



UNIVERSITAT DE
BARCELONA

Trabajo de Final de Grado

GRADO DE INGENIERIA INFORMÁTICA

**Facultad de Matemáticas
Universitat de Barcelona**

**ESTUDIO DE IMÁGENES DE RESONANCIA
MAGNÉTICA FUNCIONAL EN REPOSO PARA
LA PREDICCIÓN DE VARIABLES
PERSONALES**

Juan Luis Moral Pérez

Directora: Laura Igual
Realizado a: Departamento de Matemáticas
e Informática

Barcelona, 30 de junio de 2016

«Es preciso sacudir enérgicamente el bosque
de las neuronas cerebrales adormecidas;
es menester hacerlas vibrar con la emoción de lo nuevo
e infundirles nobles y elevadas inquietudes.»

RAMÓN Y CAJAL

Abstract

This project is focused on the creation of a classification system that separates a group of subjects according to their gender based on data from magnetic resonance images (MRI) in a resting state.

The images from MRI in a resting state are a tool to measure the brain connectivity or functioning that is currently being used for many neuroscience studies.

This project, in particular, uses the representation of facts based on the Network in a resting state to characterize the functional connectivity of the subjects for the visualization of the obtained results.

As well as evaluating the accuracy of the classification system developed, another objective of the project is to determine which of the cerebral networks are more discriminative in the task of separating men and women.

The methodology utilized combines two types of automatic learning: unsupervised learning, as in the Independent Component Analysis and the Principal Components Analysis, and supervised learning, as is the K-NN and SVM classifiers.

The results obtained are promising, because it finds a RSN that discriminates both sex and we also note that the Principal Component Analysis does not affect when classifying.

Keywords: *resting state, fMRI, Independent Component Analysis, Dual Regression, Principal Component Analysis, SVM*

Resum

Aquest projecte està enfocat a la creació d'un sistema de classificació que separi un conjunt de subjectes, segons el seu sexe, a partir de dades d'imatges per ressonància magnètica en estat de repòs.

Les imatges per ressonància magnètica en estat de repòs són una eina per mesurar la connectivitat cerebral o funcional, que actualment s'està utilitzant en nombrosos estudis de neurociència.

En particular, en aquest projecte, es fa ús de la representació de les dades basada en Xarxes en estat de repòs per caracteritzar la connectivitat funcional dels subjectes, i per a la visualització dels resultats obtinguts.

A més d'avaluar la precisió del sistema de classificació desenvolupat, un altre objectiu del projecte és esbrinar quins d'aquestes xarxes cerebrals són més discriminadors en aquesta tasca de separar homes i dones.

La metodologia utilitzada combina dues tipologies d'aprenentatge automàtic: l'aprenentatge no supervisat, com l'Anàlisi de Components Independents i l'Anàlisi de Components Principals; i aprenentatge supervisat, com els classificadors K-NN i SVM.

Les dades obtingudes són prometedores, ja que s'aconsegueix trobar una RSN que discrimina tots dos sexes i, a més, aconseguim observar que l'Anàlisi de Components Principals no afecta a l'hora de la classificació.

Paraules clau: estat de repòs, fMRI, Anàlisi de Components Independent, Regressió Dual, Anàlisi de Components Principals, SVM

Resumen

Este proyecto está enfocado a la creación de un sistema de clasificación que separe un conjunto de sujetos según su sexo a partir de datos de imágenes por resonancia magnética en estado de reposo.

Las imágenes por resonancia magnética en estado de reposo son una herramienta para medir la conectividad cerebral o funcional que actualmente se está utilizando en numerosos estudios de neurociencia.

En particular, en este proyecto, se hace uso de la representación de los datos basada en Redes en estado de reposo para caracterizar la conectividad funcional de los sujetos y para la visualización de los resultados obtenidos.

Además de evaluar la precisión del sistema de clasificación desarrollado, otro objetivo del proyecto es averiguar cuáles de estas redes cerebrales son más discriminativas en esta tarea de separar hombres y mujeres.

La metodología utilizada combina dos tipologías de aprendizaje automático: el aprendizaje no supervisado, como el Análisis de Componentes Independientes y el Análisis de Componentes Principales; y aprendizaje supervisado, como los clasificadores K-NN y SVM.

Los datos obtenidos son prometedores, ya que se consigue encontrar una RSN que discrimine ambos sexos y además logramos observar que el Análisis de Componentes Principales no afecta a la hora de clasificar.

Palabras clave: *estado de reposo, fMRI, Análisis de Componentes Independiente, Regresión Dual, Análisis de Componentes Principales, SVM*

Agradecimientos

Quisiera agradecer a varias personas la ayuda que me han prestado en la realización de este Trabajo Final de Grado. En primer lugar, a mi directora del proyecto, Laura Igual, por su excelente trabajo, el conocimiento ofrecido con generosidad, la asistencia y su apoyo constante, así como su confianza y paciencia, que me ha transmitido durante todo el proyecto. Ha sido un verdadero placer trabajar con ella.

A Marta Núñez, por la gran ayuda que me ha dado cuando lo he necesitado.

A mi familia y amigos, por la motivación que me han infundido durante el largo transcurso de este proyecto.

Para concluir este apartado de agradecimientos, debo precisar que, sin algunas de estas personas mencionadas, este trabajo no hubiera visto la luz.

Índice

1	INTRODUCCIÓN	1
2	DIAGRAMA DE GANTT	2
3	ANÁLISIS BIBLIOGRÁFICO	3
3.1	Bases del IMR	3
3.1.1	Análisis de fMRI	4
3.2	Análisis de Componentes Independiente	5
3.2.1	ICA para los datos fMRI	5
3.2.2	Tipos de ICA	6
3.3	Group ICA y Regresión dual	7
3.3.1	Propiedades de la Regresión Dual	9
3.4	Redes en estado de reposo	9
4	METODOLOGÍA	11
4.1	Sistema de clasificación	11
4.2	Extracción de las características	12
4.2.1	Análisis de Componentes Principales	13
4.3	Clasificadores	14
4.3.1	K vecinos más cercanos	14
4.3.2	Máquinas de vectores de soporte	15
4.4	Test	16
4.4.1	Estrategia de validación	16
4.5	Selección de características	17
4.6	Librerías	17
4.6.1	FSL	17
4.6.2	Python	20
4.6.3	MATLAB	23
5	DATOS	24
5.1	Adquisición de los datos	24
5.2	Organización de los datos	24
6	EXPERIMENTOS Y RESULTADOS	26
6.1	Fase de prueba	26
6.2	Correlación entre las RSN del Group ICA y las RSN estándar	27

6.3	Selección del clasificador	27
6.4	Prueba de clasificación	28
6.4.1	Prueba con 100 sujetos	29
7	CONCLUSIONES	33
7.1	Objetivos realizados	33
7.2	Mejoras posibles	33
8	REFERENCIAS	34

Lista de figuras

FIGURA 1: DIAGRAMA DE GANTT CON TODAS LAS TAREAS	2
FIGURA 2: EN REPOSO	3
FIGURA 3: ACTIVADO	3
FIGURA 4: LOS DIFERENTES PLANOS CREADOS POR LOS EJES CEREBLALES. FIGURA EXTRAÍDA DE	4
FIGURA 5: COMPARACIÓN DE 2 ENFOQUES <i>GROUP ICA</i> : 1) <i>ICA</i> INDIVIDUAL CON EL CUAL SE PUEDEN HACER CORRELACIONES O <i>CLUSTERING</i> ENTRE LOS INDIVIDUOS. 2) TEMPORAL <i>ICA</i> CON EL CUAL SE EXTRAEN PATRONES DE TIEMPO Y ESPACIALES CONJUNTOS. [2]	6
FIGURA 6: ILUSTRACIÓN DE LA ECUACIÓN DE DESCOMPOSICIÓN DE LOS DATOS. EN ELLA SE PUEDE OBSERVAR CÓMO ESTÁ COMPUESTA LA MATRIZ QUE ES EL CONJUNTO DE RS-FMRI. LA MATRIZ COMPUESTA POR COMPONENTES Y TIEMPOS, ES DE DONDE SE SACA LOS PATRONES TEMPORALES, Y LOS MAPAS ESPACIALES SON LOS PATRONES ESPACIALES EXTRAIDOS.[8]	7
FIGURA 7: EL PROCESO DE REGRESIÓN DUAL ES APLICADO PARA OBTENER LA INFORMACIÓN ESPECÍFICA DE LOS SUJETOS (SUBJECT-SPECIFIC INFORMATION). PRIMERO, OBTENEMOS LAS SEÑALES TEMPORALES ESPECÍFICAS DE CADA SUJETO (SUBJECT –SPECIFIC TIME COURSES) Y DESPUÉS LOS MAPAS ESPACIALES ESPECÍFICOS DE CADA SUJETO (SUBJECT-SPECIFIC SPATIAL MAPS). IMÁGENES EXTRAÍDAS DE [4]	8
FIGURA 8: REPRESENTACIÓN DE LAS 20 REDES EN ESTADO DE REPOSO O TAMBIÉN LLAMADAS LAS RSN ESTÁNDAR [4]	10
FIGURA 9: DIAGRAMA DE FLUJO DE UN SISTEMA DE CLASIFICACIÓN, SEPARADO POR PASOS. TAMBIÉN PODEMOS DIFERENCIAR EL CONJUNTO DE TRAIN Y DE TEST, DIFERENCIADO POR EL COLOR Y EL NÚMERO DE PASOS A REALIZAR.	11
FIGURA 10: REPRESENTACIÓN DE LAS FRONTERAS DE LOS DATOS ILUSTRADOS EN BASE LA K-NN [10].	14
FIGURA 11: REPRESENTACIÓN GRÁFICA DE LA FRONTERA CREADA POR UN ALGORITMO SVM, DONDE PODEMOS OBSERVAR EL HIPERPLANO CREADO Y LOS EJEMPLOS QUE SIRVEN PARA CREAR LOS VECTORES DE SOPORTE [10].	15
FIGURA 12: DIAGRAMA DEL FUNCIONAMIENTO DEL K-FOLD CROSS-VALIDATION. EN ESTE CASO LOS DATOS SON DIVIDIDOS EN 5-FOLDS.	16
FIGURA 13: REPRESENTACIÓN DE LAS FRONTERAS CON LOS MISMOS CONJUNTO DE DATOS CON LOS CLASIFICADORES: LINEAL, RBF, POLINOMIAL. [10]	21
FIGURA 14: VISTA DE LA COMPONENTE NÚMERO 27.	32

