UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

FACULTAD DE
INGENIERÍA
UDELAR

Diagnóstico del Trastorno depresivo mayor (MDD) utilizando imágenes de resonancias magnéticas funcionales (fMRI) y aprendizaje automático

Sebastián Volti Diano
Agustina Sierra Lima

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad de la República, en cumplimiento parcial de los requerimientos para la obtención del título de Ingeniería en Computación.

Tutor:
Pablo Rodríguez Bocca

Tribunal:
Aiala Rosá
Raquel Sosa
Juan Kalemkerian

Montevideo, Uruguay
Diciembre 2022

Índice

Resumen

En este trabajo estudiamos el trastorno depresivo mayor a través de imágenes de resonancia magnéticas de distintos pacientes, capturadas con el individuo en total estado de reposo. Buscamos encontrar una metodología que ayude a la predicción del trastorno, a partir de la actividad cerebral entre distintas regiones del cerebro. Evaluamos distintos modelos de aprendizaje automático supervisado, trabajando con diferentes agrupaciones de predictores, con el fin de caracterizar a cada individuo de la mejor manera posible, construyendo grafos y explotando sus propiedades, cálculos de correlación, matrices, etc. Los métodos que presentaron mejores resultados fueron los basados en el clasificador *Random Forest*, construyendo árboles de decisión a partir de los atributos de cada individuo, obteniendo un grafo para cada paciente a partir de su resonancia magnética, y utilizando métricas globales de la red, funciones de correlación adecuadas, y también atributos demográficos de cada paciente. Se alcanzaron resultados que superan el 70 % de precisión en todas las métricas utilizadas para medir la perfomance de nuestros modelos, tanto en *accuracy*, *recall*, *precision* y *f1-score*.
Utilizamos un conjunto de datos de magnitud considerablemente grande para esta temática (más de 2400 pacientes), analizado en la actualidad por diversos artículos científicos. Logramos buenos resultados a pesar de trabajar con un universo de pacientes grande y heterogéneo, donde habitualmente la mayoría de los artículos científicos que trabajan en esta temática, suelen perder bastante precisión, y se limitan a trabajar con un universo de pacientes reducido.

1. Introducción

El trabajo se centra en el análisis de la correlación entre distintas zonas de interés del cerebro [1], extraídas mediante resonancias magnéticas funcionales *(fMRI)* en estado de reposo [2]. Se trabaja con un conjunto de datos que contiene individuos que presentan depresión, y por otro lado individuos de control que no padecen el trastorno. Se pretende encontrar una metodología que ayude a la predicción del Trastorno Depresivo Mayor *(MDD)* [3], a partir de la actividad cerebral entre distintas zonas de interés. Las regiones del cerebro son definidas mediante agrupaciones de secciones de menor tamaño, conocidas como *voxels* [4], y la relación entre regiones se mide utilizando cierta función de correlación.

La tarea de extraer información de las *fMRI* es un problema muy complejo, estudiado por diversos artículos científicos, donde no hay consenso sobre la mejor técnica para extraer información relevante de las mismas. Tampoco partimos de bases científicas sólidas, que establezcan que ante mayor o menor actividad del cerebro en determinada región, se supone mayor o menor predisposición a presentar el trastorno estudiado, por lo que intentar predecir depresión partiendo solamente de una *fMRI* se torna un problema muy difícil de atacar. Otro aspecto que vuelve al problema desafiante es entender las resonancias en si mismas, no solo la imagen resultante del estudio, sino también cuáles son los parámetros específicos que utiliza cada tipo de *scanner* (aparato electrónico que realiza la resonancia), y luego cuáles son los mejores parámetros para pre-procesar las *fMRI* obtenidas, tratando de reducir al máximo el ruido generado durante la resonancia, y a su vez extraer la información más relevante posible. Vale aclarar que el conjunto de datos utilizado se conforma por individuos evaluados en múltiples instituciones, por lo que se utilizan distintos *scanners* y de distinta marca y prestaciones.

Gran parte de los artículos que estudian problemas similares a partir de *fMRI* obtienen buenos resultados trabajando con conjuntos de datos de pocos individuos. Pero al incrementar la cantidad de individuos, estos estudios pierden bastante la exactitud producto de la heterogeneidad de los datos observados. En este trabajo se utiliza un conjunto de datos heterogéneo de magnitud elevada para esta problemática, formado por 2428 pacientes en total, por lo que un aspecto clave de nuestra metodología debe ser la robustez y capacidad de generalización.

Implementaremos distintos modelos de aprendizaje automático tradicionales [5],

con el fin de poder clasificar de la mejor manera posible a cada individuo. Haremos foco en modelos basados en árboles de decisión, y experimentaremos con distintas metodologías para intentar extraer los atributos o propiedades (de ahora en más llamados *features*) [6] que mejor respondan a esta temática, probando con distintas variantes y modelos capaces de adaptarse tanto a pequeñas cantidades como a grandes cantidades de *features*. También realizaremos validación de hiperparámetros sobre los modelos de mejor performance, con el fin de maximizar la capacidad de los mismos para clasificar a un individuo.

1.1. Objetivos

Con motivo de hacer foco en lo que realmente buscamos, el objetivo central de nuestro estudio es realizar clasificación binaria de pacientes a través de técnicas tradicionales de aprendizaje automático supervisado, para poder etiquetar a cada sujeto como positivo (sujeto que presenta *MDD*) o negativo (sujeto de control). Además de nuestro objetivo central, también buscamos comprender y abordar de buena manera el mundo de las *fMRI* y la neurociencia, lo que a priori representa una curva de aprendizaje realmente grande.

1.2. Estructura del informe

El informe se divide en distintos capítulos, en el Capítulo 2 comenzaremos comentando el estado del arte de nuestra problemática, dando un marco teórico general a la temática, incluyendo detalle de los datos disponibles y su preprocesamiento. También se presentan brevemente los clasificadores de aprendizaje automático utilizados en el trabajo, así como los artículos previos relacionados más relevantes. Posteriormente, en el Capítulo 3 nos centraremos en describir la metodología propuesta para dar solución al problema de estudio, describiendo el motivo de cada decisión tomada y el paso a paso de la solución implementada. Luego, en el Capítulo 4, documentamos los distintos experimentos realizados, y los resultados de los distintos métodos y variantes evaluadas. Estos resultados fueron la guía principal para definir la metodología final presentada. Por último, finalizaremos el informe con el Capítulo 5, destinado a las conclusiones del trabajo, y en que aspectos se podría trabajar en un futuro para seguir iterando la solución y el estudio de las *fMRI*.

2. Estado del arte

El trastorno depresivo mayor *(MDD)* es un trastorno del estado de ánimo, y se presenta cuando los sentimientos de tristeza, pérdida, ira o frustración interfieren con la vida diaria durante un largo período de tiempo. Lo padecen más de 322 millones de personas alrededor del mundo [7].

Existen diversos estudios donde se vincula la existencia de este trastorno *(MDD)* con la actividad neuronal en el cerebro [8, 9, 10]. En estos estudios se incluyen metodologías para evaluar dicha actividad, y en muchos casos se desarrollan técnicas predictivas para inferir la existencia de depresión a partir de dicho comportamiento neuronal [11, 12, 13]. Gran parte de estos estudios se basan en el análisis de imágenes de resonancia magnética funcional *(fMRI)*, de donde se obtienen atributos relacionados a la conectividad de distintas zonas del cerebro utilizando técnicas estadísticas, para luego volcar estos atributos a técnicas de aprendizaje automático y lograr así predecir la presencia o ausencia de dicho trastorno [11].

A modo introductorio en este capítulo, en la Sección 2.1 nos centraremos en las imágenes de resonancia magnética funcional *(fMRI)*, explicando en detalle qué es lo que se mide en el cerebro de los individuos, cuales son los pasos más relevantes para la extracción de estas imágenes *(fMRI)* en estado de reposo y como se dividen las distintas partes del cerebro humano para poder ser analizadas en estudios de esta temática. Luego veremos algunos resultados obtenidos en distintos artículos que utilizaron imágenes *(fMRI)* para clasificar el trastorno del espectro autista *TEA* y el trastorno depresivo mayor *MDD*, detallando también la cantidad de datos utilizados y sus clasificadores.

En la Sección 2.2 nos enfocaremos en los datos que hacen a este trabajo, presentaremos los mismos, detallaremos sobre su particular heterogeneidad y luego nos centraremos en el filtrado de datos que realizamos, basándonos en un estudio científico [14] que trabaja con los mismos datos.

En la Sección 2.3 veremos el proceso de extracción de *features* a utilizar para caracterizar a un individuo, y explicaremos en detalle parte del método que se utiliza en uno de los artículos que encontramos más relevante para nuestro trabajo [11], que combina distintas estrategias de extracción y selección de *features*.

Luego en la Sección 2.4 presentaremos los distintos clasificadores con los que nos tocó trabajar y sus particularidades, para luego finalizar con la Sección 2.5 en

donde realizamos una pequeña recapitulación de los distintos artículos relevantes para nuestro estudio, que fueron citados en las distintas secciones de este capítulo.

2.1. Análisis de actividad cerebral

La resonancia magnética funcional *(fMRI)* es un estudio no invasivo mediante el cual se extraen imágenes del cerebro por resonancia magnética funcional. La misma permite observar la actividad neuronal de un individuo. El método más común de resonancia magnética funcional son las imágenes dependientes del nivel de oxigenación en sangre *(BOLD)*.

Para medir la oxigenación en sangre, la señal *BOLD* utiliza la hemoglobina, siendo la misma una proteína de los glóbulos rojos que lleva oxígeno de los pulmones al resto del cuerpo, razón por la cual esta señal mide la actividad neuronal indirectamente a través de su correlación hemodinámica [15]. Lo que ocurre es que cuando una región cerebral se activa, la misma necesita un nivel de oxígeno más alto de lo normal, y en consecuencia el flujo sanguíneo envía a dicha región una cantidad extra de oxígeno para suplir las demandas energéticas. Como esta cantidad de oxígeno es mayor, por un período corto de tiempo el nivel de oxígeno en sangre aumenta. Dicho aumento ocurre aproximadamente 5 segundos después de la activación de las neuronas y demora entre 10 y 20 segundos en llegar a su estado normal.

La interpretación precisa de la señal *BOLD* depende fuertemente de afirmar que la naturaleza de la actividad neuronal da lugar a la respuesta hemodinámica y que estos dos aspectos están vinculados, lo que se conoce como acoplamiento neurovascular [16]. La naturaleza exacta de este acoplamiento sigue siendo en gran parte desconocida, tanto en lo que respecta a la naturaleza como al origen de la señal de comunicación entre la neurona y el vaso [15].

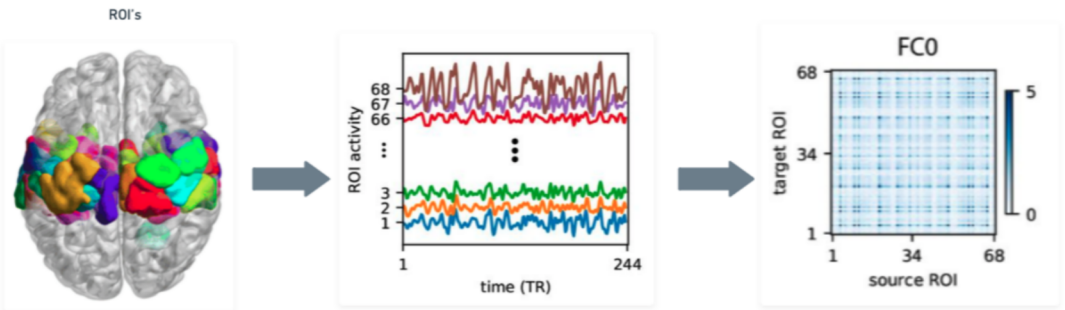
Existen diversos factores que dificultan la tarea de medir de forma precisa la señal *BOLD*, como la mencionada respuesta hemodinámica *(HR)*, el ritmo cardíaco, y hasta el movimiento del paciente a la hora de realizar la resonancia magnética, estos factores generan ruido que interfiere en la correcta identificación de actividad del cerebro. Además cabe destacar que actualmente no está claro si la *fMRI* puede diferenciar entre pequeños cambios de actividad en grandes poblaciones celulares y grandes cambios en pequeñas poblaciones [15]. Por estas razones, obtenidas las

imágenes crudas, necesitan de un procesamiento adecuado para ser utilizadas en técnicas de predicción.

Estas resonancias se pueden realizar sometiendo al paciente a ciertos estímulos o en completo estado de reposo. En nuestro caso utilizaremos solo resonancias con pacientes en estado de reposo.

En una sesión *fMRI* se extrae una determinada cantidad de volúmenes funcionales, habitualmente más de 100, a lo largo de una serie temporal, siendo cada volumen una imagen de resonancia magnética. Cada una de estas imágenes está compuesta por varios *voxels* [4]. De esta forma se puede pensar en una sesión *fMRI* como $\mathbf{v}$ volúmenes, cada uno con $\mathbf{x}$ *voxels*, donde cada uno está asociado a una serie temporal $\mathbf{t}$. Interesa entonces medir la actividad cerebral, que se traduce a niveles de oxígeno en sangre, de los distintos *voxels* a lo largo del tiempo, para poder identificar cómo se relacionan los mismos.

Figura 1: Imagen ilustrativa del proceso de extracción de una serie temporal.



Dado que el acoplamiento neurovascular puede variar en función de cómo interactúa la composición vascular de los vóxeles y este puede verse alterado de formas complejas por una enfermedad o medicación, aun no es posible separar de forma robusta la variabilidad en términos de *voxels*, región y sujeto, de la variabilidad del origen hemodinámico y la variabilidad del origen neuronal. Si bien disponer de mejores escáneres y secuencias de pulsos mejorarán la relación temporal entre la señal y el ruido, los problemas que rodean la variación de la *HR* continuarán limitando fundamentalmente las conclusiones a las que podemos llegar utilizando los datos de *fMRI* [15].

