuadro 7: Tabla de predicciones para algoritmo k-nn. Las columnas son los datos reales y las filas las predicciones.

Se ha obtenido una 'accuracy' (clasificación total correcta) del 71.43 %. La especificidad, es decir, la correcta predicción de los controles ha sido de un 83.33 %. La sensibilidad -correcta predicción de los casos- a un total del 62.5 %

K-nearest neighbor

Para llevar a cabo la implentación del algoritmo k-nn se ha utilizado la función `knn` del paquete `class` de R. Despues de varias iteraciones se ha establecido como número optimo de vecinos, 5.

Los resultados de clasificación que se han obtenido se muestran en el siguiente Cuadro:

	CB	HC
CB	6	1
HC	2	5

Cuadro 8: Tabla de predicciones para algoritmo k-nn. Las columnas son los datos reales y las filas las predicciones.

Se ha obtenido una 'accuracy' (clasificación total correcta) del 78.57 %. La especificidad, es decir, la correcta predicción de los controles ha sido de un 83.33 %. La sensibilidad -correcta predicción de los casos- a un total del 75 %

3.3.3. Evaluación de la clasificación

En este apartado se ha realizado una tabla de resumen de los distintos resultados de clasificación para los algoritmos de predicción testados.

En el Cuadro 9 se puede apreciar que los análisis realizados teniendo en cuenta únicamente el volumen total realizan peores prediciones. En cambio, al usar la información completa, la exactitud en la clasificación total aumenta considerablemente.

También cabe destacar, que entre los métodos que usan la información completa, el más preciso es k-nearest neighbor con 5 vecinos.

Datos	Método	Accuracy	Especificidad	Sensibilidad
Volumen Total	Regresión logística	50.00	25.00	83.33
Volumen Total	CART	57.14	62.50	50.00
Información completa	K-nn (5)	78.57	83.33	75.00
Información completa	SVM (lineal)	71.43	83.33	62.50

Cuadro 9: Resumen de clasificación de los datos de evaluación

Es importante aclarar que aunque los métodos utilizados para analizar el volumen total no estan indicados como Machine Learning, depende de la fuente que miremos, tanto la regresión logística como los CART pueden estar considerados de aprendizaje automático.

4. Discusión

Este trabajo partía de un estudio predictivo de desarrollo de esclerosis múltiple mediante una serie de biomarcadores genéticos (Nicolas Fissolo [11]). La idea es trabajar con las imágenes de resonancia magnética de dichos pacientes y aplicar técnicas Deep Learning (con la aceptación de los investigadores) para ver si mejoraban las predicciones. En el momento de iniciar el trabajo se nos negó el acceso a las imágenes (por parte del servicio) y se tuvo que redefinir el mismo. Se encontró un nuevo estudio disponible en la plataforma *openFMRI*. Para este estudio, finalmente, no se han aplicado técnicas Deep Learning para la clasificación debido a que una gran parte del trabajo ha consistido en el pre-procesado de los datos además de que el número de muestras disponible era muy bajo.

En el trabajo de donde se han obtenido los datos, *Grey Matter Changes Associated with Heavy Cannabis Use: A Longitudinal sMRI Study.*, los investigadores observaban i) una leve cambio en la materia gris según el grupo (Cannabis o Control) en 3 zonas del cerebro: hipocampo izquierdo, amígdala y circunvolución temporal superiory ii) a nivel longitudinal, es decir, los pacientes consumidores de cannabis seguidos a lo largo del tiempo, no detectan cambios en el volumen de la materia gris.

En este trabajo se ha realizado una revisión de técnicas de pre-procesado de imagen y una posterior aplicación clasificando los individuos en controles sanos y alto consumidores de cannabis. Se han usado métodos para clasificar en base al volumen total y métodos para clasificar basándose en la información completa. Estas segundas dan un mejor resultado y se podría concluir que resultan adecuadas para este problema.

Se han probado distintos algoritmos y distintos ajustes de estos para llevar a cabo la predicción de "nuevos" individuos clasificándolos según consumidores o no de cannabis.

Este trabajo no esta exento de limitaciones, destacamos:

El tamaño de las muestras utilizadas es bastante limitado lo que determina que el estudio tenga poca potencia por lo que probablemente habrá menos reproducibilidad y más falsos negativos de los que serían deseables si se utilizará un mayor número de muestras.

En cada paso del proceso se han tomado decisiones acerca de los métodos a seguir para el registro, normalización, segmentación, etc. La decisión de si estos métodos son los más adecuados o no, es probablemente subjetiva por lo que sería interesante saber como cambian los resultados si se tomaran otras decisiones.

Los problemas anteriores no son, sin embargo, problemas de este estudio concreto, sino en general de los estudios basados en imágenes.