Lista de tablas

TABLA 1: DIVISIÓN POR SEXO ENTRE LOS 819 SUJETOS QUE HAN REALIZADO 4 RS-FMRI	24
TABLA 2: A PARTIR DE LOS 819 SUJETOS DIVIDIDOS POR SEXO SUBDIVIDIMOS POR GRUPOS DE EDAD	24
TABLA 3: CONTIENE EL NÚMERO DE COMPONENTES (DIMENSIÓN DEL SUBESPACIO), EL TAMAÑO FINAL DE LA MATRIZ UNA VEZ REALIZADO EL GROUP-ICA, EL TIEMPO QUE TARDA EN REALIZAR ICA Y DR, LA SUMA DE LOS TIEMPOS ANTERIORES Y EL TAMAÑO DEL DATASET UTILIZADO PARA HACER GROUP ICA.	26
TABLA 4: CORRELACIÓN ENTRE LAS 20 RSNS ESTÁNDAR Y LAS 164 RNS DEL GROUP ICA	27
TABLA 5: PRECISIÓN OBTENIDA DE CADA CLASIFICADOR UTILIZADO CON LAS 20 RSN MÁS CORRELACIONADAS CON BISWAL.	28
TABLA 6: RANKING DE LAS 164 COMPONENTES QUE HAN OBTENIDO UNA PRECISIÓN MAYOR A 0.5.	¡ERROR! MARCADOR NO DEFINIDO.
TABLA 7: RANKING DE LAS 20 RSNS MÁS CORRELACIONADAS CON LAS RSN ESTÁNDAR	30
TABLA 8: SENSIBILIDAD Y ESPECIFICIDAD DEL CLASIFICADOR EN LA COMPONENTE 27.	31

1 Introducción

En el campo de la neurociencia¹, una poderosa herramienta que está siendo explotada es la utilización de Imágenes de Resonancias Magnéticas Funcionales (*fMRI*, del inglés *functional Magnetic Resonance Imaging*)[1][2].

Con esta metodología se extrae información funcional del cerebro, es decir, qué están haciendo nuestras neuronas en cada momento. En las imágenes obtenidas se puede ver como nuestro cerebro está funcionando, en otras palabras, se puede ver cómo trabajan los conjuntos de neuronas y cómo se comunican entre ellas. Así se pueden extraer las redes neuronales. Estas redes en estado de reposo, cuando un sujeto no está realizando ninguna tarea específica, reciben el nombre de redes neuronales en estado de reposo (RSN, del inglés Resting State Networks).

Las imágenes fMRI están formadas por una gran cantidad de información. El análisis, por lo tanto, se convierte en una tarea muy compleja. En este trabajo utilizamos el Análisis de Componentes Independientes (ICA, del inglés Independent Component Analysis) [3], con el cual podemos extraer las RSN [5] y entender las estructuras espacio-temporales de la señal.

Para comparar un grupo de sujetos utilizamos la regresión dual [4], con la cual extraemos la información individual de cada sujeto. Esta representación de cada sujeto puede ser información discriminativa con la que podemos compararlos entre sí.

El presente proyecto pretende realizar un estudio para encontrar diferencias entre los sexos. No nos basaremos en las características anatómicas, sino que nos basaremos en las imágenes fMRI en estado de reposo. Además, se analiza si la reducción de los datos obtenidos dificulta o favorece la búsqueda de diferencias.

En el apartado 2 se puede ver la organización, siguiendo un diagrama de Gantt, de este proyecto. A continuación, en el apartado 3 es donde veremos la descripción de algunos temas relacionados. Después, en el apartado 0, describiremos la metodología usada. En el siguiente capítulo, pasaremos a la fase de pruebas (apartado número 6) y, finalmente, en el apartado 7, presentaremos las conclusiones.

¹ Ciencia que estudia a estructura, el desarrollo y la funcionalidad del sistema nervioso.

2 Diagrama de Gantt

En este apartado expondremos detalladamente cómo se ha organizado este proyecto. En primer lugar, se muestra el listado de las tareas que se han realizado, especificando cuánto tiempo se ha empleado para cada una de ellas.

En la Figura 1, se presenta el Diagrama de Gantt, donde se enumeran todas las tareas realizadas, resaltando los distintos apartados que las contienen: Documentación teórica; FSL; Librerías e Implementación; Prueba y resultados y Documentación; además de las semanas que han sido necesarias.

De este diagrama se puede extraer qué apartados son los más costosos de tiempo:

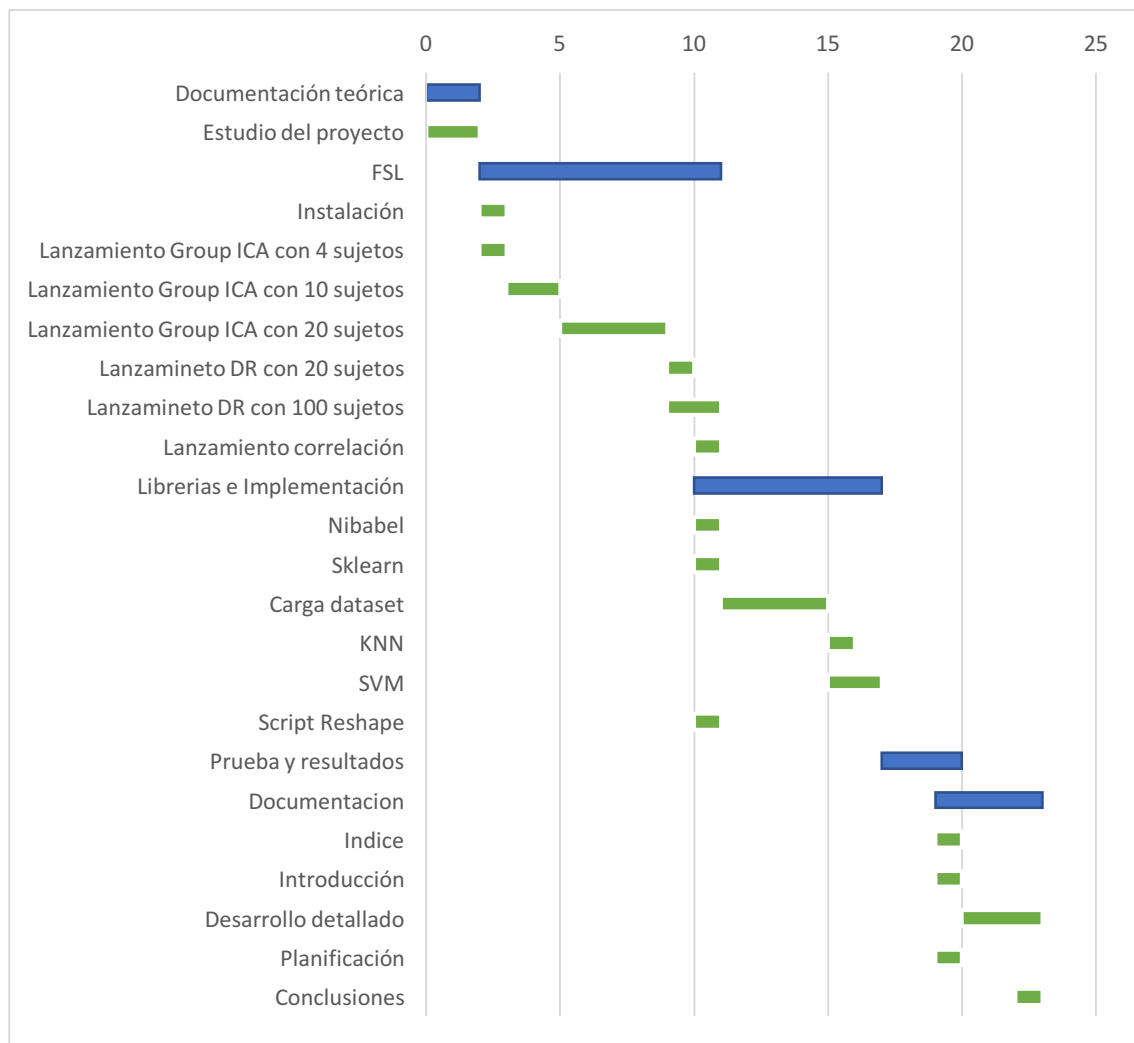


Figura 1: Diagrama de Gantt con todas las tareas

3 Análisis bibliográfico

3.1 Bases del IMR

La resonancia magnética es una técnica no invasiva mediante la cual podemos ver la estructura del cerebro, así como medir su actividad. Con este instrumento obtenemos imágenes formadas de 100 a 200.000 piezas de información cerebral —en caso que sea una sucesión de imágenes se tomarán cada segundo, aproximadamente. Este instrumento puede realizar dos tipos de técnicas:

1. Estructural: la imagen de la anatomía cerebral. Este tipo de imagen es de gran definición.
2. Funcional: es la captura de la actividad cerebral en partes del cerebro.

Los resultados obtenidos a través de una de estas dos técnicas se utilizan con el propósito de estudiar el órgano más importante del sistema nervioso encargado de coordinar todas nuestras funciones vitales: el cerebro.

Las imágenes son obtenidas mediante el contraste BOLD, el nivel de oxígeno en sangre (*Blood Oxygen Level Dependent*), ilustrado en las Figuras 2 y 3. Este es un método que mide la proporción de la transición de hemoglobina oxigenada a desoxigenada en la sangre. En otros términos, captura la demanda metabólica para la activación neuronal.

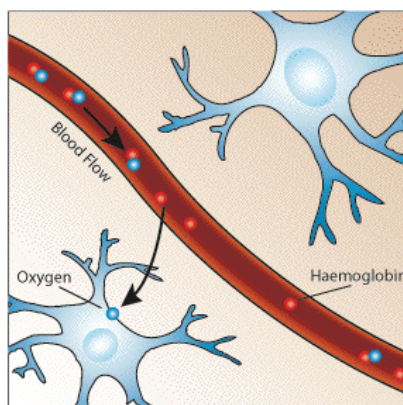


Figura 2: En reposo

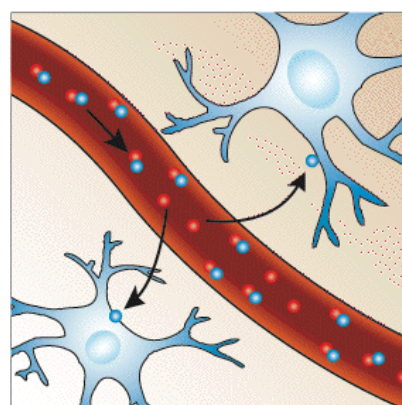


Figura 3: Activado

Figura 1 y 2: Tras la activación, el oxígeno es demandado y extraído por las células, luego, aumenta el nivel de desoxihemoglobina en la sangre. Esto se compensa por el aumento del flujo sanguíneo, acción que provoca un aumento de oxihemoglobina.