Para mitigar lo anterior, por lo general se utilizan regiones de interés del cerebro *(ROI's)*, estas pueden verse como una agrupación de *voxels*, e interesa trabajar con un conjunto de *ROI's* predefinidas. Este conjunto de *ROI's* predefinidas se agrupan de distinta forma bajo el nombre de *atlas*, siendo así un *atlas* un conjunto *ROI's*. Utilizar estas regiones de interés permite el análisis segmentado a partir de una subdivisión cerebral bajo criterios anatómicos y/o funcionales. En la Figura 1 se resume el proceso de extracción de información a partir de las *fMRI*, en la Sección 2.2 se explica este proceso en detalle.

Actualmente no está del todo claro cuál es la mejor técnica para extraer información relevante de las *fMRI*, y con ella poder predecir si un individuo posee algún trastorno cómo *MDD*, *TEA*, etc. Tampoco son del todo concluyentes los resultados obtenidos al trabajar con *fMRI*, varios estudios presentan buenos resultados al trabajar con un universo de datos acotado [17, 18], pero al trabajar con conjuntos de datos de mayor magnitud y heterogeneidad, se pierde la exactitud de los mismos [19]. Presentamos a continuación los resultados obtenidos por distintos artículos, identificando en la Tabla 1 cada uno de ellos, la cantidad de pacientes utilizados para el estudio (diferenciando pacientes que presentan algún trastorno y pacientes de control), el trastorno estudiado, el método predictivo implementado y la exactitud (*accuracy*) de los resultados obtenidos al evaluar cada modelo.

También detallamos algunos resultados extraídos del artículo que utilizamos como base para abordar el mundo de las *fMRI* y entender diversos métodos ya implementados [11], disponibles en la Tabla 2. Todos los ejemplos presentados parten del análisis de *fMRI* con el paciente en completo estado de reposo. De la Tabla 2 podemos resaltar que la mayoría de los autores utilizan como método predictivo *SVM (Support Vector Machine)*, siendo uno de los métodos de aprendizaje automático tradicionales más utilizados en esta temática (más adelante explicaremos este método). Observamos que los resultados obtenidos superan en su gran mayoría el 80 % de exactitud, aunque siempre trabajando con una pequeña cantidad de sujetos (apenas superando los 100 individuos en total en algunas ocasiones). Sin embargo, el único artículo que presenta una cantidad de sujetos un poco más grande (360 en total, sigue siendo pequeña) obtiene resultados muy bajos, en el orden de 50 % de exactitud.

Cuadro 1: Resultados obtenidos en los artículos antes mencionados y adicionalmente, el trastorno en cuestión, la cantidad de pacientes y el método de clasificación utilizado.

Artículo	Trastorno	Pacientes	Método	Accuracy
MPA	MDD	MDD=24, C=29	SVM	94 %
SNBM	MDD	MDD=30, C=30	Inverse Covariance	85 %
SNBM	MDD	MDD=19, C=19	Inverse Covariance	78 %
SNBM	MDD	MDD=30, C=30	L2-norm SVM	78 %
SNBM	MDD	MDD=19, C=19	L2-norm SVM	73 %
ABIDE	TEA	TEA=447, C=517	Leave-One-Out	60 %

* *MPA: Multivariate pattern analysis* [18]
* *SNBM: Sparse network-based models* [17]
* *ABIDE: MRI classification of autism: ABIDE results* [19]

2.2. Conjunto de datos

En este trabajo se utiliza un conjunto de datos heterogéneo de magnitud considerablemente grande para esta temática, formado por 1300 pacientes con *MDD* y 1128 pacientes de control. El *dataset* fue creado por un consorcio de 25 grupos de investigación de 17 hospitales en China, denominado *REST-meta-MDD* [32].

Un aspecto relevante a destacar es que cada centro de investigación utiliza distintos escáners para realizar la resonancia magnética, cada uno de ellos configurado con distintos parámetros de entrada, motivo por el cual la tarea de generalizar el estudio y trabajar en conjunto con todos los individuos se torna complejo. En la Tabla 3 se pueden ver algunas de las diferencias en los datos aportados por cada grupo de investigación, tales como el escáner utilizado y el tamaño de *voxel*. La mayoría de los pacientes que presentan *MDD* son de sexo femenino (826), mientras solamente encontramos 474 de sexo masculino. En las Figuras 2 y 3 se pueden ver histogramas de la distribución por sexo y por edad para pacientes de control y pacientes con *MDD* en el conjunto de datos.

Además, hay identificados 562 pacientes que tuvieron el primer episodio de *MDD*, mientras 282 son pacientes recurrentes en este trastorno. De todas formas, la información referente a episodicidad (primer episodio o recurrente) no se encuentra

Cuadro 2: Resultados obtenidos en distintos artículos que utilizan las imágenes *fMRI* con el paciente en completo estado de reposo.

Autor	Pacientes	Método	Accuracy
Yoshida et al (2017) [20]	MDD=58,C=65	PLS	80 %
Zhong et al (2017) [21]	MDD=29,C=33	SVM	91.9 %
Zhong et al (2017) [21]	MDD=46,C=57	SVM	86.4 %
Wang et al (2017) [22]	MDD=31,C=29	SVM	95 %
Sundermann et al (2017) [23]	MDD=180,C=180	SVM	45 $\sim$ 56,1 %
Bhaumik et al (2017) [24]	MDD=38,C=29	SVM	76.1 %
Zeng et al (2014) [18]	MDD=24,C=29	MMC	92.5 %
Guo et al (2014) [25]	MDD=36,C=27	NN	90.5 %
Lord et al (2012) [26]	MDD=22,C=22	SVM	99 %
Zeng et al (2012) [18]	MDD=24,C=29	SVM	94.3 %
Cao et al (2014) [27]	MDD=39,C=37	SVM	84 %

* *SVM: Support Vector Machine [28]*
* *PLS: Partial least squares regression [29]*
* *NN: Neural network [30]*
* *MMC: Maximum margin clustering [31]*

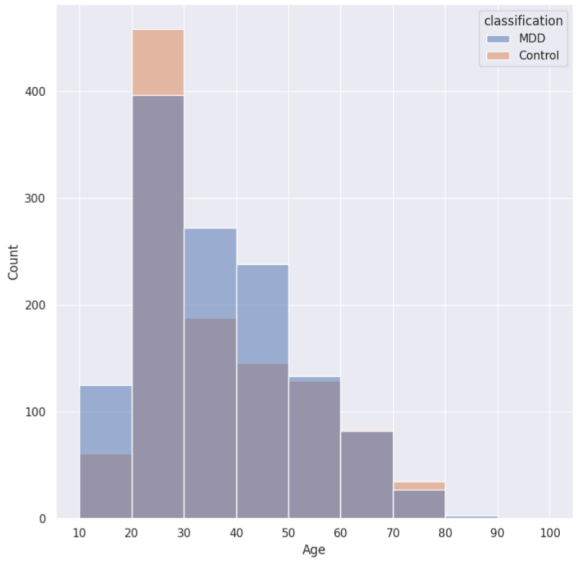
Figura 2: Distribución por sexo para pacientes de control y pacientes con *MDD* en el conjunto de datos.

Cuadro 3: Tabla descriptiva con distintas características de los datos presentados para cada una de las clínicas pertenecientes al consorcio.

Clínica	MDD	Control	Scanner	Tamaño del voxel
1	74	74	Siemens Tim Trio 3T	3.28 × 3.28 × 4.80
2	25	17	GE Excite 1.5T	4.00 × 4.00 × 4.00
3	30	30	Philips Achieva 3T	1.67 × 1.67 × 4.00
4	64	32	Siemens Tim Trio 3T	3.75 × 3.75 × 3.50
4	27	37	Siemens Magnetom Symphony 1.5T	3.75 × 3.75 × 6.25
21	50	50	Siemens Verio 3.0T MRI	3.75 × 3.75 × 4.00
4	24	24	Siemens Skyra 3T	3.75 × 3.75 × 3.50
5	31	31	GE Signa 3T	3.75 × 3.75 × 5.00
6	13	11	GE Signa 3T	3.75 × 3.75 × 5.00
17	47	44	GE Signa 3T	3.75 × 3.75 × 4.00
6	15	15	Siemens Tim Trio 3T	3.59 × 3.59 × 4.00
8	21	20	Philips Achieva 3.0 T scanner	1.67 × 1.67 × 6.00
9	38	49	GE discovery MR750	2.29 × 2.29 × 3.20
10	51	36	GE Signa 3T	3.44 × 3.44 × 4.60
11	75	75	GE Signa 3T	3.75 × 3.75 × 3.00
12	282	281	Siemens Tim Trio 3T	3.44 × 3.44 × 4.00
13	50	50	GE Discovery MR750 3.0T	3.75 × 3.75 × 4.00
14	86	70	Siemens Tim Trio 3T	3.12 × 3.12 × 4.20
15	50	33	Siemens Tim Trio 3T	3.75 × 3.75 × 4.52
16	30	20	Philips Gyroscan Achieva 3.0T	1.67 × 1.67 × 4.00
17	32	29	GE Signa 3T	3.75 × 3.75 × 5.00
19	32	30	Philips Achieva 3.0T TX	3.75 × 3.75 × 4.00
17	32	6	GE Signa 3T	3.75 × 3.75 × 4.00
18	32	31	GE Signa 1.5T	3.75 × 3.75 × 6.00
20	89	63	Siemens Verio 3T	3.75 × 3.75 × 4.00

* *1: National Clinical Research Center for Mental Disorders.*
* *3: Department of Clinical Psychology, Suzhou Suzhou Psychiatric Hospital*
* *4: The Second Xiangya Hospital of Central South University.*
* *6:Department of Psychiatry, Shanghai Jiao Tong University School of Medicine.*
* *9:Sir Run Run Shaw Hospital, Zhejiang University School of Medicine.*
* *11:Department of Psychiatry, First Affiliated Hospital, China Medical University.*
* *13:The First Affiliated Hospital of Jinan University.*
* *15:First Hospital of Shanxi Medical University.*
* *17:Department of Psychiatry, The First Affiliated Hospital of Chongqing Medical University.*
* *2:The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an Central Hospital.*
* *21:Department of Psychosomatics and Psychiatry, Zhongda Hospital, School of Medicine, Southeast University.*
* *5:MR Research Center, West China Hospital of Sichuan University.*
* *8:Department of Radiology, The First Affiliated Hospital, College of Medicine, Zhejiang University.*
* *10:Anhui Medical University.*
* *12:Faculty of Psychology, Southwest University.*
* *14:Beijing Anding Hospital, Capital Medical University.*
* *16:The Institute of Mental Health, Second Xiangya Hospital of Central South University.*
* *19:Mental Health Center, West China Hospital, Sichuan University.*
* *18:First Affiliated Hospital of Kunming Medical University.*
* *20:Department of Neurology, Affiliated ZhongDa Hospital of Southeast University.*

Figura 3: Distribución por edad para pacientes de control y pacientes con *MDD* en el conjunto de datos.



disponible para todo el *dataset*, existiendo 456 pacientes sin estos datos.

Este *dataset* fue utilizado en diversos artículos, en su mayoría para analizar y comparar distintas métricas de pacientes que presentan depresión *vs* pacientes de control, con foco en el análisis de la actividad cerebral de los pacientes, pero sin involucrar modelos de clasificación. Sin embargo, al igual que nosotros, algunos científicos se embarcaron en la tarea de clasificación, para intentar predecir si determinado paciente presenta el trastorno [33].

El artículo [33] es interesante por diversos motivos, ya que además de utilizar el mismo *dataset*, implementa métodos de aprendizaje automático tradicionales similares a los utilizados en nuestro estudio, con el fin de evaluar distintos clasificadores. Otro aspecto relevante a mencionar, es que el artículo fue publicado recientemente (Julio 2021), y fue realizado por científicos de renombre para esta área de estudio, por lo que representa una visión bastante actual del estado del arte. Si bien las similitudes con nuestro trabajo son notorias en cuanto al *dataset* utilizado y a los métodos de clasificación, los *features* utilizados difieren rotundamente, ya que no se basan en correlación de *Pearson*, construcción de grafos, ni

extracción de métricas relevantes de esos grafos. En este artículo trabajan directamente con las matrices extraídas a partir de las imágenes *fMRI*, aplicando sobre las mismas técnicas matemáticas como *t-test filter(TF)* [34] y descomposición en valores singulares *SVD* [35] para reducir la dimensión de los *features*.

Detallamos en la Tabla 4 los resultados obtenidos por el artículo en cuestión, identificando el clasificador y su correspondiente exactitud.

Cuadro 4: Resultados obtenidos en artículo [33], que utiliza el mismo dataset que nuestro estudio

Clasificador	Accuracy
Support Vector Machine	63,7 %
Regresión Logística	59,8 %
SVD + Support Vector Machine	68,9 %
SVD + Regresión Logística	65,7 %
Árboles de decisión	59,3 %
XGBoost	72,8 %

Una observación importante a tener presente es que dejaron por fuera del estudio algunos pacientes del total del *dataset* (307), trabajando entonces con 1021 sujetos con *MDD*, y 1100 sujetos de control.

La exactitud de los resultados obtenida por estos científicos parece ser muy buena para la magnitud del *dataset* utilizado, superando valores de 70 % de *Accuracy*, número pocas veces alcanzado en otros trabajos con cantidad similar de sujetos.

2.2.1. Pre-procesamiento de datos

Una vez que tenemos las *fMRI* en estado de reposo para cada paciente, es necesario preprocesar las mismas, con el fin de poder transformar cada imagen de resonancia magnética en un objeto manejable, del cual poder extraer ciertas métricas e información que nos ayude a caracterizar a cada individuo.

Además de transformar las resonancias en objetos manejables, el preprocesamiento es útil para remover el ruido generado durante la resonancia, por ejemplo

ante movimientos involuntarios de un paciente. Habitualmente se ejecutan una serie de pasos básicos para preprocesar estas imágenes, como lo son la extracción del cráneo, corrección de distorsión y corrección de movimiento.

En la actualidad existen muchas herramientas que reciben como entrada *fMRI*, y retornan como salida matrices que representan la correlación entre determinados *ROI's* del cerebro, a lo largo de una serie temporal. Algunos de los software comunmente utilizados para este tipo de preprocesamiento son *FSL* [36], *SPM12* [37], *AFNI* [38], *fMRIprep* [39], teniendo cada uno de ellos una curva de aprendizaje importante, ya que además de la imagen de resonancia magnética funcional, es necesario definir como entrada determinados parámetros relevantes para el análisis.