Asimismo, cabe destacar la importancia del uso correcto de las técnicas estadísticas en análisis de neuroimágen y estudios científicos.

Como futuras líneas de trabajo sería interesante i) realizar dichos análisis teniendo en cuenta las distintas partes del cerebro (analizando tanto el volumen total como con la información completa) y ii) aplicar técnicas de clasificación Deep Learning.

5. Conclusiones

Se ha realizado un resumen de las principales técnicas de pre-procesado para imágenes de resonancia magnética estructurales.

Los controles de calidad llevados a cabo mediante visualización en la primera parte de este estudio han permitido establecer que los datos con los que se ha trabajado eran validos.

Los análisis llevados a cabo sobre los datos registrados y normalizados han permitido proceder con la adecuada extracción de los distintos tipos de tejido.

La mejor técnica de predicción para la predicción de personas consumidoras o no de cannabis según el volumen de materia gris, en términos de predicción absoluta, ha sido el algoritmo k-nearest neighbor (5 vecinos). Seguido de cerca por SVM.

Aparentemente los algoritmos utilizados tienen más dificultad (en estos datos) de predecir correctamente los casos (consumidores de cannabis) que los controles (controles sanos) a excepción de la regresión logística que sucede lo contrario. Esta conclusión se ve sin embargo limitada por el bajo número de individuos.

Se ha observado que el uso de la información completa (todos los vóxeles) produce mejores clasificaciones que teniendo en cuenta unicamente el volumen total de materia gris.

6. Glosario

- **Resonancia magnética**: Es un examen imagenológico que utiliza imanes y ondas de radio potentes para crear imágenes del cuerpo.
- **Materia gris**: Tejido del cerebro y la médula espinal que consta principalmente de cuerpos de células nerviosas y dendritas de ramificación.
- **Materia blanca**: tejido del cerebro y la médula espinal que consta principalmente de fibras nerviosas y sus vainas de mielina.
- **CSF, Cerebrospinal Fluid**: un líquido acuoso que fluye en los ventrículos y alrededor de la superficie del salvado y la médula espinal
- **Registro**: El proceso de transformar imágenes en un sistema de coordenadas estándar
- **Segmentación**: El proceso de partición de la imagen de acuerdo con diferentes tipos de tejidos.
- **Capa**: la sección transversal del cerebro que se visualiza.
- **Imagen plantilla**: una imagen estandarizada utilizada como objetivo en el proceso de registro de imagen
- **Voxel**: un elemento de volumen. La unidad básica de medida de una resonancia magnética.
- **Axial**: Corte que divide el cerebro en inferior y superior, es decir, corte horizontal.
- **Sagital**: Corte que, visto el cuerpo de frente, divide el cerebro en dos mitades (derecha y izquierda).
- **Coronal**: Corte que divide el cerebro en anterior y posterior.

7. Bibliografía

Referencias

[1] Brain Extraction/Segmentation.

[2] Neuroconductor, https://neuroconductor.org/.

[3] The NIFTI file format — Brainder.

[4] John Ashburner and Karl J Friston. Rigid Body Registration.

[5] B B Avants, C L Epstein, M Grossman, and J C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, feb 2008.

[6] Nick Tustison Brian B. Avants and Gang Song. Advanced Normalization Tools (ANTS). 2011.

[7] Jan Copeland, Sally Rooke, and Wendy Swift. Changes in cannabis use among young people. *Current Opinion in Psychiatry*, 26(4):325–329, jul 2013.

[8] Louisa Degenhardt and Wayne Hall. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *Lancet (London, England)*, 379(9810):55–70, jan 2012.

[9] Jimit Doshi, Guray Erus, Yangming Ou, Bilwaj Gaonkar, and Christos Davatzikos. Multi-atlas skull-stripping. *Academic radiology*, 20(12):1566–76, dec 2013.

[10] Benjamin M. Ellingson, Taryar Zaw, Timothy F. Cloughesy, Kourosh M. Naeini, Shadi Lalezari, Sandy Mong, Albert Lai, Phioanh L. Nghiemphu, and Whitney B. Pope. Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas. *Journal of Magnetic Resonance Imaging*, 35(6):1472–1477, jun 2012.

[11] Nicolas Fissolo, Béatrice Pignolet, Clara Matute-Blanch, Juan Carlos Triviño, Berta Miró, Miriam Mota, Santiago Perez-Hoyos, Alex Sanchez, Patrick Vermersch, Aurélie Ruet, Jérôme de Sèze, Pierre Labauge, Sandra Vukusic, Caroline Papeix, Laurent Almoyna, Ayman Tourbah, Pierre Clavelou, Thibault Moreau, Jean Pelletier, Christine Lebrun-Frenay, Xavier Montalban, David Brassat, and Manuel Comabella. Matrix metalloproteinase 9 is decreased in natalizumab-treated multiple sclerosis patients at risk for progressive multifocal leukoencephalopathy. *Annals of Neurology*, 82(2):186–195, aug 2017.