Figuras extraídas de [9]

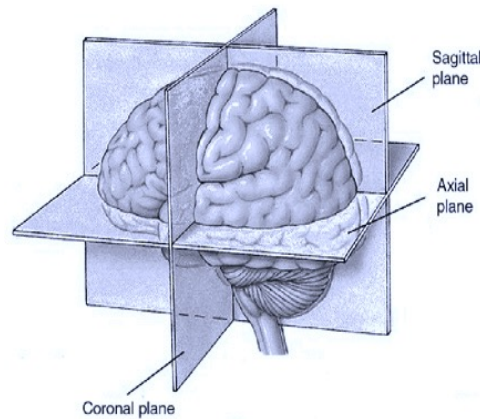


Figura 4: Los diferentes planos creados por los ejes cerebrales. Figura extraída del blog [F.Traver](#)

Como se ha comentado anteriormente, una imagen está formada por un conjunto de piezas llamadas vóxel. Un vóxel es una medida volumétrica que abastece un espacio 3D en el cerebro. Cada uno de ellos localiza y representa, a través de una asociación numérica, la intensidad de la demanda de oxígeno en una zona determinada.

Como podemos observar en la Figura 4, están representados los tres ejes cerebrales: el eje coronal, correspondiente a un corte vertical ortogonal a la cara; el eje sagital que, al igual que el anterior, es un corte vertical paralelo a la cara; y el eje axial, representado por un corte horizontal paralelo a los ojos.

3.1.1 Análisis de fMRI

Hay dos enfoques en la adquisición de resonancias magnéticas funcionales (fMRI - *functional Magnetic Resonance Imaging*):

- En tarea (t-fMRI, del inglés *task – fMRI*) son una serie de fotografías hechas de un sujeto cuando este está realizando una tarea, p.ej.: escribir en el ordenador.
- En reposo (rs-fMRI, del inglés *resting state – fMRI*) son una sucesión de imágenes capturadas de un sujeto cuando este no está realizando ninguna tarea específica, es decir, en estado de reposo.

En este proyecto se ha centrado en utilizar las rs-fMRI. Hay tres métodos principales de análisis de este tipo de datos:

1. Localización: su uso más frecuente es la localización de áreas cerebrales en respuesta a una cierta tarea. En esta regla se utiliza el Modelo General Lineal (GLM, del inglés *General Linear Model*). Este es un método estadístico para evaluar las relaciones entre las tareas realizadas y la actividad cerebral. Dicho método puede ser usado para estimar la respuesta cerebral con un suceso

simple y así comparar diferentes tipos de sucesos, con el fin de evaluar las correlaciones entre actividad y comportamiento.

2. Conectividad: describe cómo interactúan las diferentes regiones del cerebro. Se distinguen tres tipos:
 - a. Anatómica: describe cómo están conectadas físicamente las diferentes regiones cerebrales.
 - b. Funcional: que compara las correlaciones extrayendo un patrón temporal de la activación neuronal.
3. Predicción: empleo de la actividad general de un sujeto para percepciones, comportamientos o salud. En este apartado se utilizan clasificadores de patrones mediante estadística y técnicas de aprendizaje automático.

3.2 Análisis de Componentes Independiente

El Análisis de Componentes Independientes (ICA, del inglés *Independent Component Analysis*) es una herramienta para descubrir características de un conjunto de datos espacio-temporales. Este modelo asume que las observaciones son una mezcla lineal de fuentes independientes.

Considerando un vector aleatorio M-dimensional observado, por ejemplo $x = [x_1, x_2, \dots, x_M]^T$, ICA lo descompone en 2 partes: la parte espacial y la temporal, como se puede ver en la ecuación (1) [3]:

$$x = A s \quad (1)$$

Grosso modo, ICA descompone una señal muy compleja en partes simples.

Para resolver existen varios algoritmos, como Infomax, FastICA y diagonalización aproximada [3].

3.2.1 ICA para los datos fMRI

En el área de neuroimagen se aplica ICA usando diferentes configuraciones de los datos:

- ICA espacial (sICA, del inglés *spatial ICA*) cuando el número de vóxeles es mayor al número de puntos en el tiempo.
- ICA temporal (tICA, del inglés *temporal ICA*), en caso contrario, cuando el número de puntos en el tiempo son mayores al número de vóxeles.

ICA puede ser utilizado para descubrir, ya sea espacial o temporalmente, componentes independientes. El objetivo del análisis de componentes fMRI es, entonces, factorizar la matriz de datos en un producto de un conjunto de señales temporales, y un conjunto de patrones espaciales.

Sin embargo, mayoritariamente los estudios de neuroimagen se basan en sICA, ya que actualmente las rs-fMRI contienen más número de vóxeles que número de puntos en el tiempo. sICA encuentra regiones del cerebro que no se superponen, temporalmente coherentes.

3.2.2 Tipos de ICA

Para realizar un análisis sobre un grupo de sujetos ICA tiene 2 tipos según el enfoque dado:

- El primer tipo, Figura 5 (1) consiste en la aplicación de ICA por individuos. La ventaja de este enfoque es la extracción de las características espaciales y temporales únicos, con la desventaja que los componentes de datos ruidosos no son descompuestos de la misma forma si se rehace este enfoque para el mismo sujeto.
- La ventaja del segundo tipo, Figura 5 (2) —concatenación temporal— es que permite señales temporales únicas por cada sujeto pero un mapa espacial único.

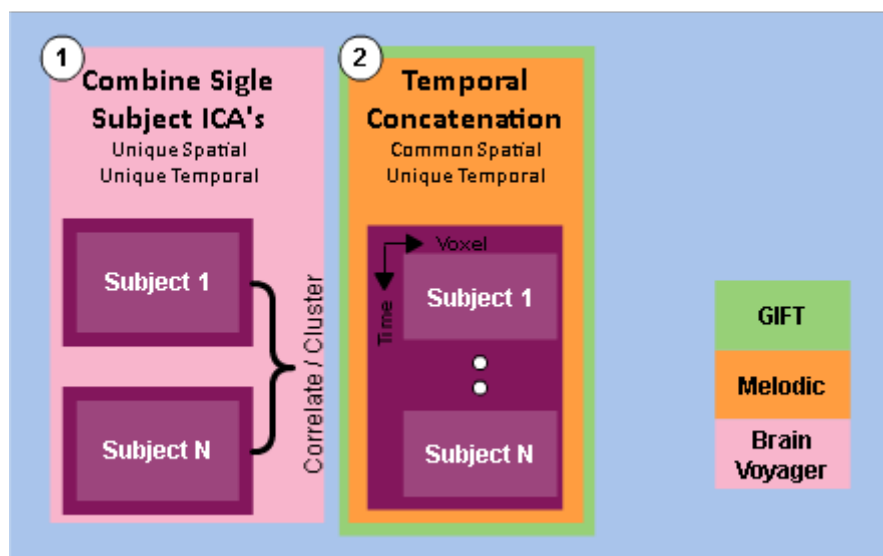


Figura 5: comparación de 2 enfoques *Group ICA*: 1) ICA individual con el cual se pueden hacer correlaciones o *clustering* entre los individuos. 2) Temporal ICA con el cual se extraen patrones de tiempo y espaciales conjuntos. [2]

3.3 Group ICA y Regresión dual

ICA se ha utilizado ampliamente con el fin de identificar los patrones de conectividad funcional en estado de reposo. Estas técnicas han demostrado ser útiles en la caracterización de estas fluctuaciones funcionales a nivel de sujeto individual. En este apartado se comentará un método de análisis que combina ICA sobre múltiples sujetos con la regresión dual para estimar las diferencias individuales en la conectividad funcional.

El procedimiento estándar sigue tres etapas:

1. Concat-ICA:

Múltiples conjuntos de datos fMRI son concatenados temporalmente y se forma una matriz *Espacio 2D * Tiempo * nSujeto*. ICA se aplica con el fin de identificar los patrones a gran escala de la conectividad funcional.

En la Figura 6 ilustra la ecuación de descomposición de los datos. En ella, se puede observar cómo está compuesta la matriz, que es el conjunto de rs-fMRI. La matriz compuesta por componentes y tiempos es de donde se sacan los patrones temporales, y los mapas espaciales son los patrones espaciales extraídos.

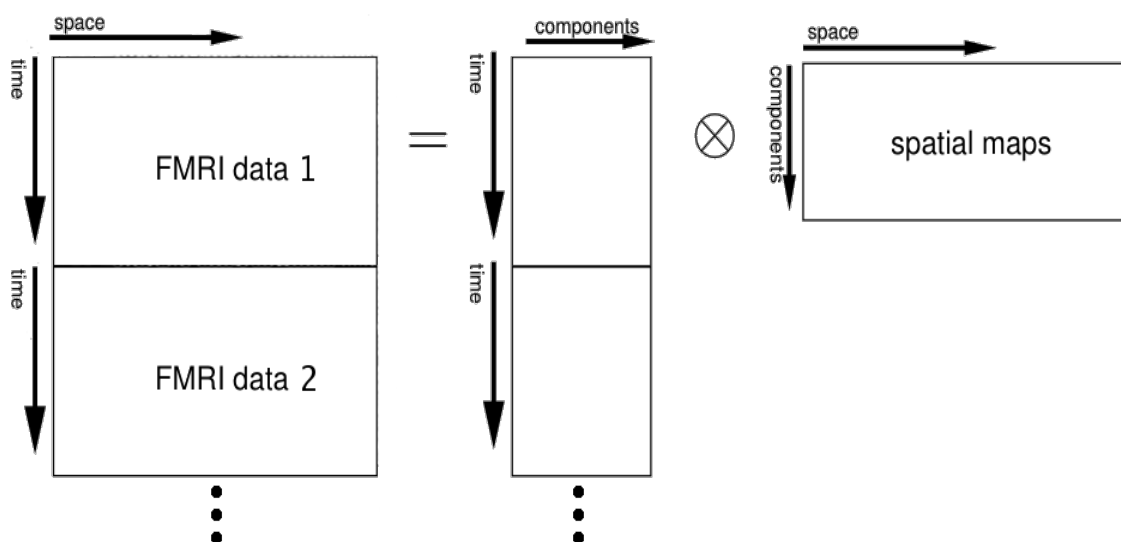


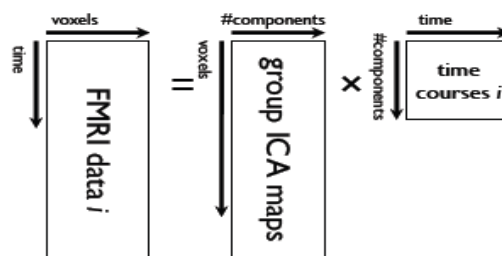
Figura 6: Ilustración de la ecuación de descomposición de los datos. En ella se puede observar cómo está compuesta la matriz que es el conjunto de rs-fMRI. La matriz compuesta por componentes y tiempos, es de donde se saca los patrones temporales, y los mapas espaciales son los patrones espaciales extraídos.[8]

2. Regresión dual:

En esta etapa, es donde se identifican los mapas espaciales y los patrones temporales en cada uno de los N sujetos individuales. Para cada conjunto de datos, es decir, para cada rs-fMRI o sujeto se realizan los siguientes pasos:

- En el primer problema de regresión, se utilizan los mapas espaciales extraídos a nivel de grupo como el conjunto de regresores espaciales en un GLM. Su objetivo es encontrar una dinámica temporal asociada con cada mapa a nivel de grupo.
- La normalización de estas señales temporales a la unidad de varianza
- El segundo problema de regresión es la utilización de estas señales temporales, resultado del primer problema de regresión, como el conjunto de regresores temporales en un GLM, para encontrar mapas específicos de los sujetos (*specific-subject maps*).