El *dataset* que utilizamos para este proyecto tiene la ventaja de que además de las *fMRI* para cada paciente, presenta también el *dataset* preprocesado con la herramienta *DPARSF* [40], que particularmente utiliza varios de los software detallados anteriormente, y donde obtenemos como salida una matriz de correlación para cada individuo. Lo interesante es que se cuenta con la correlación de *9 atlas* distintos para cada paciente. En nuestro estudio vamos a trabajar justamente con las matrices de correlación que brinda esta herramienta, utilizando varios de los *atlas* disponibles.

2.2.2. Filtrado de datos

A lo largo de este trabajo varias veces nos preguntamos si contar con un conjunto de datos tan heterogéneo era positivo a la hora de clasificar los sujetos. Dado que los mismos provienen de distintos grupos de investigación, donde la forma en la que fueron extraídas estas imágenes *fMRI* fue diferente, ya sea por la utilización de distintos escáners, o por parámetros que escapan a nuestro conocimiento ingenieril y pertenecen al terreno de neurociencia. Nos topamos entonces con uno de los primeros artículos en donde se utilizó el *dataset* en cuestión [14], y a pesar de enfocar el trabajo puramente al análisis de los datos en el plano de la neurociencia, ejecutan una reducción del *dataset* que nos resultó muy útil.

En este artículo se analizan los sujetos desde el punto de vista de las *Functional Brain Networks*, donde estas últimas detallan la conectividad funcional mediante el análisis estadístico de la señal *BOLD*. La conectividad funcional se define como la dependencia temporal de los patrones de activación neuronal de las dis-

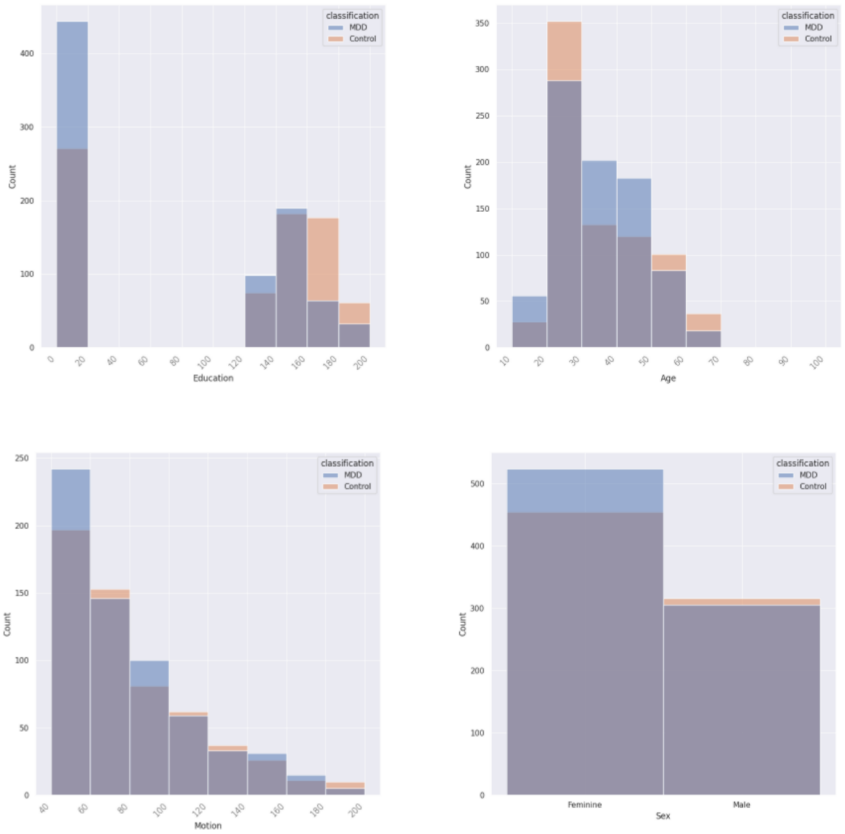
tintas regiones de interés del cerebro *(ROI's)*, y se utilizan para mostrar patrones correlacionados entre regiones cerebrales separadas [41]. Podemos decir que este artículo realizó un procesamiento y un análisis de los datos basado fuertemente en neurociencia, lo cual no nos compete a efectos de nuestro estudio.

Lo interesante entonces es que realizó un recorte en los datos, partiendo de 1300 sujetos con *MDD* y 1128 de control, y seleccionando 848 con *MDD* y 794 de control, trabajando con 17 de los 25 grupos de investigación disponibles. Algunos de los criterios de exclusión fueron por ejemplo, información demográfica incompleta, mala normalización espacial, movimiento excesivo de la cabeza durante la resonancia, grupos de investigación con menos de 10 sujetos, etc.

Logramos establecer contacto con el experto en neurociencia que realizó este trabajo, *YAN Chao-Gan*, y nos entregó exactamente cuales sujetos fueron los excluídos del *dataset*, lo que nos permitió ejecutar nuestra solución final con este conjunto de datos de mejor comportamiento y al mismo tiempo continuar con un conjunto de datos de gran tamaño para la temática.

En la Figura 4 podemos ver un cuadro con cuatro histogramas que representan algunos *features* demográficos que se utilizaron para clasificar a los individuos luego del filtrado de datos, discriminando por sujetos con *MDD* y pacientes de control 4. Aquí se agrega un nuevo *feature* llamado *Motion*, la misma representa el movimiento de la cabeza de los pacientes al momento de la resonancia.

Figura 4: Histogramas con la distribución de algunos de los *features* utilizados en el clasificador final, discriminados por individuos de control e individuos con *MDD*.



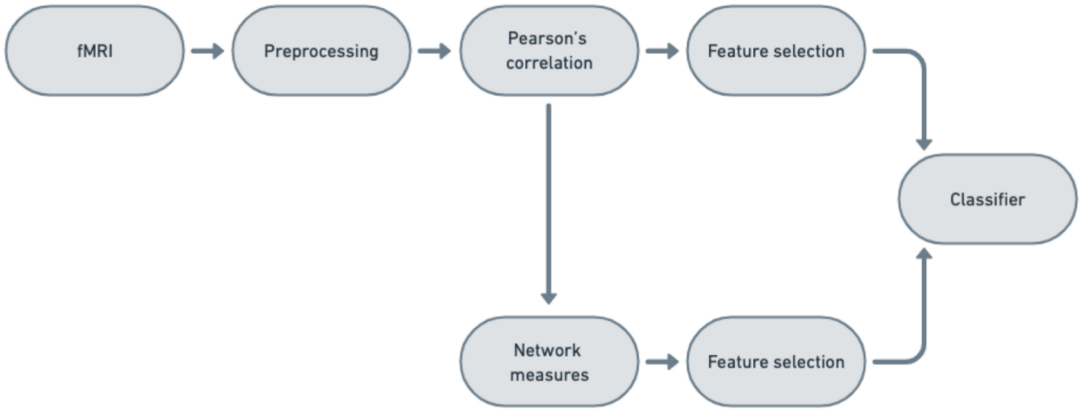
2.3. Extracción de predictores - *features*

La etapa de seleccionar los *features* a utilizar para caracterizar a un individuo es quizás uno de los aspectos más relevantes a la hora de construir un modelo de aprendizaje automático tradicional, teniendo un gran impacto en los resultados obtenidos por cada clasificador.

Luego de analizar varios artículos, optamos por aplicar a nuestro trabajo parte del método presentado en [11], que combina distintas estrategias de selección de *features* para intentar predecir a partir de una *fMRI* si determinado paciente posee o no depresión. Este método nos llamó particularmente la atención, no solo por obtener buenos resultados (en un conjunto de datos reducido), sino por combinar

distintas técnicas comúnmente utilizadas para analizar *fMRI*. Como forma de introducir el método, podemos basarnos en el diagrama de la Figura 5 que explica algunas de las etapas que lo componen. Para profundizar en cada una de las etapas recomendamos ir al artículo original.

Figura 5: Secuencia de pasos realizados para lograr la clasificación de un sujeto partiendo desde las imágenes de *fMRI* crudas.



Lo interesante del método es que se utilizan dos variantes distintas para la extracción de *features*, ambas basadas previamente en la matriz de *correlación de Pearson*. Una de ellas utiliza directamente los coeficientes absolutos de *Pearson* como *features*, seleccionando los mejores $\boldsymbol{k}$ coeficientes, a partir en un criterio de selección específico que en breve detallaremos. La otra variante, a partir de la matriz de correlación de *Pearson* construye un grafo (llamado habitualmente "red" en el análisis de datos), binarizando las aristas que conectan cada nodo en base a distintos criterios, y luego calcula distintas métricas de esta red para utilizar como *features*. Una vez calculadas las métricas en cuestión, también se seleccionan los mejores $\boldsymbol{k}$ *features* utilizando el mismo criterio de selección que la primer variante. Como último paso, se combinan los mejores *features* obtenidos de ambas variantes, y con ellos se procede a construir el modelo de clasificación adecuado.

Este artículo [11] trabaja con un total de 82 pacientes (49 adolescentes con MDD y 33 sujetos de control sanos), sin diferencias significativas entre los grupos con respecto a la edad y al sexo. Los resultados obtenidos con las diferentes

configuraciones de sus predictores se presentan en la Tabla 5.

Cuadro 5: Resultados obtenidos utilizando distintas variantes de *features*, siguiendo el método detallado en artículo [11]

Features	Cantidad	Accuracy	Sensitivity	Specificity
Correlation	10	74 %	81 %	64 %
Network	10	67 %	86 %	43 %
Combination	20	**79 %**	86 %	**70 %**
Anatomical	16384	78 %	**90 %**	55 %

En nuestro trabajo replicaremos parte de este método, en particular utilizando la selección de mejores *features* planteada y la combinación de distintas heurísticas para construir un clasificador, pero lo aplicaremos sobre nuestro conjunto de datos de gran tamaño.

2.3.1. Correlación de Pearson

Podemos definir el coeficiente de *correlación de Pearson* como un índice que puede utilizarse para medir la relación lineal de dos variables siempre y cuando ambas sean cuantitativas y continuas [42]. En este caso se utiliza la *correlación de Pearson* como medida de similitud entre pares de vértices, que corresponden a pares de *ROI's*. Dados dos vértices i, j pertenecientes a un grafo, su *correlación de Pearson* se calcula como $P_{i,j}$:

$$P_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}.$$

Adicionalmente a la *correlación de Pearson*, con el fin de aproximar los datos a una distribución normal, se puede aplicar la transformación de Fisher [43] a la matriz de correlación previamente obtenida:

$$Z_{ij} = \frac{1}{2}\ln(\frac{1+P_{ij}}{1-P_{ij}}).$$

2.3.2. Métricas de red

Describimos a continuación las métricas habituales que se calculan sobre la red para extraer distintas características de un grafo. Por más detalles recomendamos ver [44].

1. **Eficiencia Local y Global:** La eficiencia de una red representa que tan eficiente es el intercambio de información entre los nodos, y también se denomina eficiencia de comunicación. El principal postulado es que cuanto más distantes estén dos nodos en la red, menos eficiente será su comunicación. El concepto de eficiencia se puede aplicar tanto a escala local como global, siendo global la eficiencia que cuantifica el intercambio de información en todo el grafo, y local de un nodo respecto a sus vecinos.

2. **Modularidad:** La modularidad de una red mide la fuerza de la división de la misma en módulos o agrupamientos. Las redes con alta modularidad tienen conexiones sólidas entre los nodos dentro de los módulos, pero escasas conexiones entre nodos en diferentes módulos.

3. **Centralidad de intermediación:** La centralidad de intermediación es una medida que cuantifica la frecuencia o el número de veces que un nodo se encuentra entre los caminos más cortos de otros nodos.

4. **Coeficiente de Clustering:** El coeficiente de clustering de un vértice en una red cuantifica qué tan interconectado se encuentra sus vecinos entre sí. Se puede decir que si el nodo está agrupado como un clique (grafo completo) su valor es máximo, mientras que un valor pequeño indica un vértice poco agrupado en la red.

5. **Promedios:** Para las métricas mencionadas que se calculan a nivel de nodo, además de trabajar con el valor máximo, o con todos sus valores, se puede utilizar como *feature* el promedio del valor obtenido de todos los nodos.

2.3.3. Selección de atributos

En este paso se implementa la selección de *features* de manera independiente para el conjunto de valores absolutos de *correlación de Pearson* y para los *features*

basados en la red, utilizando el método conocido como *Relevancia Máxima Redundancia Mínima (mRMR)* [45]. El método *mRMR* selecciona los *features* más importantes basándose en la correlación entre *features* y la clasificación de cada sujeto (*MDD* o control), mientras que la correlación se calcula utilizando la ganancia de información mutua *(I-mutual information)* [46].

Este algoritmo encuentra un subconjunto de *features* $\{S\}$ de modo que los *features* extraídos tengan la máxima información mutua con las etiquetas y_i, donde $y_i \in \{MDD,\ Healthy\}$.

La máxima relevancia $Max\{D(S, y_i)\}$ se calcula:

$$D(S, y_i) = \frac{\sum_{x_i \in S} I(x_i, y_i)}{|S|}.$$

La mínima redundancia $Min\{R(S)\}$ se calcula:

$$R(S) = \frac{\sum_{x_i, x_j \in S} I(x_i, x_j)}{|S|^2}.$$

Finalmente, se usa un operador $\Phi(D, R) = D - R$ para maximizar la relevancia y minimizar la redundancia de los *features* seleccionados.

Veremos más sobre este método en el Capítulo 3.

2.4. Métodos de clasificación

Planteamos la utilización de distintos métodos de aprendizaje automático supervisado para intentar predecir si determinado paciente posee o no depresión, muchos de los cuales se utilizaron también en diversos artículos ya referenciados en este informe.