[12] Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044, jun 2016.

[13] Luis Guerra, Laura M McGarry, Víctor Robles, Concha Bielza, Pedro Larrañaga, and Rafael Yuste. Comparison between supervised and unsu-

pervised classifications of neuronal cell types: a case study. *Developmental neurobiology*, 71(1):71–82, jan 2011.

[14] Karimollah Hajian-Tilaki. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine*, 4(2):627–35, 2013.

[15] Hiba A. Mohammed. The Image Registration Techniques for Medical Imaging (MRI-CT).

[16] Joanna Jacobus and Susan F Tapert. Effects of cannabis on the adolescent brain. *Current pharmaceutical design*, 20(13):2186–93, 2014.

[17] John Hopkins University. Introduction to Neurohacking In R.

[18] G. V. Kass, Kass, and G. V. Automatic Interaction Detection (AID) Techniques. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2006.

[19] Junghoon Kim, Brian Avants, Sunil Patel, John Whyte, Branch H. Coslett, John Pluta, John A. Detre, and James C. Gee. Structural consequences of diffuse traumatic brain injury: A large deformation tensor-based morphometry study. *NeuroImage*, 39(3):1014–1026, feb 2008.

[20] Laura Koenders, Janna Cousijn, Wilhelmina A. M. Vingerhoets, Wim van den Brink, Reinout W. Wiers, Carin J. Meijer, Marise W. J. Machielsen, Dick J. Veltman, Anneke E. Goudriaan, and Lieuwe de Haan. Grey Matter Changes Associated with Heavy Cannabis Use: A Longitudinal sMRI Study. *PLOS ONE*, 11(5):e0152482, may 2016.

[21] Valentina Lorenzetti, Dan I. Lubman, Sarah Whittle, Nadia Solowij, and Murat Yücel. Structural MRI Findings in Long-Term Cannabis Users:

What Do We Know? *Substance Use & Misuse*, 45(11):1787–1808, jun 2010.

[22] Valentina Lorenzetti, Nadia Solowij, Alex Fornito, Dan Ian Lubman, and Murat Yucel. The association between regular cannabis exposure and alterations of human brain morphology: an updated review of the literature. *Current pharmaceutical design*, 20(13):2138–67, 2014.

[23] R J Marshall. The use of classification and regression trees in clinical epidemiology. *Journal of clinical epidemiology*, 54(6):603–9, jun 2001.

[24] Susumu Mori and Jiangyang Zhang. Principles of Diffusion Tensor Imaging and Its Applications to Basic Neuroscience Research. *Neuron*, 51(5):527–539, sep 2006.

[25] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the Demon's Algorithm: 3D Non-rigid Registration by Gradient Descent. pages 597–605. Springer, Berlin, Heidelberg, 1999.

[26] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.

[27] John J. Kim Robert J. Grissom. Effect Sizes for Research: Univariate and Multivariate Applications - Robert J. Grissom, John J. Kim - Google Libros.

[28] Shaswati Roy and Pradipta Maji. A simple skull stripping algorithm for brain MRI. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6. IEEE, jan 2015.

[29] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19:221–248, jun 2017.

[30] Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, for the Australian Imaging Biomarkers Lifestyle Flagship Study of Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, and the Alzheimer's Disease Neuroimaging Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage. Clinical*, 6:9–19, 2014.

[31] Stephen M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, nov 2002.

[32] M Symms, H R Jäger, K Schmierer, and T A Yousry. A review of structural magnetic resonance neuroimaging. *J Neurol Neurosurg Psychiatry*, 75:1235–1244, 2004.

[33] Hans-Ulrich Wittchen, Christine Fröhlich, Silke Behrendt, Agnes Günther, Jürgen Rehm, Petra Zimmermann, Roselind Lieb, and Axel Perkonigg. Cannabis use and cannabis use disorders and their relationship to mental disorders: a 10-year prospective-longitudinal community study in adolescents. *Drug and alcohol dependence*, 88 Suppl 1:S60–70, apr 2007.

8. Anexos

Este trabajo ha sido realizado con la utilización de R y `knitr`. Knitr es un paquete de R que permite integrar código R con Latex, de esta manera, el estudio es fácilmente reproducible.

El pre-procesado de las imágenes se ha realizado en un script de R a parte, debido al coste computacional. Una vez obtenidos los archivos finales se han utilizado para la realización de los análisis estadisticos mediante ".Rnw" (R + knitr).

Los 'scripts' de los análisis estan disponibles en Github github.com/miriammota.