1r problema de regresión



2n problema de regresión

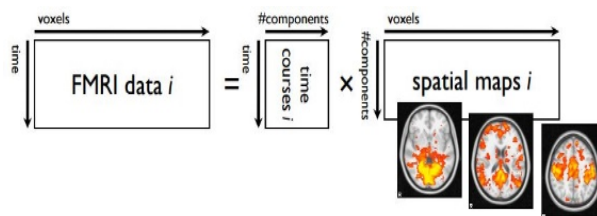


Figura 7: El proceso de regresión dual es aplicado para obtener la información específica de los sujetos (*subject-specific information*). Primero, obtenemos las señales temporales específicas de cada sujeto (*subject-specific time courses*) y después los mapas espaciales específicos de cada sujeto (*subject-specific spatial maps*). Imágenes extraídas de [4]

En la Figura 7 se observa el proceso de regresión dual que es aplicado para obtener la información específica de los sujetos (*subject-specific information*). Primero, obtenemos las señales temporales específicas de cada sujeto (*subject-specific time courses*) y después los mapas espaciales específicos de cada sujeto (*subject-specific spatial maps*).

3. Inferencias:

En la última etapa, los diferentes conjuntos de mapas espaciales se recogen a través de matices 4D y son analizados mediante métodos no paramétricos (p.ej.: permutaciones). Los mapas espaciales resultantes entre sujeto y diferencias grupales.

3.3.1 Propiedades de la Regresión Dual

La regresión dual está relacionada con los enfoques alternativos para el análisis del grupo:

- No se basa en una única ubicación sino que integra la información temporal de los datos fMRI distribuidos a través de las RSN². Los resultados de la descomposición inicial definen las regiones o las redes de interés que son relevantes a nivel de grupo para la población en particular.

En lugar de utilizar la regresión individual, emplea la regresión lineal múltiple; por lo tanto, puede estimar diferencias significativas. Además, este método también puede abordar la interacción dinámica entre las redes.

- El enfoque presentado estima la dinámica espacial y temporal a nivel de sujeto con los datos originales. Este nuevo enfoque estima que los mapas espaciales se encuentran necesariamente dentro del espacio definido por los principales espacios propios de los sujetos específicos iniciales. Por lo tanto, la comparación final entre sujetos se hace dependiente de las etapas iniciales de reducción de sujetos específicos.

3.4 Redes en estado de reposo

Las redes en estado de reposo (RSN, del inglés *Resting State Network*) no fueron descubiertas hasta el 1995. Primer, en los estudios donde se utilizaban rs-fMRIs surgían fluctuaciones de frecuencia baja, es decir, perturbaciones de la señal menores a 0,1 Hz. Durante años, estas variaciones fueron ignoradas como ruido de la señal ya que no estaban relacionadas con ninguna tarea en particular.

² Este concepto se explicará en el apartado 3.4.

No fue años hasta más tarde cuando se confirmó la existencia de patrones en el funcionamiento del cerebro que estaban correlacionados temporalmente en algunas áreas del cerebro. Estas señales de baja fluctuación se pudieron descomponer en:

- Señales cardiovasculares (0,6 – 1,2 HZ)
- Señales respiratorias (0,1 – 0,5 Hz)
- Rango de baja señal específica (0.01 -0.08 Hz)

Después de una gran recopilación de datos pudieron extraer zonas separadas, físicamente, pero funcionalmente conectadas, que seguían un patrón durante el estado de reposo. Estas 20 RSNs fueron aceptadas por la comunidad médica, las cuales son llamadas RSN estándar (Figura 8) [4][5].

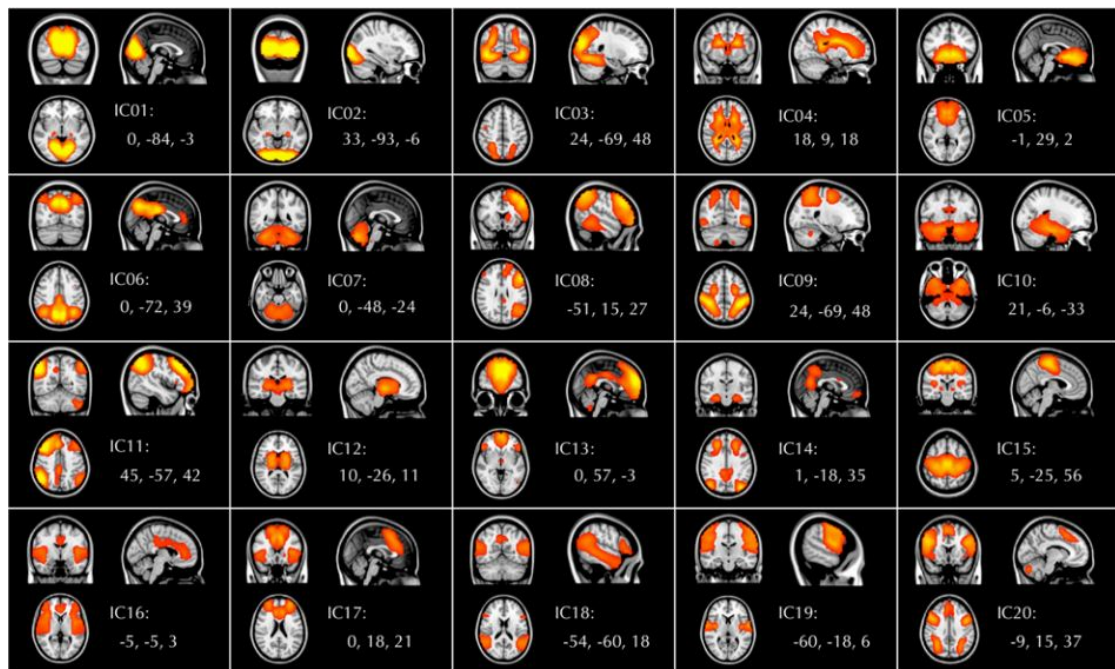


Figura 8: Representación de las 20 Redes en Estado de Reposo o también llamadas las RSN estándar [4]

4 Metodología

En este proyecto se desarrolla un sistema de clasificación aplicado a datos de rs-fMRI para predecir el sexo de un conjunto sujetos. Además, se quiere descubrir cuáles son las características más discriminativas para este objetivo de clasificación.

Esta metodología difiere de la sección 3.3 en la etapa de las inferencias, ya que es donde los sujetos son comparados para extraer resultados.

4.1 Sistema de clasificación

La metodología seguida en el sistema de clasificación que proponemos contiene los pasos que se pueden ver en la Figura 9 el diagrama de flujo de un sistema de clasificación, separado por pasos. También se puede diferenciar el conjunto de *train* y el de test.

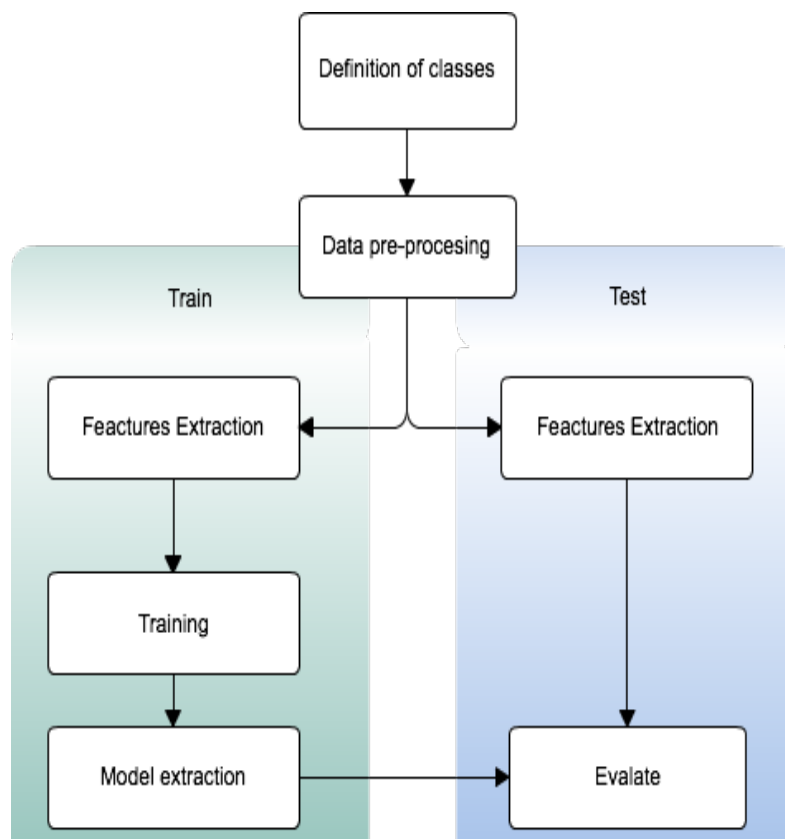


Figura 9: Diagrama de flujo de un sistema de clasificación, separado por pasos. También podemos diferenciar el conjunto de *train* y de test, diferenciado por el color y el número de pasos a realizar.

Los detalles de los pasos del procedimiento son los siguientes:

1) Definición de las clases:

En nuestro caso consideramos dos clases: hombre y mujer del conjunto de sujetos del *Human Connectome Project* (HCP). Los sujetos seleccionados, para el aprendizaje y el test del clasificador, han sido escogidos en la franja de 26 a 30 años. (Ver más detalles sobre los datos en la sección 5).