Dividimos los clasificadores utilizados en tres subconjuntos, el primero de ellos refiere a clasificadores de aprendizaje automático tradicionales que utilizamos para las primeras experimentaciones realizadas, luego detallamos clasificadores basados en *Gradient Boosting Methods*, y por último entramos en detalle del clasificador *Random Forest*, que si bien entraría en el grupo de clasificadores de aprendizaje automático tradicionales, lo diferenciamos en una sección independiente por ser el clasificador que utilizaremos en la solución final, motivo por el cual haremos foco en su descripción.

A continuación presentamos el conjunto de clasificadores implementados en nuestro estudio.

2.4.1. Clasificadores de aprendizaje automático tradicionales

En esta sección detallamos algunos clasificadores de aprendizaje automático supervisado que utilizamos en nuestras primeras experimentaciones.

Cada uno de estos algoritmos, al pertenecer al grupo de clasificadores de aprendizaje automático supervisado, parten de un conjunto de datos conocido, que incluye tanto las entradas como la salidas esperadas para nuestro modelo. Luego estos modelos identifican patrones en los datos, aprenden comportamientos de los mismos, y luego realizan predicciones, repitiendo y corrigiendo sobre este proceso con el fin de alcanzar un nivel de precisión y rendimiento satisfactorio para un conjunto de datos de prueba [47].

Realizamos pruebas con los algoritmos más comunes y populares, que utilizan distintas estrategias para llevar adelante el proceso de aprendizaje y realizar la predicción de la salida para cada entrada recibida, algunos de ellos: algoritmos de regresión, otros algoritmos bayesianos basados en probabilidad, y también experimentamos con varios algoritmos basados en árboles de decisión, que detallaremos más a fondo en las siguientes secciones.

1. **Regresión logística:** Este modelo logístico binario se utiliza para estimar la probabilidad de una respuesta binaria basada en una o más variables predictoras o independientes (*features*). Es una técnica de aprendizaje automático para clasificación, que puede verse como una red neuronal de una sola neurona [48].

2. **Support Vector Machine:** Formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta. Intuitivamente, una SVM para clasificación binaria es un modelo que representa a los puntos de muestra en el espacio, separando los puntos en dos clases lo más amplias posibles mediante un hiperplano. Cuando las nuevas muestras se evaluan con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas en una u otra clase [28].

3. **K-Nearest Neighbors:** El método de los K vecinos más cercanos es otro método de clasificación supervisado, en este caso no paramétrico. Estima el

valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento pertenezca a determinada clase a partir de la información proporcionada por sus K vecinos más cercanos [49].

4. **Naive Bayes:** El clasificador Naive Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y en el supuesto de que las características son independientes entre sí [50].

2.4.2. Gradient Boosting Methods

Gradient Boosting es una técnica de aprendizaje automático utilizada en tareas de regresión y clasificación, que proporciona un modelo de predicción en forma de conjunto de modelos de predicción débiles, que suelen ser árboles de decisión.

Utilizamos este tipo de clasificadores en algunas experimentaciones, ya que suelen obtener resultados similares a los modelos basados en *Random Forest*, técnica también basada en árboles de decisión.

1. **XGBoost:** XGBoost (Extreme Gradient Boosting), es uno de los algoritmos de aprendizaje automático de tipo supervisado más uilizado en la actualidad. Es un algoritmo predictivo supervisado que utiliza el principio de *boosting*. Utilizando *boosting* se pretende generar múltiples modelos de predicción "débiles" de manera secuencial, y luego cada uno de estos toma los resultados del modelo anterior, con el objetivo de generar un modelo más "fuerte", utilizando para esto último el algoritmo de optimización de descenso por gradiente [51].

2. **LightGBM:** Al igual que XGBoost, LightGBM es un algoritmo de refuerzo de gradientes (Gradient Boosting), basado en modelos de árboles de decisión. Persigue la misma idea que XGBoost, maximizar o minimizar una función objetivo combinando clasificadores débiles más sencillos [52].

2.4.3. Random Forest

Random Forest [53] es un algoritmo de aprendizaje automático supervisado que se basa en *árboles de decisión*. Los árboles de decisión son un modelo que intenta aproximar funciones discretas, y se pueden aplicar en distintas áreas de interés. Dado un conjunto de instancias clasificadas, cada una de ellas con una determinada cantidad de atributos, se construye un árbol en el cual cada rama

representa una restricción sobre los posibles valores de las instancias, expresada como una conjunción. Cada nodo del árbol se corresponde con un atributo de la realidad, y representa las posibles decisiones a tomar. Por ende dependiendo de la cantidad de valores que tiene un atributo, resultan las posibles bifurcaciones que tendrá una rama [53]. Existen distintos algoritmos para construir árboles de decisión, siendo uno de ellos y quizás el más popular el algoritmo denominado *ID3* [54]. El algoritmo *ID3* construye un árbol desde la raíz a las hojas, en un enfoque *top-down*. En cada paso se decide por cuál atributo se debe preguntar y se genera una rama por cada posible valor del atributo seleccionado. Se repite recursivamente este proceso para cada una de las ramas generadas, tomando aquellas instancias que tienen el valor de la rama en el atributo seleccionado. Este algoritmo sigue un enfoque *greedy* [55] puesto que nunca vuelve hacia atrás en una decisión tomada y no verifica que el atributo seleccionado en un momento dado haya sido realmente el mejor. Para seleccionar el mejor atributo existen distintas variantes a utilizar, muchas de ellas basadas en la *entropía*, siendo esta última una manera de medir la heterogeneidad de los datos. Cuanto más homogéneos son, menor será la entropía. Uno de los fenómenos más habituales en el área del aprendizaje automático es el *sobreajuste*, y se da cuando un modelo dado, en este caso un árbol de decisión, se ajusta demasiado a los datos de entrenamiento, perdiendo así eficacia en la clasificación de nuevas instancias. Para mitigar el sobreajuste en árboles de decisión existen distintos métodos, muchos de ellos basados en definir la profundidad de cada rama del árbol y utilizar distintas técnicas de poda. *Random Forest* es una de las técnicas utilizadas para atenuar la sensibilidad de un árbol en el conjunto de entrenamiento, formando una especie de bosque de árboles (poco correlacionados) y ensamblando los mismos por votación. Los árboles seleccionan datos de entrenamiento de forma aleatoria y por tanto ven distintas porciones de los datos de entrada, por ende cada uno de ellos se entrena con distintas muestras de datos para un mismo problema.

2.5. Resumen de trabajos relacionados

A lo largo de este capítulo fuimos presentando una serie de trabajos científicos que atacan completa o parcialmente nuestro problema predictivo. Para su rápida referencia, en esta sección resumimos los aportes de cada uno de los trabajos

previos que tienen impacto en nuestra proyecto:

- Sin duda uno de los trabajos más importantes en el cual nos hemos apoyado fue *Classification of Major Depressive Disorder from Resting-State fMRI* [11]. Aquí se presenta una metodología de extracción de features dividida en dos etapas. La primera de ellas trabaja sobre la creación de un grafo para cada individuo, utilizando las imágenes *fMRI* (etapa comunmente utilizada a efectos de esta temática). Pero luego decide utilizar también los coeficientes de *Pearson* como predictores del clasificador, además agregando la metodología de selección de *features* con el algoritmo *mRMR*. Podemos decir que este artículo fue el puntapié inicial para nuestro trabajo.

- El segundo artículo que también fue de gran importancia en este trabajo fue el denominado *Reduced default mode network functional connectivity in patients with recurrent major depressive disorder* [14]. Como se comentó en la Sección 2.2.2, éste artículo analiza la conectividad funcional del cerebro mediante el análisis de la señal *BOLD*, tratando de mostrar patrones correlacionados entre regiones cerebrales separadas. En este artículo se trabaja desde el punto de vista de la neurociencia, lo cual excede a nuestro campo de estudio. Pero es aquí donde se define un mecanismo para filtar el conjunto de datos, en búsqueda de reducir su heterogeneidad, utilizando criterios de exclusión como datos demográficos incompletos o movimiento excesivo del paciente durante la resonancia.

- Otro artículo relevante para nuestro estudio y para nuestra comprensión del mundo de las *fMRI* fue el denominado *Machine learning in major depression: From classification to treatment outcome prediction* [8]. Este artículo presenta una revisión de trabajos similares de aprendizaje automático utilizando imágenes cerebrales. Aquí se reúnen varios trabajos específicos para clasificación de trastorno *MDD* utilizando imágenes de resonancia magnética *fRMI*, aunque también reúnen algunos artículos sobre otros trastornos del estado de ánimo. El objetivo aquí es ofrecer una descripción general que pueda ayudar a los lectores a comprender mejor las aplicaciones de la extracción de neuroimagen en la depresión. En lo que compete a nuestro trabajo, este artículo fue importante a la hora de tener un pantallazo general sobre el universo de estudio.

- El artículo llamado *Multivariate Classification of Major Depressive Disorder Using the Effective Connectivity and Functional Connectivity* [13] presenta un método para clasificar individuos con trastorno depresivo mayor basándose en la conectividad cerebral y aquí también se utilizan imágenes de resonancia magnética funcional *(fMRI)* en estado de reposo. Aquí los clasificadores que se utilizan son: *Linear Support Vector Machine*, *non-linear SVM*, *k-Nearest Neighbor (KNN)* y *Regresión logística*, dándonos un buen punto de partida sobre que clasificadores son utilizados en esta temática.

- En el artículo *Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory* [12], al igual que los anteriores, los investigadores desarrollan técnicas predictivas para inferir la existencia de depresión a partir de su comportamiento neuronal utilizando imágenes *(fMRI)* en estado de reposo entre otro tipo imágenes. Aquí se utiliza un clasificador *SVM* con un conjunto de datos pequeño. Algo que encontramos interesante aquí es que se realiza un análisis para tratar de revelar cuales son las métricas extraídas del grafo resultante más robustas para clasificar a los individuos, utilizando un clasificador *SVM*. Dentro de las métricas analizadas se encuentran Global efficiency, Modularity, Global betweenness, las cuales también fueron utilizadas en el artículo [11] y en nuestro estudio.

- Por último, otro de los artículos que nos gustaría resaltar es *Multivariate Machine Learning Analyses in Identification of Major Depressive Disorder Using Resting-State Functional Connectivity* [33]. Aquí se trabaja con el mismo dataset utilizado por nuestro estudio, y los resultados obtenidos por estos científicos fueron muy buenos dada la magnitud de los pacientes. Si bien no construyen grafos a partir de las *fMRI* para caracterizar a los individuos, utilizan técnicas matemáticas y de optimización, como por ejemplo *descomposición SVD* [56]. Además, utilizan varios de los clasificadores que nosotros terminamos utilizando. Otro detalle no menor, es que el artículo es muy reciente (Julio 2021), por lo que representa una noción del estado del arte bastante actual.

3. Solución

En esta sección describimos nuestra propuesta de metodología para resolver el problema de clasificación planteado. La solución final resulta luego de investigar y desarrollar distintas opciones, detallando el proceso general implementado, los *features* utilizados y las distintas etapas construídas para obtener la solución definitiva. A continuación un breve resumen del paso a paso implementado, para luego entrar en detalle de cada una de las etapas relevantes del proceso.

Partimos de las *fMRI* en estado de reposo para cada paciente, queriendo transformar las mismas en una matriz que nos indique la correlación entre las regiones de interés del cerebro *(ROI's)* identificadas en las imágenes de resonancia magnética. En nuestra solución utilizamos el *dataset* preprocesado con la herramienta *DPARSF*, que provee como salida la matriz deseada para cada individuo. Una vez obtenida la matriz, seleccionamos el *atlas* a utilizar de los 9 disponibles y calculamos una nueva matriz utilizando *correlación de Pearson*, en este caso una matriz cuadrada simétrica, que indica la correlación entre todas las regiones definidas en el *atlas*. A partir de nuestra matriz de correlación construimos un grafo, en el cual los vértices son las regiones de interés, y las aristas indican si entre dos regiones existe o nó una correlación superior a cierto umbral predefinido. Luego de calcular el grafo, planteamos dos variantes principales para trabajar con distintos *features* y métricas relativas al grafo. La primer variante implica trabajar directamente con los coeficientes absolutos de *Pearson* (sin necesidad de construir un grafo, basta con la matriz de correlación), seleccionando los mejores coeficientes mediante el método de selección de atributos *mRMR*, y sumar algunos atributos demográficos de los individuos. La segunda variante implica utilizar como *features* algunas métricas extraídas de grafos, así como también incluir ciertos atributos demográficos de los individuos brindados por el *dataset*. Por último, clasificamos a los individuos utilizando *Random Forest*, siendo este último el modelo de aprendizaje automático tradicional que presentó los mejores resultados a lo largo de la experimentación realizada.

3.1. Filtrado de datos

Como se describió en el estado del arte, realizamos un filtrado de datos luego de experimentar un tiempo con el *dataset* completo. Nos cuestionamos si trabajar

con un conjunto de datos heterogéneo era preciso a la hora de clasificar los sujetos. Como se comentó los datos provienen de distintos grupos de investigación lo cual engloba un conjunto de datos de universos un poco diferentes (uno de los factores que identificamos influyente es que se utilizó distinta maquinaria para extraer las imágenes).

Este *dataset* filtrado o reducido, parte del dataset completo y selecciona 848 sujetos con *MDD* y 794 sujetos de control, formado por 17 de los 25 grupos de investigación. Dado que uno de los criterios de exclusión fue información demográfica incompleta de los pacientes, el conjunto final de datos filtrados contiene los datos demográficos de todos los sujetos, siendo estos la edad, el sexo, el nivel educativo y el movimiento de los pacientes durante la resonancia (motion), lo que nos permitió utilizarlos como predictores.

Dado que obtuvimos mejores resultados partiendo del *dataset* reducido, agregamos este filtro a la solución final de nuestro trabajo. Entonces en las dos variantes finales implementadas se trabajó sobre el *dataset* filtrado, utilizando 1618 pacientes, 848 con *MDD* y 794 pacientes de control.