2) Pre-procesamiento de los datos:

Los datos proporcionados por el proyecto *Human Connectome* ya vienen pre-procesados, es decir, las rs-fMRI han sido procesadas para corregir el movimiento y ángulo de posición de la cabeza de cada sujeto.

3) Extracción de las características:

En este apartado podemos diferenciar 2 partes. Primero realizamos un Group ICA, con el que extraeremos las características comunes a los sujetos, y a continuación, realizamos un Regresión Dual, método que nos permite extraer las características individuales de cada sujeto. Los detalles de este paso se presentarán en el apartado 4.2.

4.2 Extracción de las características

Para la extracción de las características, procederemos a realizar un *Group ICA*, o más concretamente, un espacial *Group ICA* con concatenación temporal. Obtenemos una matriz de tamaño *Tiempo Concatenado * nSujeto × Espacio2D*, como se ilustra en la Figura 6.

Primero se realiza un PCA con el cual reduciremos el número de vóxeles. A continuación, se procederá a aplicar el ICA, que factorizará la matriz en el producto de dos matrices. Como se formaliza en la ecuación (2) se muestra la descomposición de M_i , que es la matriz perteneciente del sujeto i , en dos matrices A, matriz de mezcla, y S, matriz de fuentes, que en este caso contiene los mapas espaciales. Estos mapas espaciales son las características de los n sujetos que pierden detalles de forma individual.

$$M_{t_i \times V} = A_{t \times R} \cdot S_{R \times V} \quad (2)$$

A continuación, realizaremos la regresión dual para definir las componentes individuales a partir de las componentes del grupo. La matriz X, que es la rs-fMRI de un sujeto, se transforma siguiendo el siguiente esquema:

1º problema de regresión:

$$X_{V \times t_i}^T = S_{V \times R}^T \times G_{R \times t_i} \quad (3)$$

2º problema de regresión:

$$X_{t_i \times V} = G_{t_i \times R} \times SS_{R \times V} \quad (4)$$

Este proceso consiste en resolver dos problemas consecutivos de regresión lineal, formalizado en las ecuaciones (3) y (4), y es aplicado para obtener la información específica de los sujetos (*subject-specific information*) de forma individual. Resolviendo el primer problema de regresión obtenemos los *subject-specific time courses*, es decir, conseguimos los patrones temporales, y después los *subject-specific spatial maps*, en otras palabras, los patrones espaciales.

Finalmente, el vector de características del sujeto i-ésimo es la concatenación de la información de las r RSNs, se puede ver en la ecuación (5).

$$S_i = \{RSN_1, RSN_2, \dots, RSN_r\} \quad (5)$$

Este vector de características resultante contiene muchas dimensiones $r \times v$. Este número de vóxeles es variable, es decir, este número cambia dependiendo del número de sujetos que se haya utilizado en el Group-ICA; en nuestro caso es de 227.441.

Como solución al problema de alta dimensión seguiremos el procedimiento:

1. Una máscara que estará compuesta por las r RSNs extraídas del Group ICA, que a su vez se escogerán los valores diferentes a 0, para elegir los vóxeles donde se haya encontrado un patrón espacio-temporal.
2. Las r RSNs de cada sujeto, o los mapas espaciales específicos de los sujetos, se filtrarán por una máscara y se eliminarán los valores iguales a 0.

Además, probaremos de reducir esta alta dimensión añadiendo el análisis de componentes principales explicado en el siguiente apartado.

4.2.1 Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA, en inglés *Principal Component Analysis*) es una técnica de aprendizaje automático no supervisado.

Esta técnica de reducción de la dimensión reduce el número de variables. Las nuevas componentes principales o factores serán una combinación lineal de las variables de origen.

4.3 Clasificadores

Durante la realización de este proyecto se han empleado los clasificadores que explicaremos a continuación. Estos clasificadores son métodos de aprendizaje supervisado, es decir, que se pueden dividir en 2 partes:

1. La fase de aprendizaje que es donde se diseñará el clasificador utilizando el algoritmo y los datos de entrenamiento.
2. La fase de test o clasificación aplicará el clasificador para predecir la clase del conjunto test.

4.3.1 K vecinos más cercanos

Este método (K-NN, del inglés *K-Nearest Neighbors*) es un clasificador no parametrizado, que calcula la distancia de una muestra con el resto. A continuación, se escogerán las K observaciones más cercanas al individuo, y se asignará la clase con más presencia en estas K observaciones.

Este método es un tipo de *Lazy Learning* en inglés, pues rompe con el esquema explicado anteriormente (sección 4.3), ya que en la fase de aprendizaje solo recolectará la información, y es en la fase de clasificación donde crea un modelo por cada ejemplar

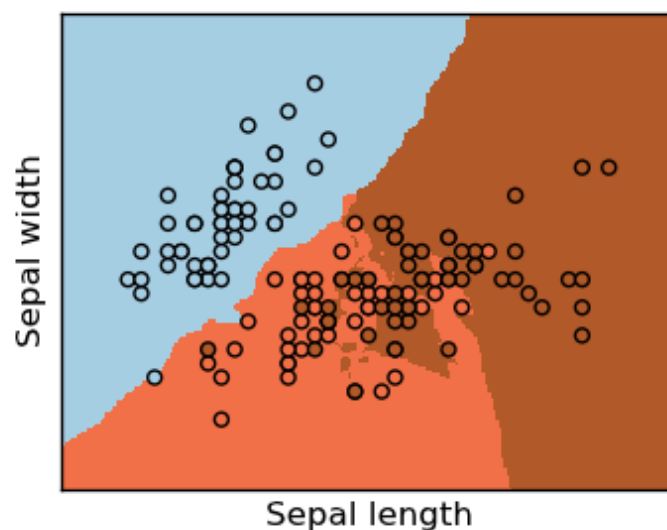


Figura 10: Representación de las fronteras de los datos ilustrados en base la K-NN [10].

introducido. La distancia que más se utiliza para la clasificación es la distancia euclídea que se define de la siguiente manera:

Dado los ejemplares $x_1 = (a_1, b_1)$ y $x_2 = (a_2, b_2)$ la ecuación (6) muestra la distancia entre estos:

$$d_e(x_1, x_2) = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2} \quad (6)$$

En la Figura 10 podemos observar cómo este algoritmo crea las fronteras de cada clase.

4.3.2 Máquinas de vectores de soporte

Este método SVM (del inglés *Support Vector Machines*) mapea los datos de entrada a un espacio mayor de características para encontrar el hiperplano que separe y maximice el margen entre las clases. Pertenece a la categoría de los clasificadores lineales ya que mueve estos hiperplanos a espacios de características de mucha dimensionalidad (realizado por el núcleo o *kernel*) y traza una frontera lineal.

Al contrario de la mayoría de métodos de aprendizaje que se centran en minimizar el error del entrenamiento, los SVM busca un hiperplano de separación equidistante a los ejemplos de clases diferentes y más cercanos entre ellos. Como resultado, obtiene un

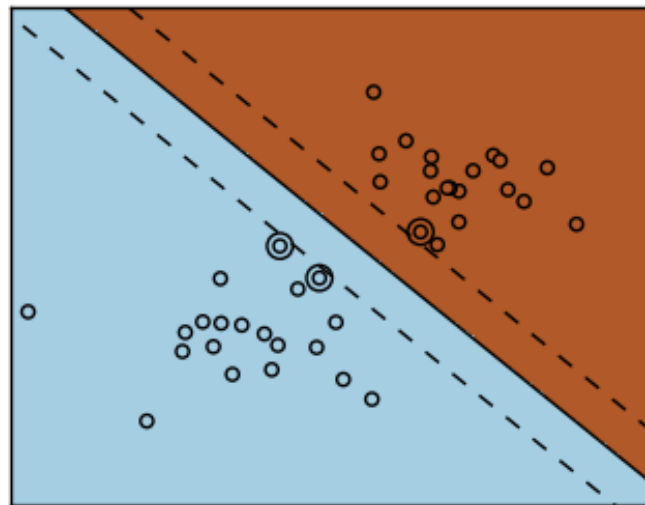


Figura 11: Representación gráfica de la frontera creada por un algoritmo SVM, donde podemos observar el hiperplano creado y los ejemplos que sirven para crear los vectores de soporte [10].

margen máximo. A estos ejemplos más cercanos y de diferente clase se les considera para definir la frontera y reciben el nombre de *vectores de soporte*. Este método tiene gran capacidad de generalización, así evita el sobreajustamiento³.

En la Figura 11, se puede ver la frontera creada por este algoritmo para los datos introducidos de distinto color. Además, el hiperplano óptimo representado por la línea continua y por las líneas discontinuas son los vectores de soporte asociados a un ejemplo que maximiza la distancia entre ellos.

4.4 Test

4.4.1 Estrategia de validación

La validación cruzada (*Cross-validation* en inglés) es una técnica que separa el conjunto de muestras que tenemos en dos subconjuntos, el de aprendizaje (*train set*) y el de prueba (*test set*). Esta técnica evalúa los resultados de un análisis estadístico o análisis de aprendizaje automático.

En este proyecto, utilizaremos el *K-fold cross-validation* para analizar los resultados obtenidos. Este método divide los datos en K subconjuntos. Uno de los subconjuntos se utilizará para el *test set* y todos los demás para el aprendizaje. El proceso consiste en ir cambiando el *test set* y el *train set* K veces. En la Figura 12 podemos ver, de forma esquemática, el funcionamiento de este validador.



Figura 12: Diagrama del funcionamiento del *K-Fold cross-validation*. En este caso los datos son divididos en 5-folds.

Imagen extraída del blog de [Bryce Christensen](#)

³ Consecuencia cuando un modelo se ajusta mucho a los datos de entrenamiento.

4.5 Selección de características

Por último, en este proyecto, queremos saber cuál de las RSN es la más discriminativa para diferenciar hombre y mujer. Por ello, entrenaremos tantos clasificadores como RSN tengamos utilizando los vectores de características correspondientes. La RSN más discriminativa será la que consiga mejor precisión en esta tarea clasificación.

4.6 Librerías

En este apartado explicaremos el *software* y los lenguajes de programación utilizados. Además, se hará una pequeña explicación del código para el desarrollo de este proyecto.

4.6.1 FSL

FSL, del inglés *FMRIB Software Library*, es una biblioteca de herramienta para el análisis de fMRI realizada en el centro FMRIB de Oxford. Esta biblioteca se puede ejecutar en todas las plataformas, sin embargo, la única que no utiliza una máquina virtual para su ejecución es Mac OS.

Nos ofrece una infinitud de funcionalidades de neuroimagen, de las cuales solo describiré los utilizados durante el proyecto.

1. Para poder realizar el Group ICA hemos utilizado la instrucción `melodic`. Para poder utilizar esta instrucción es necesario disponer de 10 veces más espacio de lo que ocupan todos los archivos para analizar:

- a. Instrucción utilizada:

```
>>melodic -i path_to_data.txt --nobet --tr=2.0 --Ostats -o  
ICA_result --report -v
```

- b. Morfología:

```
>>melodic -i <filename> <options>
```

- c. Análisis morfosintáctico:

- Parámetros:

-i, -in: su traducción es input, después de este parámetro, que es obligatorio, se le introduce los nombres de archivo para realizar ICA. Estos nombres pueden ser: un solo archivo, listado separado por comas, o un archivo que contenga todas las rutas.

--nobet: este argumento desactiva la realización de la función BET⁴ antes de hacer ICA.

--tr=2.0: es para indicar a la variable TR, cada cuantos segundos se realiza un corte en la fMRI.

--Ostats: este argumento hace la creación de mapas *thresholded* y mapas de probabilidad

-o, --outdir: para indicarle la carpeta donde se guardará el *output*

--report: este argumento opcional genera una web donde saldrá el informe de Melodic

-v, --verbose: para encender los mensajes de diagnóstico.

- Inputs:

Path_to_data.txt: este archivo lo creamos nosotros, aquí dentro habrá todas las rutas de los archivos que se quieren utilizar para aplicar ICA

ICA_result: este es el ejemplo del nombre de la carpeta donde se guardará la salida generada por el proceso ICA. También añadiremos que esta carpeta se crea en el directorio donde se llama la instrucción.

2. Para poder realizar el DR hemos utilizado la instrucción `dual_regression`:

a. Instrucción utilizada:

```
>> dual_regression melodic_IC.nii 1 design.mat design.con 1  
DR_result_50 `cat path_to_data.txt`
```

b. Morfología:

```
>> dual_regression <group_IC_maps> <des_norm> <design.mat>  
<design.con> <n_perm> <output_directory> <input1> <input2>  
<input3> ....
```

c. Análisis morfosintáctico:

- Parámetros:

⁴ BET en inglés *Brain Extraction Tool*, es una herramienta para la extracción del cerebro eliminando tejido que no pertenece al cerebro de una imagen.

- `group_IC_maps_4D`: Imagen 4D contenedora de los mapas IC espaciales, por ejemplo, el fichero `melodic_IC.nii`, que es la salida del análisis group ICA.
- `des_norm`: los valores que puede coger este parámetro es 0 o 1, pero se recomienda 1. Este parámetro sirve para normalizar los timecourses usados en la segunda etapa de la regresión.
- `design.mat`: matriz diseñada para poder realizar la regresión a todos los sujetos.
- `n_perm`: número de permutaciones, puesto que 1 tiene como salida `tstats`.
- `output_directory`: es el directorio donde se guardarán todos los archivos de salida y *logfiles*.
- `<input1> <input2> ...`: Lista de todos los sujetos preprocesados.

- Inputs:

`cat path_to_data.txt`: esta instrucción la utilizamos como input para no poner la entrada sujeto a sujeto y sea automático. Devuelve un *string* con todos los elementos guardados en el fichero `path_to_data.txt`

3. Para poder correlacionar las RSN estándar con las RSN extraídas por el Group ICA

a. Instrucción utilizada:

```
>> fslcc -t 0.2 melodic_IC.nii Resampled_Biswal.nii
```

b. Morfología:

```
>> fslcc -t <threshold (default=0.1)> <first-NIFTI> <second-NIFTI>
```

c. Análisis morfosintáctico:

- Parámetros:

`threshold`: mínimo valor necesario para mostrarse en la salida

`<first-NIFTI>/<second-NIFTI>`: archivos que contienen las RSNs que se quieren comparar

4.6.2 Python

Python es un lenguaje interpretado con el objetivo de hacer una sintaxis legible. A continuación, se explicarán las librerías de Python que se han utilizado en el proyecto

4.6.2.1 Nibabel

Nibabel es una librería que nos permite la lectura y escritura de los formatos más comunes de los archivos utilizados en la neuroimagen.

Su instalación es muy simple, solo hay que tener en cuenta los requisitos, que son:

- *Python* 2.6 o posterior
- *NumPy* 1.5 o posterior, es el paquete fundamental para la computación científica con Python
- *SciPy* es un ecosistema basado en Python de software de código abierto para las matemáticas, la ciencia y la ingeniería.
- *PyDICOM*
- *Python Imaging Library* (PIL) añade capacidades de procesamiento de imágenes para el intérprete de Python. Esta biblioteca es compatible con muchos formatos de archivo, y proporciona potentes capacidades de procesamiento de imágenes y gráficos
- *Nose* 0.11 o posterior para hacer correr los tests
- *Sphinx* para construir la documentación

4.6.2.2 Scikit-learn

Scikit-learn, abreviado SKlearn, es una biblioteca libre de Python para facilitar la implementación de algoritmos de clasificación, regresión, *clustering*, entre otros. El lenguaje para poder utilizar esta librería es Python. Nos centramos en los clasificadores K-NN y SVM, y para la validación el K-fold utilizados en el proyecto.

4.6.2.2.1 K-NN en Scikit-learn

Como se ha comentado en la sección 4.3.1, este algoritmo busca los K vecinos más cercanos. En la instanciación de este algoritmo se ha modificado K, que es la medida de la muestra de los vecinos más cercanos que queremos mirar.

```
KNeighborsClassifier(n_neighbors)
```

Donde:

- `N_neighbors`: que especifica el número de vecinos por defecto es igual a 5, y hemos utilizado esta configuración para clasificar.

El problema de encontrar la K óptima es que si la K es muy pequeña se ajusta demasiado al modelo, por lo tanto, el ruido de la muestra o las muestras que pueden ser ruido afectan al modelo. Por el contrario, si K es muy grande, es decir, si K coge el valor del número de muestras, rompe con la filosofía de este algoritmo, y si el número de muestras, de cada clase utilizada, son iguales o similares no podrá clasificar bien, ya que la función de densidad que permite clasificar dará un resultado muy similar entre las clases empleadas.

Por este motivo, no quise modificar el valor de K , ya que se debería hacer un *K-fold cross-validation* para escoger que K otorga más precisión. Esto supone un alto coste computacional, y además esa K solo serviría para un volumen de muestras concreto.

4.6.2.2.2 SVC en Scikit-learn

El clasificador de vectores de soporte (SVC, del inglés *Support Vector Classifier*), como se ha explicado en el punto 4.3.2, es el *kernel* o núcleo es el encargado de generar el hiperplano. A continuación, les mostramos morfología de la instanciación de la clase:

SVC (kernel, degree, gamma)

En la parametrización solo se muestran las variables que hemos ido modificando durante la realización de este proyecto:

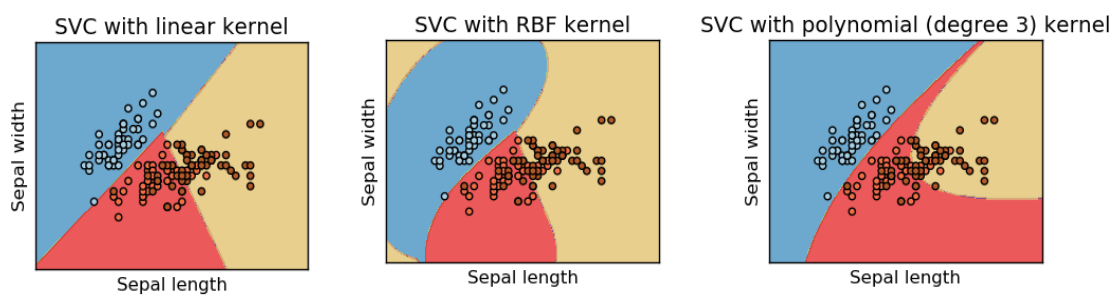


Figura 13: Representación de las fronteras con los mismos conjunto de datos con los clasificadores: Lineal, RBF, Polinomial. [10]

- *Kernel*: especifica el tipo de núcleo para ser utilizado en el algoritmo. Nosotros hemos hecho servir 'poly', 'linear', 'rbf'⁵
- *Degree*: indicar el grado de la función polinomio núcleo ('poly'). Hemos utilizado la que viene por defecto, es decir, de 3º grado. Cuando se utiliza otro tipo de *kernel* que no sea polinomial este parámetro será ignorado.
- *Gamma*: solo es utilizado para el *kernel* RBF, cuando se utiliza el automático es $\frac{1}{\text{numero de características}}$. Este valor nos indica que influencia tiene un ejemplo del train, es decir, con valores bajos indica lejanía y con valores altos proximidad. Aparte del automático se ha utilizado una gamma con un valor de 0,5.

En la Figura 13 se puede ver una representación de los tres clasificadores utilizados. Estas fronteras se han creado con el mismo conjunto de datos.

4.6.2.2.3 K-flod en Scikit-learn

Esta clase nos facilita la creación de un iterador. Este iterador divide los índices de las muestras en dos grupos, el de entrenamiento y el de test. Como su propio nombre indica hace K particiones en los índices proporcionados para poder hacer la validación del K-fold, explicada en el punto 4.4.1. La instanciación de esta:

KFold(n, n_fold, shuffle)

Donde:

- *n* es el número de ejemplares utilizadas.
- *n_fold* el número de carpetas a dividir la muestra, en el proyecto *n_fold* obtiene el valor de $\frac{\text{numero de muestras}}{4}$. De esta forma la precisión obtenida aumenta en un cuarto de punto, es decir, que cada acierto que obtiene es un 0,25.
- *shuffle* es un *booleano* para indicar si queremos barajar los índices utilizados, en otras palabras, no se aceptarán índices consecutivos para crear cada carpeta. En nuestro caso este valor lo activamos.

4.6.2.2.4 PCA en Scikit-learn

Esta clase nos permite reducir el número de variables, explicado en la sección 4.2.1. En su instanciación el parámetro *n_components* es el número de componentes al que

⁵ Función de base radial (RBF, del inglés *Radial Basis Functions*) es una función real cuyo valor depende sólo de la distancia del origen.

queremos reducir. Este parámetro te da la opción de que sea nulo y, como consecuencia, el número de dimensiones es el mínimo entre número de ejemplos y número de características.