3.2. Relación entre regiones de interés: generación del grafo

A continuación comentaremos el proceso ya mencionado de generación de grafos implementado para cada individuo. Este conjunto de pasos lo consideramos el más tradicional pues fue el denominador común de los distintos artículos que tratan la temática. Luego de obtener los datos preprocesados a partir de cada sesión de *fMRI*, para poder obtener el grafo de cada paciente se estudia la correlación lineal entre un par de regiones de interés del cerebro, utilizando la *correlación de Pearson* como métrica. Aplicando esta función de correlación a cada par de regiones de interés, obtenemos una matriz simétrica cuadrada donde cada coordenada *i,j* contiene el valor de la *correlación de Pearson*, $P_{ij} = \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$, para las regiones de interés *i,j*.

En nuestro trabajo decidimos dejar parametrizado el *atlas* a utilizar, de forma tal de poder analizar los resultados utilizando la mayoría de los disponibles por la herramienta *DPARSF*.

Por último, definimos los nodos de nuestro grafo como cada región de interés y las aristas cómo el valor de la correlación calculada. Adicionalmente se debe

definir un umbral de correlación para poder decidir cuando existe o no una arista entre estos nodos. En este trabajo llamaremos a este valor como coeficiente de *binarización*, valor que también será parametrizable. A partir de estas matrices generadas se obtiene el grafo resultante para cada sujeto.

3.3. Extracción de predictores

Luego de construidos los grafos de cada sujeto a clasificar necesitamos extraer los predictores que luego utilizará el clasificador *Random Forest*. La intención de esta sección es mostrar el métodos de extracción de predictores que fueron parte de la solución de nuestro proyecto.

La extracción consiste en el cálculo de métricas del grafo, aquí se extraen valores del grafo resultante de cada sujeto, donde su construcción se menciona en la sección anterior. Resulta de interés por ejemplo calcular la *Eficiencia Global* de los grafos, donde cuantificamos el intercambio de información entre nodos. Otro conjunto de predictores utilizados para la clasificación son los coeficientes absolutos de *Pearson*. Aquí no hace falta construir un grafo para extraer estos valores, simplemente se utilizan los atributos de *Pearson* que se calcularon a partir de las matrices de correlación que devuelve la herramienta *DPARSF*. Es aquí donde se utiliza el método de selección de *features mRMR*. Entonces, luego de obtener el conjunto de atributos de *Pearson*, estos son colocados dentro del algoritmo que selecciona los más relevantes, y el resultado lo terminamos utilizando como predictores en nuestra solución.

Cabe destacar que no se utiliza el método de selección *mRMR* en el conjunto de métricas extraídas de la red, dado que el número de valores extraídos de la red es pequeño en comparación con coeficientes de *Pearson*.

3.3.1. Cálculo de métricas sobre el grafo

Para la extracción de atributos de los grafos resultantes, a nivel de nodo local en el grafo calculamos tres atributos: *local efficiency*, *clustering coefficient*, y *betweenness centrality*. Luego, a nivel global calculamos: *global efficiency* y *modularity*.

Con la intención de evaluar la capacidad que tienen los atributos para discriminar los pacientes de *MDD*, realizamos una serie de histogramas. Estos primeros

histogramas fueron realizados con la totalidad de los datos, es decir, sin realizar el filtrado del cual se habló en secciones previas. Con los atributos globales a los grafos resultantes, las gráficas se implementaron calculando el promedio de los valores obtenidos para cada nodo. El objetivo aquí es tener una primera visión del comportamiento de los *features* antes de clasificar a los sujetos.

Si para determinado *feature*, el comportamiento del mismo para pacientes con *MDD* es totalmente diferente que para los pacientes de control, este *feature* podría ser de utilidad a la hora de clasificar a un sujeto.

Los resultados aquí no fueron muy alentadores, las Figura 6 muestra los histogramas correspondientes. Se puede apreciar como el comportamiento de las métricas extraídas del grafo no es sustancialmente diferente entre pacientes con trastorno *MDD* y pacientes de control.

En las próximas secciones intentamos realizar investigaciones similares con el resto de los *features* y con los datos filtrados. El objetivo aquí es poder tener una visión más general del comportamiento de los *features*, intentar visualizarlos de alguna forma y quizás encontrar algún comportamiento diferente en los pacientes con *MDD vs* pacientes de control. Este comportamiento podría ser utilizado por el algoritmo clasificador a la hora de clasificar.

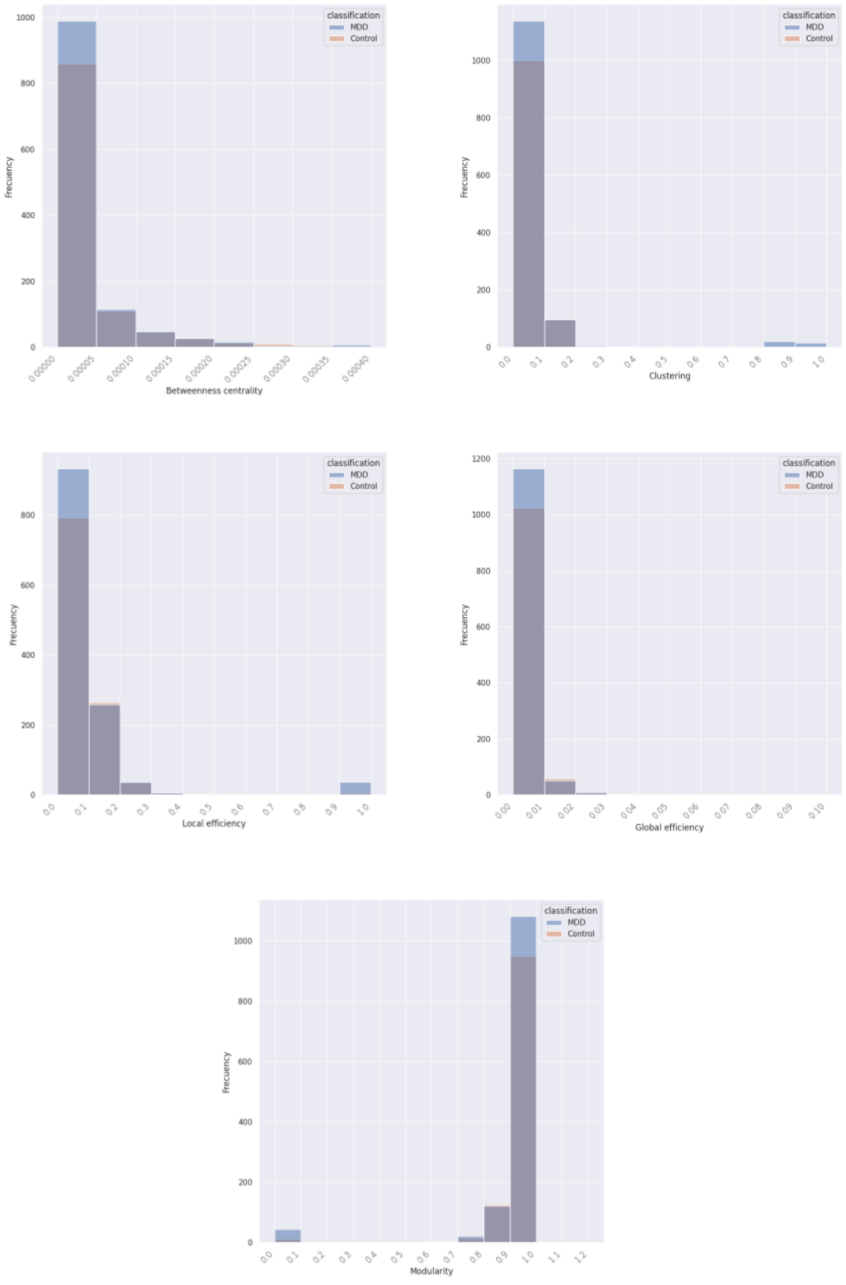
3.4. Selección de mejores predictores

Siguiendo con nuestro método, utilizamos el algoritmo de selección de *features* llamado *Minimum Redundancy and Maximum Relevance (mRMR)* detallado en el capítulo anterior. La selección de atributos es un método que engloba múltiples objetivos:

- Mejorar la precisión de la predicción mediante la eliminación de atributos irrelevantes.
- Acelerar el entrenamiento del modelo y la velocidad de predicción.
- Proporcionar una mejor capacidad de interpretación y diagnóstico del modelo.

No obstante, seleccionar un subconjunto de características óptimo dentro de un espacio de características de gran magnitud se considera un problema NP-

Figura 6: Histogramas con la distribución de la distintas métricas extraídas de los grafos resultantes, discriminados por individuos de control e individuos con *MDD*.



completo. El método de selección de atributos *mRMR* resuelve este problema seleccionando los atributos relevantes mientras controla la redundancia dentro de los atributos seleccionados.

3.4.1. Selección de mejores atributos para la correlación de Pearson

En esta biblioteca, el procedimiento que selecciona los mejores atributos recibe un parámetro K, que representa el número de mejores atributos que se desea seleccionar.

```
selected_features = mrmr_classif(X=X, y=y, K),
```

Si bien K parece a priori un hiperparámetro a definir a la hora de realizar la validación cruzada de hiperpárametros, esto podría tomar más tiempo de lo deseado. Por esta razón luego de algunas pruebas manuales decidimos seleccionar un K pequeño, con por ejemplo los mejores 5 ó 10 valores absolutos de *Pearson* para la solución final.

Siguiendo con lo comentado en el final de la Sección 3.3.1, continuamos con la visualización del comportamiento de los *features* y realizamos, al igual que con las métricas del grafo, histogramas con el total del *dataset*, en este caso de los 5 mejores coeficientes de *Pearson* seleccionados por el algoritmo *mRMR*, discriminando por pacientes con *MDD* y pacientes de control. Al igual que en el caso de los predictores calculados a partir de grafos, el comportamiento a priori no identifica nada que parezca diferente y de utilidad para el algoritmo clasificador. En la Figura 7 se pueden observar los gráficos.

3.4.2. Evaluación del filtrado de datos sobre los predictores

Continuando con nuestro análisis, nos pareció pertinente completar la investigación del comportamiento de los *features* graficando los mismos valores que en las secciones anteriores, métricas extraídas del grafo y coeficientes de *Pearson*, pero ahora con el *dataset* filtrado (*dataset* utilizado en nuestra solución definitiva), nuevamente con el fin de poder observar el comportamiento de los *features* seleccionados. Presentamos estos histogramas en las Figuras 8 y 9.

Figura 7: Histogramas con los mejores coeficientes seleccionados por el algoritmo *mRMR*, discriminados por individuos de control e individuos con *MDD*.

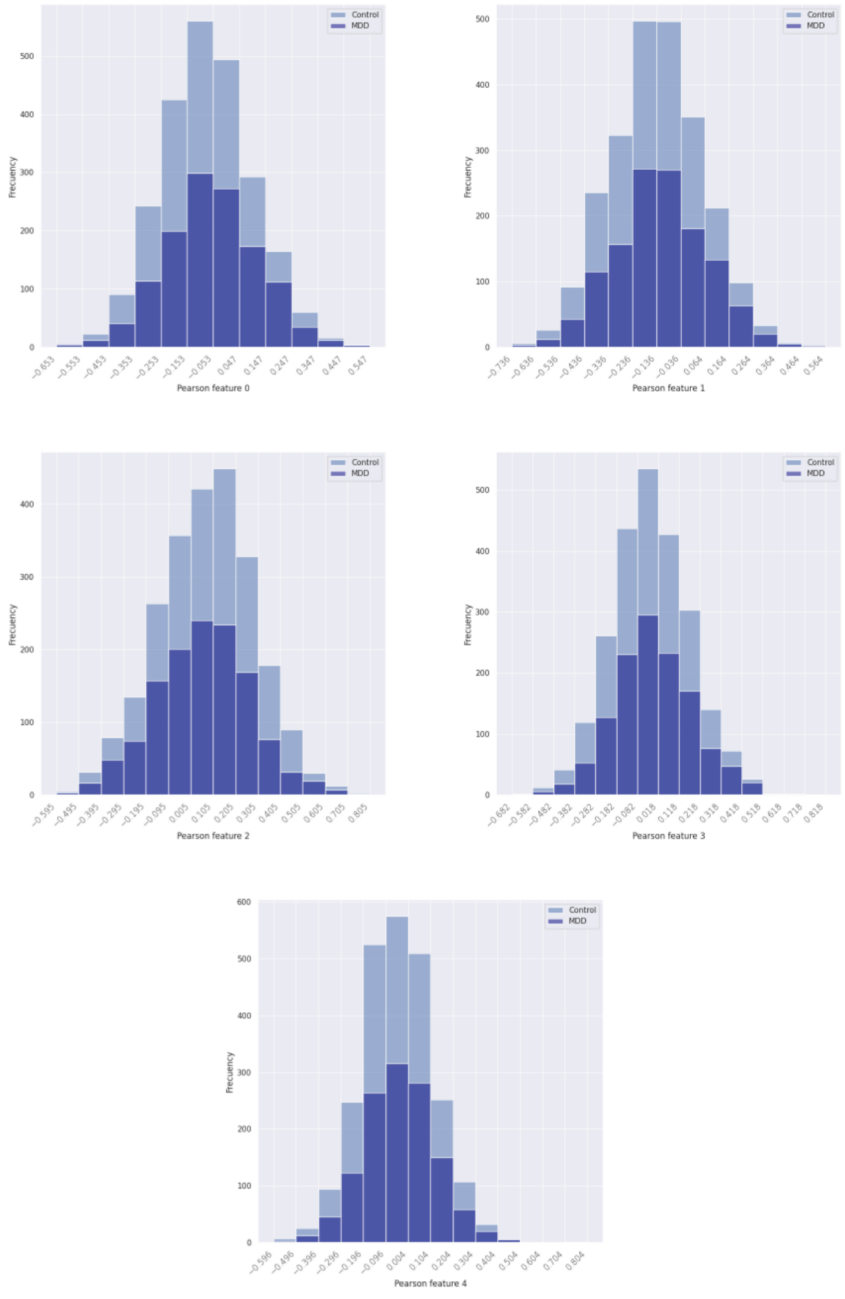


Figura 8: Histogramas con distribución de las distintas métricas extraídas de los grafos resultantes, discriminados por individuos de control e individuos con *MDD* para el *dataset* filtrado.