PCA(n_components)

4.6.3 MATLAB

MATLAB es un lenguaje de computación técnica de alto nivel y un entorno interactivo para desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. Con MATLAB, se pueden resolver problemas de cálculo técnico más rápidamente que con lenguajes de programación tradicionales, tales como C, C++ y FORTRAN.

Se puede usar MATLAB en una amplia gama de aplicaciones que incluyen procesamiento de señales e imágenes, comunicaciones, diseño de sistemas de control, sistemas de prueba y medición, modelado y análisis financiero y biología computacional.

4.6.3.1 SPM

En este proyecto, al trabajar con las RSN estándar se ha necesitado de la librería SPM implementada en MATLAB. Esta librería es un conjunto de herramientas para el análisis de secuencia de datos de imágenes cerebrales. Normalmente se emplea para el manejo de los múltiples formatos de neuroimagen.

5 Datos

5.1 Adquisición de los datos

La base de datos facilitada por el HCP (<http://www.humanconnectome.org/>) está formada por 970 sujetos de los cuales 819 tienen 4 rs-fMRI. Prestaremos especial atención a estos 819 individuos, ya que han completado el seguimiento, y los subdividiremos por sexo como podemos ver en la Tabla 1.

NUMERO DE SUJETOS	
MUJERES	453
HOMBRES	366

Tabla 1: División por sexo entre los 819 sujetos que han realizado 4 rs-fMRI

Estos conjuntos de sujetos están repartidos en 4 grupos de edad. En la Tabla 2 se puede observar la cantidad de sujetos que hay en los diferentes grupos de edad y sexo:

INTERVALOS DE EDAD	MUJERES	HOMBRES
22-25	70	105
26-30	198	152
31-35	182	106
36+	3	3

Tabla 2: A partir de los 819 sujetos divididos por sexo subdividimos por grupos de edad

5.2 Organización de los datos

Como hemos comentado anteriormente, los sujetos seleccionados tienen 4 rs-fMRI. Estas 4 rs-fMRI están realizados durante 2 sesiones, en relación con la BBDD: las 2 primeras rs-fMRI están realizadas en la primera sesión, y las 2 siguientes en la sesión restante. Cada una de ellas es la captura de la actividad cerebral, en estado de reposo, durante 15 minutos aproximadamente. Dentro de cada sesión, alternaron la dirección de la codificación de datos, que son: derecha-izquierda (RL) e izquierda-derecha (LR).

La relación comentada anteriormente, la división de las rs-fMRI en sesiones, se ve plasmada en la distribución de los datos a la hora de descargarse los datos, es decir, las rs-fMRI están agrupadas en las sesiones, y estas, divididas en dos grandes bloques:

- 1.El primer bloque que contiene las dos primeras rs-fMRI, ocupa 5216,93 GB.

- 2.La segundo, con las siguientes rs-fMRI correspondientes, ocupa 4975,56 GB

La suma de los bloques de información produce que el volumen de datos facilitado por el proyecto sea inmenso, en total 10192,49 GB, en otras magnitudes, 10TB aproximadamente.

En los dos bloques, los archivos de un sujeto en la BBDD están formados por:

1. La carpeta *MNINonLinear/Results/*, dónde se encuentra: Las dos rs-fMRI del bloque correspondiente y los datos del preprocesamiento.

2. *release-notes/*, que es el directorio donde se encuentra la versión, las notas y la licencia de los datos.

3. El directorio *T1W/Results*, es el directorio donde se encuentran los escáneres estructurales de cada sujeto.

Toda la información de un sujeto oscila entre los 5,5GB y los 7GB aproximadamente. Sin embargo, nos percatamos que los archivos que necesitamos ocupan más o menos 1GB, es decir, que las rs-fMRI de un sujeto pesa 1GB aproximadamente.

6 Experimentos y resultados

6.1 Fase de prueba

Para empezar, procedemos a una fase de pruebas donde realizaremos la primera toma de contacto con la metodología (*Group* ICA + Regresión dual) y con la librería FSL. Se observa como oscilan diferentes valores –tamaño de los datos, tiempo de ejecución de los diferentes procesos, etcétera— para diferentes cantidades de sujetos, con la finalidad de poder tener una aproximación de tiempo para la prueba final.

Como se puede observar en la

Tabla 3, tenemos los datos recogidos en tres pruebas. En esta tabla podemos ver el tiempo de duración de los dos procesos (*Group* ICA y Regresión Dual) junto al total. También el tamaño de la matriz, una vez se haya realizado el *Group* ICA, y el número de dimensiones del subespacio, que es el número de redes que ha encontrado, es decir, el número de componentes. Por último, estimaremos el tamaño del conjunto de sujetos utilizado para realizar *Group* ICA.

En las 2 primeras pruebas, la de 4 sujetos y la de 10 sujetos, contienen un número similar de RSNs. Estas RSNs se han podido observar que están formadas por datos muy ruidosos. Es el *Group* ICA de 20 sujetos, en las que estas RSN ruidosas no aparecen, que contiene algo de ruido.

Nº DE SUJETOS	4	10	20
TAMAÑO INICIAL DE LOS DATOS	1200x227441		
DIMENSIÓN SUBESPACIO	154	157	164
TAMAÑO FINAL DE LOS DATOS	616x227441	1570x227441	3280x227441
TIEMPO GICA	03:50	05:13	25:24
TIEMPO DR	00:15	00:38	15:44
TIEMPO TOTAL	04:05	05:52	41:08
TAMAÑO DATA SET	3,93GB	9,27GB	18,45

Tabla 3: contiene el número de componentes (Dimensión del subespacio), el tamaño final de la matriz una vez realizado el *Group*-ICA, el tiempo que tarda en realizar ICA y DR, la suma de los tiempos anteriores y el tamaño del *dataset* utilizado para hacer *Group* ICA.

6.2 Correlación entre las RSN del *Group* ICA y las RSN estándar

En el apartado 3.4 se ha explicado la existencia de las 20 RSN estándares, en esta sección miraremos que RSNs de la salida del *Group* ICA son más parecidas con estas 20 RSN estándar. Para obtener esta correlación tuvimos que seguir este procedimiento:

1. Se tuvo que remodelar los 2 conjuntos de RSN para que poseyeran el mismo tamaño, ya que las dos estaban formadas por dimensiones diferente.
2. Todo seguido utilizamos la instrucción `fs1cc`, la explicación del uso está en la sección 4.6.1

Se puede observar en la Tabla 4 las 20 correlaciones que hay entre los dos tipos de RSNs:

Estándar	<i>Group</i> ICA	Estándar	<i>Group</i> ICA	Estándar	<i>Group</i> ICA	Estándar	<i>Group</i> ICA
1	8	6	12	11	1	16	28
2	11	7	25	12	37	17	30
3	3	8	10	13	21	18	24
4	2	9	17	14	13	19	6
5	46	10	55	15	23	20	27

Tabla 4: Correlación entre las 20 RSNs estándar y las 164 RNS del *Group* ICA

6.3 Selección del clasificador

Como se ha explicado anteriormente, en la sección 4.3, utilizaremos tres clasificadores. Todas las experiencias realizadas han sido testeadas con las mismas variables, es decir, hemos utilizado el mismo número de sujetos, los mismos sujetos y la misma estrategia *K-fold cross-validation*.

En la Tabla 5, resumimos los resultados obtenido de las pruebas. Los índices de la primera columna de la tabla son las RSN que hemos extraído del ICA y con más correlación con las RSN Biswal. Observando los resultados, nos damos cuenta de la mala actuación de los clasificadores SVM con un *kernel* polinomial y RBL, así, que estos clasificadores son descartados. También podemos ver que es el SVM con un *kernel* lineal es el que mejor clasifica, por lo tanto, escogemos este para pruebas posteriores.

	K-NN	POLY	LINEAL	RBL 1/NºF	RBL 0.5
8	0,52	0,42	0,63	0,47	0,43
11	0,52	0,43	0,49	0,47	0,43
3	0,51	0,51	0,72	0,47	0,43
2	0,51	0,34	0,51	0,47	0,43
46	0,49	0,51	0,58	0,47	0,43
12	0,55	0,41	0,54	0,47	0,43
25	0,49	0,47	0,61	0,47	0,43
10	0,54	0,41	0,52	0,47	0,43
7	0,55	0,34	0,63	0,47	0,43
55	0,50	0,33	0,53	0,47	0,43
1	0,59	0,32	0,60	0,47	0,43
37	0,49	0,43	0,55	0,47	0,43
21	0,52	0,32	0,52	0,47	0,43
13	0,50	0,51	0,55	0,47	0,43
23	0,49	0,43	0,55	0,47	0,43
28	0,45	0,41	0,54	0,47	0,43
30	0,54	0,32	0,51	0,47	0,43
24	0,48	0,51	0,51	0,47	0,43
6	0,49	0,49	0,54	0,47	0,43
27	0,47	0,47	0,49	0,47	0,43
MEDIA	0,51	0,41	0,55	0,47	0,43

Tabla 5: Precisión obtenida de cada clasificador utilizado con las 20 RSN más correlacionadas con Biswal.

6.4 Prueba de clasificación

Una vez hemos seleccionado el clasificador que utilizaremos (SVM lineal), realizaremos una prueba considerando todas las componentes extraídas con ICA (164). También comprobaremos si la componente extraída por ICA que obtiene más precisión es una de las 20 RSN más correlacionadas con las RSN de Biswal. Además, miraremos si reduciendo la alta dimensión de los datos, con un PCA, afecta a la precisión de la clasificación, es decir, si la alta dimensión de las RSN tiene influencia en la separación de sexos.

A continuación, dividiremos esta prueba en sub-pruebas:

- 1) 164 componentes sin PCA
- 2) 164 componentes aplicando PCA
- 3) 20 componentes, que han sido las que tienen mayor correlación con las componentes de Biswal, sin PCA
- 4) 20 componentes, las que tienen mayor correlación comparando las redes de Biswal, aplicando PCA

Las pruebas se identificarán por el número de sujetos utilizados. Este identificador es el nombre de sujetos empleados y se divide en 2 partes, ya que, se utiliza la misma cantidad de hombres que de mujeres.

Las tablas que se muestran, a continuación, contienen la media de los resultados en los 10 K-Fold ordenadas de forma descendiente, es decir, creamos un ranking de las mejores medias de precisión. Las tablas han sido divididas a ambos lados: las celdas de la izquierda son las que ocupan unas posiciones más elevadas en el ranking que las celdas de la derecha.