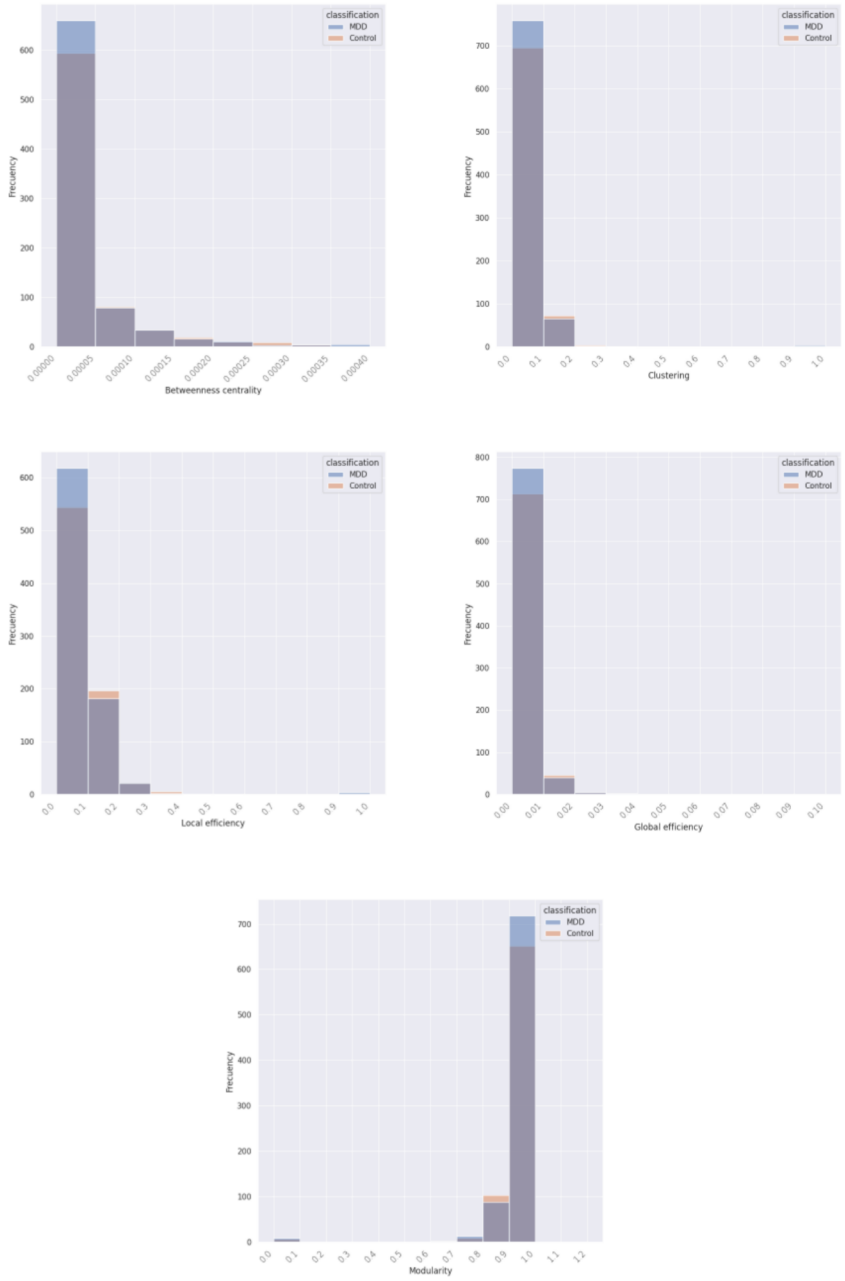
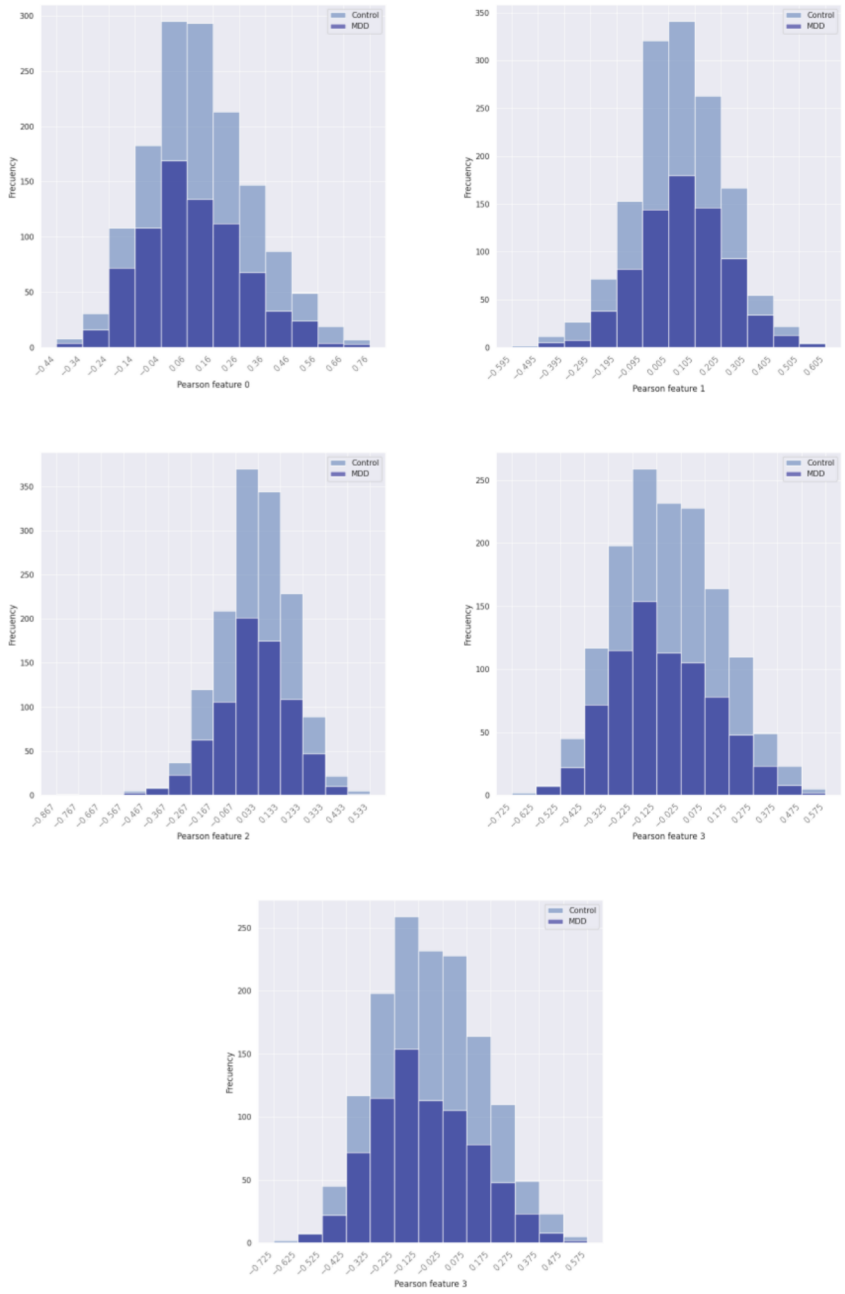


Figura 9: Histogramas con distribución de los coeficientes de Pearson más importantes, discriminados por individuos de control e individuos con *MDD* para el *dataset* filtrado.



Aquí la situación no es muy diferente a la de los casos anteriores, para los predic-

tores extraídos de los grafos el comportamiento es casi idéntico para pacientes con *MDD* y pacientes de control. Para los atributos más relevantes de *Pearson*, quizás el comportamiento no es tan parecido, pero de igual forma no parece influyente. Se pueden ver ligeros cambios en donde se agrupan la mayor cantidad de pacientes, pero en general responde a la misma tendencia que en los casos anteriores.

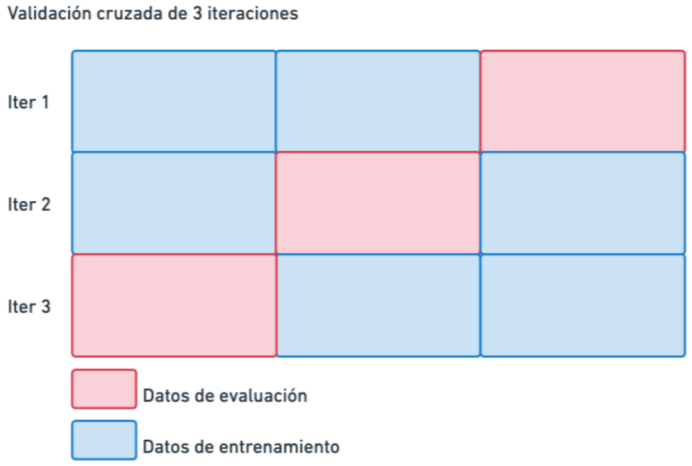
3.5. Entrenamiento del clasificador mediante validación cruzada

Seleccionados los *features* que utilizaremos para clasificar, debemos entrenar un clasificador, en nuestro caso *Random Forest*, utilizando un conjunto de datos de entrenamiento. El conjunto de entrenamiento incluye el 90 % de las instancias disponibles en el *dataset*, reservando el otro 10 % de las instancias para la evaluación final de resultados.

En este caso, proponemos realizar el entrenamiento mediante validación cruzada: donde dado un conjunto de entrenamiento I, la validación cruzada de n subconjuntos, consiste en dividir el conjunto de entrenamiento en n particiones, y realizar k iteraciones distintas en las cuales se ejecuta *Random Forest*. Para cada una de las iteraciones se provee $\frac{I}{n} * (n-1)$ datos para la generación del bosque y las $\frac{I}{n}$ instancias restantes para utilizar como datos de evaluación del bosque generado. Una vez realizada la validación cruzada, se toman los resultados de las k iteraciones y se promedian para obtener el valor final asignado al hiper-parámetro que se estaba validando.

En nuestra solución proponemos realizar una validación cruzada de ***3 folds***, aplicando el filtrado de datos sobre el *dataset*, y trabajando como comentamos con el 90 % de las instancias. Un ejemplo de la partición de los datos se puede visualizar en la Figura 10.

Figura 10: Boceto de toma de datos para la implementación de validación cruzada de $k = 3$ iteraciones.



3.6. Elección de hiper-parámetros

El clasificador *Random Forest* nos permite configurar ciertos parámetros de entrada para alterar el funcionamiento por defecto del algoritmo. En esta sección definimos una metodología para definir estos hiper-parámetros.

La elección de hiper-parámetros es la última etapa para terminar de especificar nuestra metodología, donde habitualmente se espera una pequeña mejora en los resultados obtenidos.

3.6.1. Hiper-parámetros relevantes

Identificamos a continuación los hiper-parámetros de *Random Forest* que determinamos mediante validación cruzada, utilizando la nomenclatura provista por la popular biblioteca que lo implementa, *scikit learn* [57]:

1. **criterion:** Función de ganancia utilizada para decidir cuál atributo es el mejor localmente y seleccionarlo como raíz en un momento dado, siendo los valores posibles *gini* ó *entropy* [58].

2. **n_estimators:** Cantidad de árboles a utilizar para la generación del bosque, siendo los valores posibles 10, 20, 50, 100.

3. **max_depth:** Establece la profundidad máxima de cada árbol, siendo los valores posibles 5, 10, 20, 30, 40, 50, 100.

4. **max_features:** Número de *features* considerados al buscar la mejor división del árbol posible, siendo los valores posibles *sqrt(features)* ó *log2(features)* .

5. **warm_start:** Valor *booleano*, cuando se establece como verdadero, reutiliza la solución de la llamada anterior ejecutada, para ajustar y agregar más estimadores al conjunto.

Como comentamos en distintas instancias del informe, tanto los *features* a utilizar como los hiper-parámetros seleccionados para la ejecución de *Random Forest*, están sumamente relacionados con el atlas definido como punto de partida, por lo que ejecutamos la validación cruzada para los atlas que mejores resultados presentaron en las distintas experimentaciones ejecutadas, describiendo los mismos a continuación.

1. **Automated Anatomical Labeling:** Atlas constituído por 117 regiones de interés predefinidas.

2. **Harvard-Oxford cortical:** Atlas constituído por 96 regiones de interés predefinidas.

3. **Dosenbach ROI's:** Atlas constituído por 116 regiones de interés predefinidas.

Un punto interesante a resaltar es que estos atlas poseen una cantidad de regiones de interés similar. Detallamos a continuación el resto de los atlas disponibles por el *dataset*, para poder apreciar con mayor precisión la cantidad de regiones de interés de cada uno de ellos.

1. **Harvard-Oxford subcortical:** Atlas constituido por 16 regiones de interés predefinidas.

2. **Craddock clustering:** Atlas constituido por 200 regiones de interés predefinidas.

3. **Zalesky random parcelations:** Atlas constituído por 980 regiones de interés predefinidas.

4. **Power Rois:** Atlas constituido por 264 regiones de interés predefinidas.

3.6.2. Búsqueda de hiper-parámetros

Debido a la cantidad de valores de hiper-parámetros que queremos validar, sumado esto a los distintos *atlas* a estudiar, y también las distintas variantes planteadas en relación a los *features*, necesitamos realizar muchas ejecuciones para finalizar todas las combinaciones planteadas y obtener los mejores hiper-parámetros. Como se verá en la sección experimental, esto se logra ejecutar en paralelo utilizando distintos *clusters* de computadores.

4. Experimentación

A la hora de buscar el mejor método posible para intentar predecir si un paciente presenta o no *MDD*, nos enfrentamos a muchas decisiones que pueden afectar la performance de nuestro algoritmo, y por consiguiente los resultados del trabajo. Podemos optar por utilizar distintos *atlas*, distinta cantidad de *features*, distintas formas de binarizar las matrices de correlación obtenidas, distintas funciones de correlación propiamente dichas, distintos hiper-parámetros en los algoritmos, etc. Por lo que vale la pena invertir trabajo en validar que combinación puede ser la más provechosa. Detallamos a continuación todas las líneas de trabajo y experimentos realizados con este fin.

Si bien siempre partimos del mismo *dataset* original, en todos los experimentos realizados también trabajamos con el *dataset* filtrado, en donde se descartan algunos individuos en base a su edad y a la calidad de su resonancia entre otras razones. También trabajamos agrupando a los individuos según el centro de investigación donde se obtuvieron las resonancias.

Todos los experimentos planteados se basan en un esquema bien definido, en el cual a partir de las distintas regiones de interés de determinado atlas, se aplica alguna función de correlación (como por ejemplo *Pearson*). En cualquiera de las variantes implementadas trabajamos con todos los atlas disponibles de manera independiente. Adicionalmente con el fin de mejorar la normalidad de los datos obtenidos experimentamos aplicando la transformación de Fisher a la matriz de correlación previamente obtenida, pero no obtuvimos cambios significativos en ninguna variante implementada.

Luego de obtenida la matriz, se extraen *features* o métricas relevantes de la misma, para la mayoría de los casos convirtiendo esa matriz previamente en un grafo. Para convertir la matriz a un grafo, trabajamos con distintos coeficientes de binarización, muy atado cada uno de ellos al atlas seleccionado. El coeficiente de binarización determina el umbral ante el cual corresponde colocar una arista entre dos vértices. El rango de los coeficientes de binarización utilizados varía entre 0 y 1, colocando una arista entre dos vértices, siempre y cuando la correlación entre los mismos sea mayor al coeficiente de binarización, recordando que la correlación de *Pearson* vive en el intervalo $[-1, 1]$.
Para calcular las métricas de red sobre el grafo, utilizamos la librería *networkX* [59]

de *Python*. Para no quedarnos solamente con los grafos y las métricas brindadas por la librería *networkX*, también generamos los grafos con otra herramienta denominada *igraph* [60], y aunque no obtuvimos cambios significativos, sirvió para validar las métricas que estábamos utilizando.

Nuestra línea de trabajo sigue con la selección de *features*, donde también trabajamos con distintas alternativas, utilizando como *features* métricas de los grafos construídos, atributos demográficos, combinaciones de los mismos y también selección de *features* basada en *Minimum Redundancy and Maximum Relevance (mRMR)*, criterio detallado en capítulos anteriores implementado para seleccionar los *features* más relevantes. Debimos realizar una pequeña investigación sobre que herramientas de selección de atributos *mRMR* podíamos utilizar. En nuestro lenguaje de desarrollo, *Python 3.8*, decidimos trabajar con la biblioteca *mrmr_selection* [61], en el repositorio de la misma se hace referencia a un artículo [62] publicado por los ingenieros de la compañía *Uber* en el año 2019, que describe cómo llevar a cabo la implementación de *mRMR* en su plataforma y los beneficios de la misma.

Luego de decidir qué *features* utilizar para caracterizar a un individuo, construímos uno o varios modelos de aprendizaje automático para clasificar a nuestro *dataset*.

A continuación brindamos un pequeño resumen de las secciones de experimentación que abordaremos. Comenzamos con la Sección 4.1, con una descripción general de los experimentos realizados, repasando los clasificadores implementados disponibles para nuestro estudio, distintos *atlas* disponibles, etc. Luego saltaremos a la Sección 4.2, donde nos enfocamos en las distintas configuraciones que podemos realizar de predictores, repasamos *features* disponibles y combinaciones realizadas en distintas experimentaciones. Entraremos en detalle de las 3 configuraciones principales que utilizamos como punto de partida, utilizando todos los *features* disponibles al mismo tiempo (Sección 4.2.1), combinando *features* calculados de los grafos construidos con atributos demográficos (Sección 4.2.2), y también coeficientes absolutos de *Pearson* con atributos demográficos (Sección 4.2.3). Luego pasamos a la Sección 4.3, donde comentamos en detalle la búsqueda de hiper-parámetros implementada, mencionando también la ejecución realizada en *Cluster.uy* (Sección 4.3.1) y las distintas combinaciones de hiper-parámetros (Sección 4.3.2), junto con los resultados obtenidos. Posteriormente damos lugar a la

Sección 4.4, donde estudiamos el impacto de los centros de investigación en la calidad de los resultados alcanzados, trabajando con cada clínica o centro de investigación de manera independiente, sin combinar todos los individuos del *dataset*. Por último tenemos la Sección 4.5, en donde evaluamos técnicas adicionales para utilizar la completitud de los *features* disponibles, y trabajamos con nuevos clasificadores basados en *Gradient Boosting Methods*.

4.1. Exploración de los métodos de clasificación y de los *atlas*

Tenemos a disposición distintos clasificadores con los cuales podemos construir diferentes modelos para entrenar con nuestro *dataset* y luego clasificar a una porción del mismo. Implementamos algunos experimentos base con el fin de identificar qué clasificadores logran adaptarse mejor a nuestro método, a los *features* y a los datos disponibles. Inicialmente planteamos la utilización de 4 modelos distintos, utilizando los siguientes clasificadores: *Random Forest*, *K-nearest Neighbors*, *Naive Bayes* y *Support Vector Machine*.

Si bien en los primeros acercamientos y pruebas logramos resultados similares con estos clasificadores, analizando el comportamiento de cada uno de ellos en la clasificación, encontramos algunos resultados interesantes. Como primer punto a destacar, podemos comentar que los modelos de *Support Vector Machine* y *Naive Bayes* utilizados no obtienen buenos resultados. Además de presentar valores de *accuracy* distantes de los otros dos clasificadores utilizados, para el caso de *SVM* identificamos que prácticamente la totalidad de las instancias siempre son clasificadas como positivas o negativas, sin realizar ningún tipo de aprendizaje. Al trabajar con un *dataset* estratificado, *SVM* retorna resultados siempre cercanos al 50 % de exactitud. Si bien acompañaremos los experimentos realizados con *SVM*, no resultará de mucha utilidad. En el caso del clasificador probabilístico de *Naive Bayes*, la clasificación de las instancias fue bastante estratificada, no caímos en el mismo problema que *SVM*, pero de todas formas los resultados no lograron equiparar a los obtenidos con los otros clasificadores.

Los clasificadores que presentaron mejores resultados fueron *Random Forest* y un poco por debajo *K-nearest Neighbors*. Además, el clasificador *Random Forest* (basado en árboles de decisión) también nos permite observar cuales fueron los

features más relevantes utilizados al momento del entrenamiento para construir los árboles de decisión, dado que en cada fase del algoritmo, se selecciona el atributo de mayor ganancia para utilizar como posible nodo del árbol. Por tales motivos nos parece interesante hacer foco en *Random Forest* a lo largo de la experimentación, ya que podemos identificar cuales *features* son relevantes al momento de clasificar a los individuos, sumando una herramienta extra para intentar reconocer que *features* utilizar. Recordemos que la mayoría de los artículos obtienen buenos resultados trabajando con pocos individuos, pero al incrementar la dimensión de los mismos se pierde bastante la exactitud.

Otro aspecto relevante (que también comentamos en el Capítulo 3) es el *atlas* a utilizar en nuestros experimentos. Si bien realizamos distintas iteraciones utilizando todos los *atlas* a disposición, parecen performar un poco mejor los modelos que utilizan una cantidad pequeña de regiones de interés. Recordemos que para todos los experimentos, a partir de las regiones de interés construimos matrices de correlación, y nos basamos en las mismas para obtener distintos *features*. Cuanto más grande es la cantidad de regiones de interés, más grande es la matriz de correlación resultante, y por ende se torna más costoso calcular las métricas y *features* que nos interesan. A lo largo de la experimentación mostraremos resultados en base al *atlas* denominado *Harvard-Oxford cortical*, formado por 96 regiones de interés.

4.2. Distintas configuraciones de predictores

Luego de seleccionado el algoritmo clasificador, está claro que contamos con dos grandes conjuntos de *features*, por un lado las métricas resultantes de construir un grafo para cada individuo (a partir de su matriz de correlación de regiones de interés), y por otro lado los coeficientes de *Pearson* extraídos del cálculo de *Correlación de Pearson*. Adicionalmente contamos con los datos demográficos de los sujetos, siendo los mismos la edad, sexo, nivel educativo y *motion* de la resonancia. Inspirados en el artículo [11], implementamos distintos experimentos combinando los tres conjuntos de *features* existentes, utilizando los clasificadores que mencionamos anteriormente (con foco en el clasificador *Random Forest*), y teniendo a disposición tanto el *dataset* completo como el *dataset* reducido, obtenido luego del filtrado de datos. Resumiendo, partimos de tres tipos de predictores para realizar los experimentos:

- Seleccionar los **mejores coeficientes de *Pearson*** a partir del algoritmo *mRMR*;
- **Métricas extraídas de grafos** tales como eficiencia global, eficiencia local, modularidad, coeficiente de clustering y centralidad de intermediación;
- Datos **demográficos** del sujeto, edad, sexo, nivel educativo y *motion* de la resonancia.

4.2.1. Utilizando la totalidad de los predictores

Este experimento es el que más se parece al método resumido en el estado del arte, más precisamente en la etapa de selección de *features*, combinando las dos formas de extraer *features* de una imagen de resonancia magnética, y adicionando también los atributos demográficos disponibles.

Sólo trabajamos con los coeficientes absolutos de *Pearson* resultantes luego de aplicar binarización y construir el grafo, aunque una alternativa válida podría ser utilizar todos los coeficientes obtenidos sin necesidad de transformar la matriz a un grafo. De todas formas, los coeficientes absolutos de *Pearson* luego de obtener el grafo se acercan al orden de centenas de valores (dependiendo del *atlas* utilizado). Por esta razón, nos interesa seleccionar los mejores coeficientes dentro de los disponibles, dado que trabajamos con clasificadores de aprendizaje automático tradicionales (como *Random Forest*) que suelen responder bien ante cantidades de *features* acotadas. Es por esto que aplicamos el criterio de selección *mRMR*, que selecciona los atributos intentando maximizar relevancia y minimizar la redundancia de los mismos. También ejecutamos distintas pruebas para intentar identificar el número de coeficientes a seleccionar, probando con valores dentro del conjunto $[5, 10, 20, 30, 40, 50]$.

Presentamos a continuación los resultados obtenidos de este primer experimento, trabajando con los *features* extraídos del grafo resultante de cada sujeto, con los mejores coeficientes de *Pearson*, y con los atributos demográficos como predictores. Si bien nos enfocaremos en los resultados obtenidos por el clasificador de *Random Forest*, también presentamos los resultados obtenidos por el resto de los clasificadores. Además, detallamos los resultados tanto para el *dataset* completo (ver Tabla 6) como para el *dataset* reducido, obtenido luego del filtrado de

datos (ver Tabla 7). Adicionalmente, para el modelo que utiliza el *dataset* reducido agregamos un gráfico que indica la importancia de los *features* al momento de construir los árboles de decisión por *Random Forest* para clasificar a los individuos (ver Figura 11).

Cuadro 6: Tabla con los resultados tomando como predictores *features* extraídos del grafo resultante de cada sujeto, los mejores coeficientes de *Pearson* y datos demográficos, trabajando con el *dataset* completo.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	50 %	56 %	58 %	56 %
K-nearest Neighbors	51 %	55 %	52 %	53 %
Naive Bayes	46 %	61 %	42 %	42 %
Random Forest	**56 %**	**55 %**	**61 %**	**64 %**

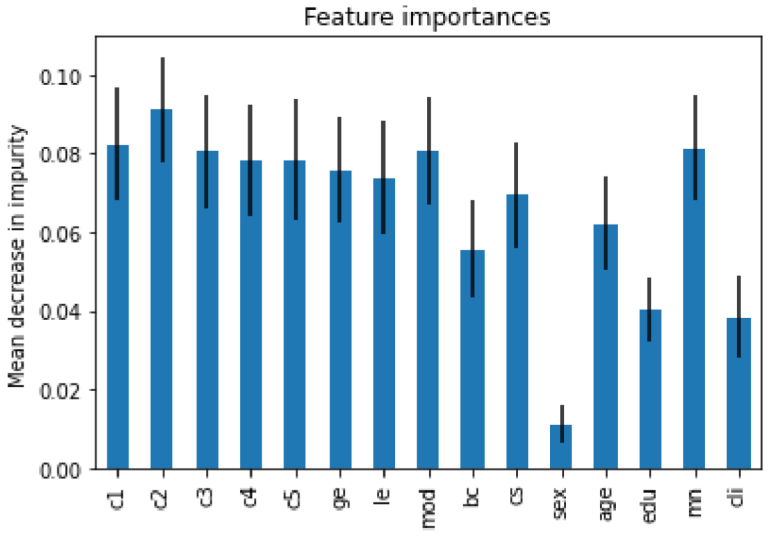
∗ *Atlas: Harvard-Oxford cortical*
∗ *Coeficiente de binarización:* 0.7

Cuadro 7: Tabla con los resultados tomando como predictores *features* extraídos del grafo resultante de cada sujeto, los mejores coeficientes de *Pearson* y datos demográficos, trabajando con el *dataset* reducido.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	52 %	58 %	58 %	58 %
K-nearest Neighbors	50 %	57 %	57 %	58 %
Naive Bayes	45 %	60 %	42 %	40 %
Random Forest	**57 %**	**58 %**	**64 %**	**66 %**

∗ *Atlas: Harvard-Oxford cortical*
∗ *Coeficiente de binarización:* 0.7

Figura 11: Relevancia de atributos utilizando un clasificador *Random Forest* y tomando como predictores *features* extraídos del grafo resultante de cada sujeto, los mejores coeficientes de *Pearson* y datos demográficos



* *c1,c2,c3,c4,c5: 5 mejores coeficientes de Pearson*
* *ge,le,mod,bc,cs: features extraídos de la red*
* *sex,age,edu,mn,cli: atributos demográficos*

Analizando el gráfico de importancia de *features* solamente podemos concluir que el sexo de los individuos parece importar muy poco al momento de construir los árboles de decisión para clasificar a los individuos. La importancia del resto de los *features* parece estar bastante repartida. Luego de obtener estos resultados poco alentadores, decidimos separar la experimentación en dos grandes configuraciones de *features*. Tratando así de poder analizar mejor el comportamiento de los distintos conjuntos de *features* por separado, siguiendo con foco en el clasificador de *Random Forest*, y trabajando tanto con el *dataset* completo como con el *dataset* reducido. Las configuraciones de *features* mencionadas son las siguientes:

- Los mejores coeficientes de *Pearson* seleccionados mediante el algoritmo de *mRMR*, más atributos demográficos (edad, sexo, nivel educativo, clínica y motion).