6.4.1 Prueba con 100 sujetos

En esta prueba la muestra empleada se divide en 50 hombres y 50 mujeres. 20 de 100 sujetos han sido utilizados para la extracción de características comunes del proceso Group ICA. Como se puede ver en las Tablas 6 y 7, son los rankings de las mejores precisiones obtenidas en la prueba.

En la Tabla 6 se observa la mayor precisión que se obtiene, en las componentes, es el número 27 con un 0,683.

La componente número 27 fue correlacionada con una RSN estándar número 20 (dicha correlación se puede observar en la Tabla 4), obteniendo el mismo resultado de precisión en la clasificación con las RSN (Tabla 7).

Nº COMPONENTE	NO PCA	PCA	Nº COMPONENTE	NO PCA	PCA
27	0,683	0,683	32	0,545	0,545
4	0,675	0,675	35	0,542	0,542
40	0,66	0,66	29	0,54	0,54
9	0,62	0,62	7	0,539	0,539
13	0,603	0,603	137	0,538	0,538
2	0,6	0,6	139	0,532	0,532
26	0,596	0,596	19	0,531	0,531
20	0,592	0,592	38	0,527	0,527
46	0,592	0,592	1	0,527	0,527
39	0,591	0,591	107	0,524	0,524
42	0,591	0,591	25	0,523	0,523
18	0,589	0,589	121	0,521	0,521
23	0,589	0,589	90	0,521	0,521
76	0,58	0,58	44	0,52	0,52
33	0,568	0,568	10	0,508	0,508
8	0,564	0,564	12	0,507	0,507
15	0,562	0,562	31	0,506	0,506
14	0,558	0,558	79	0,506	0,506

111	0,557	0,557	96	0,505	0,505
50	0,556	0,556	104	0,503	0,503
17	0,549	0,55	21	0,502	0,502
5	0,548	0,548	62	0,502	0,502

Tabla 6: Ranking de las 164 componentes que han obtenido una precisión mayor a 0.5

Nº COMPONENTE	NO PCA	PCA	Nº COMPONENTE	NO PCA	PCA
27	0,683	0,683	10	0,508	0,508
13	0,603	0,603	12	0,507	0,507
2	0,6	0,6	21	0,502	0,502
46	0,592	0,592	24	0,499	0,499
23	0,589	0,589	28	0,484	0,484
8	0,564	0,564	30	0,482	0,482
17	0,549	0,55	3	0,471	0,471
6	0,547	0,547	37	0,447	0,447
1	0,527	0,527	11	0,428	0,428
25	0,523	0,523	55	0,352	0,352

Tabla 7: Ranking de las 20 RSNs más correlacionadas con las RSN estándar

En la Figura 14 se puede observar la componente 27. Esta componente de las RSN del Grupo ICA también es seleccionada como la que más se correlaciona con la componente número 20 de las 20 RSN (Tabla 4) y se puede ver similitud en la Figura 8.

Además, en dichas tablas, se puede observar la precisión obtenida aplicando PCA en el vector de características. La precisión, reduciendo el vector de características mediante el uso de PCA, no difiere de la precisión del vector de características sin ser reducido.

En esta parte del análisis hemos obtenido precisiones iguales. A partir de ahora, nos fijaremos en su sensibilidad y especificidad, he aquí su definición:

- La sensibilidad nos indica la capacidad de nuestro clasificador para dar como casos positivos los casos realmente negativos.
- la especificidad nos indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente positivos.

En nuestro caso estas definiciones se pueden ajustar como:

- La sensibilidad nos indicará la capacidad de nuestro clasificador para dar casos que son hombres en casos que son mujeres.
- La especificidad nos indicará la capacidad de nuestro estimador para dar casos que son mujeres los casos que son hombres

En la Tabla 8 se puede observar la sensibilidad y la especificidad obtenidos del clasificador para la RSN número 27. Podemos concluir de esta tabla que nuestro clasificador tiende a catalogar como mujer los casos que son hombres.

	164 COMPOENTES		20 RSN MÁS CORRELACIONADAS	
	NO PCA	PCA	NO PCA	PCA
SENSIBILIDAD	0,576	0,576	0,576	0,576
ESPECIFICIDAD	0,436	0,436	0,436	0,436

Tabla 8: Sensibilidad y Especificidad del clasificador en la componente 27.

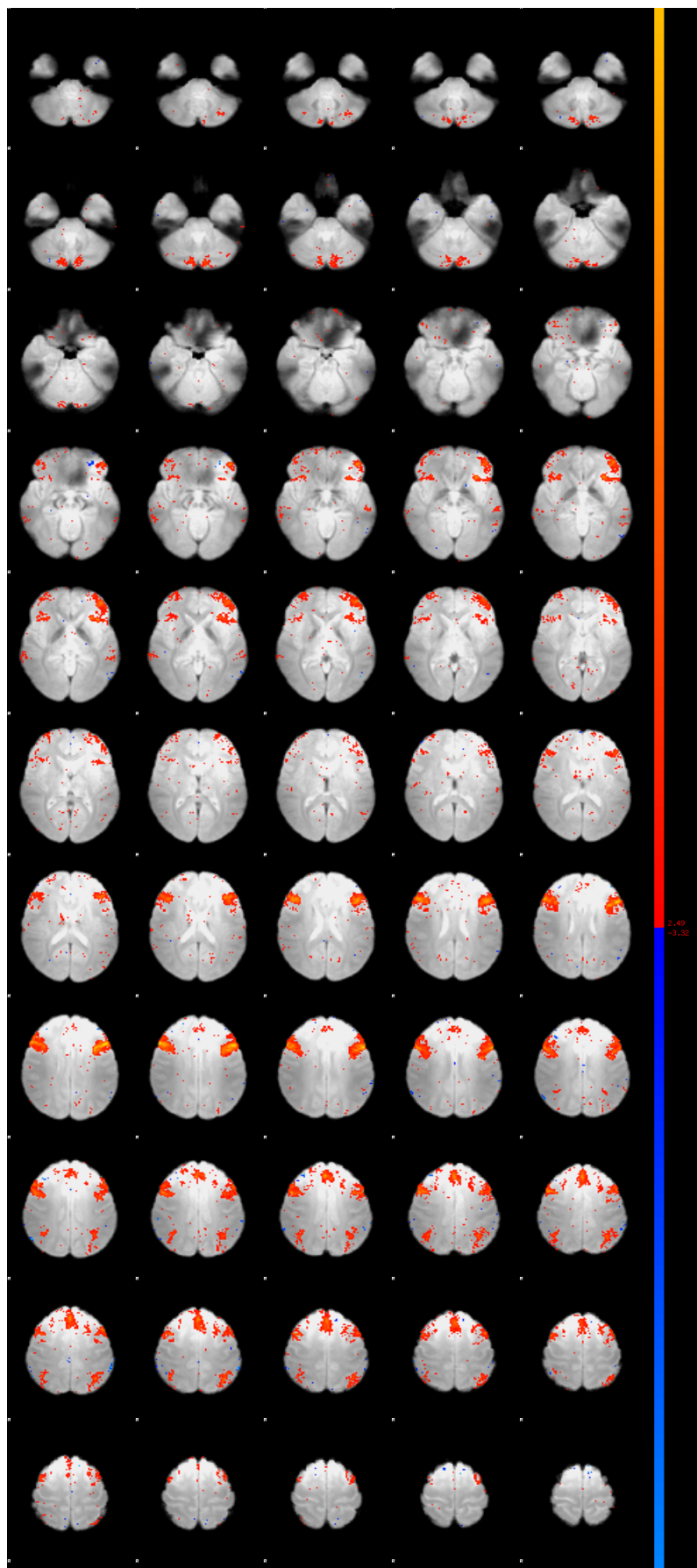


Figura 14: Vista de la componente número 27.

7 Conclusiones

Al inicio de este trabajo, las expectativas albergaban la posibilidad de encontrar alguna o algunas RSN/s que pudieran discriminar ambos sexos. Esas expectativas se han cumplido, ya que se ha logrado alcanzar este objetivo utilizando técnicas que actualmente están en funcionamiento. En este apartado, se detallará qué objetivos hemos llevado a cabo con éxito, y cuáles pueden ser mejorados.

7.1 Objetivos realizados

En el marco de la extracción de características comunes y específicas de sujeto, se ha logrado entender toda la base de la materia que sustenta el FSL software. Su dificultad no ha sido ningún impedimento a la hora de realizar este proyecto.

En cuanto a la implementación, se ha intentado elaborar un código limpio y fácil de entender, asimismo, que hiciera la función de clasificar. En este sentido, esta codificación diáfana posee la finalidad de que otros investigadores puedan usarlo y/o mejorarlo.

Los resultados obtenidos son bastante prometedores. Hemos encontrado una RSN que clasifica los sujetos de diferente sexo en base a nuestros datos con una precisión de 0,683.

7.2 Mejoras posibles

Una de las mejoras más importantes que se debe realizar es un *Group ICA* con un gran número de sujeto. Se ha intentado aplicar un *Group ICA* con 50 sujetos, pero los requisitos que se requerían no eran satisfechos, ya que esta tarea es algo complicada y se necesita un ordenador con altas prestaciones para poder realizarla con éxito.

Otra mejora posible es probar esta metodología en los diferentes grupos de edades para obtener si esta RSN discriminativa obtenida se mantiene a lo largo de los rangos de edad, o en caso contrario, la RSN va cambiando según el rango de edad.

También queda pendiente discutir los resultados con neuropsicólogos y hacer algún test estadístico para ver si el resultado es estadísticamente significativo.

8 Referencias

- [1] [Online]. Disponible: <https://www.coursera.org/course/fmri1>
- [2] [Online]. Disponible: <https://www.coursera.org/course/fmri2>
- [3] V. D. Calhoun, J. Lui, and T. Adali, "A review of group ica for fMRI data and ICA for joint inference of imaging, genetic, and ERP data", *NeuroImage*, vol. 45, no. 1, pp S163-S1972, 2009
- [4] Group comparison of resting-state FMRI data using multi-subject ICA and dual regression, 2009
- [5] B.B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckman, J.S. Adelstein, R.L. Buckner, S. Colcombe et al., "Toward discover science of human brain function", *Proceedings of the National Academy of Science*, vol. 107, no. 10, pp 4734-4739, 2010
- [6] [Online]. Disponible: <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL>
- [7] [Online]. Disponible: <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/DualRegression>
- [8] [Online]. Disponible: <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC>
- [9] [Online]. Disponible: http://users.fmrib.ox.ac.uk/~stuart/thesis/chapter_3/section3_3.html
- [10] [Online]. Disponible: <http://scikit-learn.org/stable/index.html>
- [11] [Online]. Disponible: <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>