- Métricas extraídas de grafos construidos para cada sujeto (eficiencia global,

eficiencia local, modularidad, coeficiente de clustering y centralidad de intermediación), más atributos demográficos (edad, sexo, nivel educativo, clínica y motion).

En las dos siguientes secciones se profundizará sobre estas dos configuraciones, y es aquí donde encontramos los mejores resultados en general, validando hiperparámetros exhaustivamente (mediante validación cruzada) con el fin de calibrar al máximo nuestros modelos.

4.2.2. Atributos extraídos del grafo resultante y datos demográficos

Para esta configuración, trabajamos exclusivamente con los atributos demográficos que nos brinda el *dataset* de estudio y los atributos calculados del grafo, trabajando entonces con los siguientes *features*: eficiencia global, eficiencia local, modularidad, coeficiente de clustering, centralidad de intermediación, sexo, edad, nivel educativo, motion y clínica. Al igual que para la configuración inicial que utilizaba todos los *features* a disposición, si bien nos enfocaremos en los resultados obtenidos por el clasificador de *Random Forest*, también presentamos los resultados obtenidos por el resto de los clasificadores. Además, detallamos los resultados tanto para el *dataset* completo (ver Tabla 8) como para el *dataset* reducido, obtenido luego del filtrado de datos (ver Tabla 9). Adicionalmente, para el modelo que utiliza el *dataset* reducido agregamos un gráfico que indica la importancia de los *features* al momento de construir los árboles de decisión por *Random Forest* para clasificar a los individuos (ver Figura 12). Al igual que en el primer experimento, el atributo demográfico del sexo parece ser el menos relevante para clasificar a los individuos. Mientras tanto, los restantes *features* utilizados parecen ser importantes para la clasificación, tanto las cinco métricas calculadas sobre las redes, cómo también tres de los cinco atributos demográficos (sobresaliendo el nivel educativo, la edad y el atributo motion obtenido durante la resonancia). Finalmente, agregamos una comparativa de los resultados obtenidos por *Random Forest* al trabajar por un lado con el *dataset* completo, y por otro lado con el *dataset* reducido (ver Figura 13). Se ve una notoria mejoría en los resultados cuando utilizamos el conjunto de datos filtrado.

Cuadro 8: Tabla con los resultados tomando como predictores *features* extraídos del grafo resultante de cada sujeto y datos demográficos, trabajando con el *dataset* completo.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	55 %	53 %	87 %	66 %
K-nearest Neighbors	59 %	58 %	62 %	60 %
Naive Bayes	52 %	68 %	58 %	40 %
Random Forest	**62 %**	**60 %**	**67 %**	**64 %**

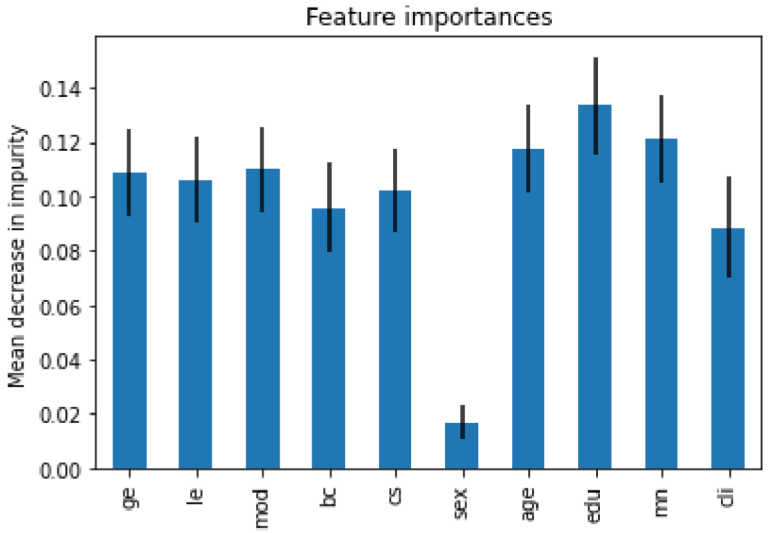
* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Cuadro 9: Tabla con los resultados tomando como predictores *features* extraídos del grafo resultante de cada sujeto y datos demográficos, trabajando con el *dataset* reducido.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	60 %	58 %	71 %	64 %
K-nearest Neighbors	61 %	60 %	67 %	62 %
Naive Bayes	57 %	55 %	64 %	59 %
Random Forest	**69 %**	**65 %**	**77 %**	**71 %**

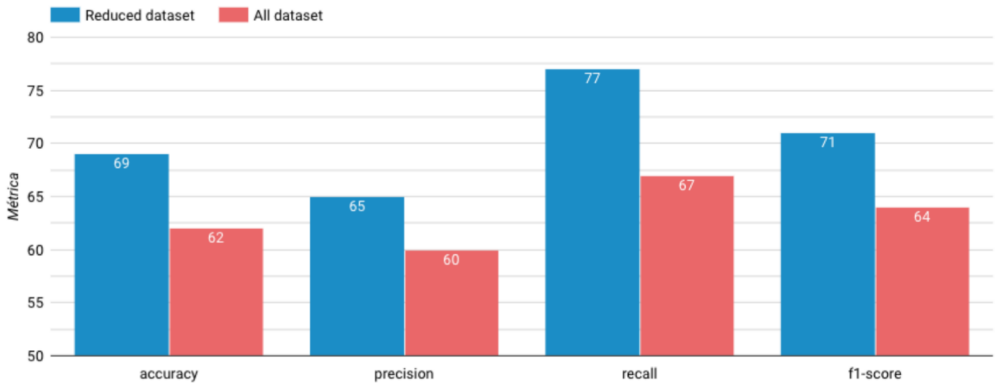
* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Figura 12: Relevancia de features utilizando un clasificador *Random Forest* y tomando como predictores features extraídos del grafo resultante de cada sujeto y datos demográficos.



* *ge,le,mod,bc,cs: features extraídos de la red*
* *sex,age,edu,mn,cli: atributos demográficos*

Figura 13: Comparativa de los resultados obtenidos por *Random Forest* con predictores extraídos del grafo y demográficos, al trabajar por un lado con el *dataset* completo, y por otro lado con el *dataset* reducido.



4.2.3. Coeficientes absolutos de Pearson y datos demográficos

En este experimento, planteamos la selección de los cinco mejores coeficientes absolutos de *Pearson* mediante el método de selección *mRMR*. Observamos que no hace falta transformar la matriz de correlación a un grafo, ya que solamente trabajaremos con los coeficientes de *Pearson* calculados. Al igual que para las configuraciones anteriores, presentamos los resultados obtenidos por todos los clasificadores (con foco en los obtenidos por el clasificador de *Random Forest*). Detallamos los resultados tanto para el *dataset* completo (ver Tabla 10) como para el *dataset* reducido (ver Tabla 11). Y presentamos el gráfico que indica la importancia de los *features* al momento de construir los árboles de decisión por *Random Forest* para clasificar a los individuos del *dataset* reducido (ver Figura 14). La importancia parece comportarse de manera similar a los experimentos anteriores, en donde sin tener en cuenta al atributo demográfico del sexo, y un poco más arriba (en grado de relevancia) la clínica donde se obtuvieron las resonancias magnéticas, el resto de los *features* se visualizan útiles para la clasificación. En la Figura 15 también agregamos la comparativa de los resultados obtenidos por *Random Forest* al trabajar por un lado con el *dataset* completo, y por otro lado con el *dataset* reducido. Nuevamente encontramos una mejora importante en los resultados cuando utilizamos el conjunto de datos filtrado.

Cuadro 10: Tabla con los resultados tomando como predictores los mejores coeficientes absolutos de *Pearson* y datos demográficos, trabajando con el *dataset* completo.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	60 %	60 %	58 %	60 %
K-nearest Neighbors	59 %	68 %	60 %	62 %
Naive Bayes	58 %	66 %	52 %	56 %
Random Forest	**60 %**	**58 %**	**73 %**	**65 %**

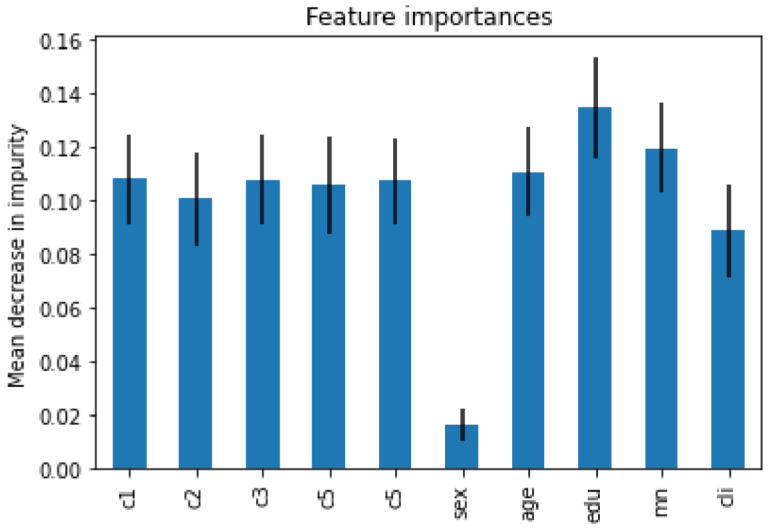
∗ *Atlas: Harvard-Oxford cortical*
∗ *Coeficiente de binarización:* 0.7

Cuadro 11: Tabla con los resultados tomando como predictores los mejores coeficientes absolutos de *Pearson* y datos demográficos, trabajando con el *dataset* reducido.

Clasificador	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	60 %	69 %	55 %	61 %
K-nearest Neighbors	61 %	68 %	60 %	64 %
Naive Bayes	59 %	68 %	52 %	59 %
Random Forest	**71 %**	**77 %**	**71 %**	**74 %**

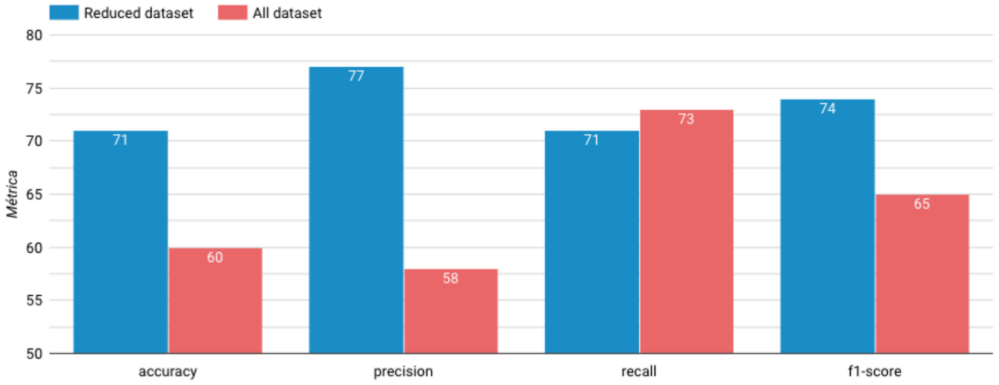
* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Figura 14: Relevancia de *features* utilizando un clasificador *Random Forest* y tomando como predictores los mejores coeficientes absolutos de *Pearson* y datos demográficos.



* *c1,c2,c3,c4,c5: 5 mejores coeficientes de Pearson*
* *sex,age,edu,mn,cli: atributos demográficos*

Figura 15: Comparativa de los resultados obtenidos por *Random Forest* al trabajar por un lado con el *dataset* completo, y por otro lado con el *dataset* reducido.



4.3. Búsqueda de hiper-parámetros

Debido a la cantidad de valores de hiper-parámetros que queremos evaluar, sumado esto a los distintos *atlas* a estudiar, y también las distintas variantes planteadas en relación a los *features*, necesitamos más de 5000 entrenamientos de *Random Forest* para finalizar todas las combinaciones planteadas. Tal cual describimos previamente, en cada entrenamiento se utiliza validación cruzada, con el 90 % del *dataset* reducido, dejando de lado un 10 % de los datos para la evaluación final del modelo.

En nuestro estudio utilizamos el clasificador de *Random Forest* provisto por *scikit learn* [57], biblioteca de aprendizaje automático desarrollada para Python. Se decidió ejecutar la búsqueda de hiper-parámetros en el Centro Nacional de Supercomputación *Cluster UY* [63]. La misma es una plataforma de computación de alto desempeño que posee la capacidad de gestionar en forma coordinada múltiples recursos de cómputo. Está formado por una infraestructura computacional que tiene un poder superior a 10.000 computadores tradicionales, y es capaz de realizar operaciones complejas y de largo aliento en poco tiempo.

4.3.1. Ejecución de opciones

Para esto utilizamos SLURM [64], encargado de gestionar los recursos de cómputo en *Cluster UY* y planificar la ejecución de trabajos. Al avanzar en la validación

de hiper-parámetros, los árboles de decisión de *Random Forest* se tornan cada vez más profundos (ya que la profundidad de las ramas es uno de los hiper-parámetros que validamos), y por este motivo no podemos predecir a priori cuánto demorará el total de las ejecuciones. Utilizamos entonces trabajos fuera de línea, que fuimos ejecutando cada 5 días. También fue necesario utilizar Conda [65] para configurar el entorno de ejecución con Python 3.8 y todas las librerías necesarias. A continuación se puede ver uno de los *jobs* ejecutados a la hora de realizar la validación de hiper-parámetros:

```
#!/bin/bash
#SBATCH --job-name=pgrado
#SBATCH --ntasks=1
#SBATCH --mem=64GB
#SBATCH --time=5-00:00:00
#SBATCH --tmp=9G
#SBATCH --partition=besteffort
#SBATCH --qos=besteffort
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxxxx
source /clusteruy/home/agustina.sierra.lima/miniconda3/bin/activate
cd /clusteruy/home/agustina.sierra.lima/pgrado
python3 cross_val.py
```

El resultado de esta validación cruzada fue un archivo CSV con 4033 filas y con 4 columnas correspondientes a las métricas de *Accuracy, Precision, Recall y F1-Score*.

4.3.2. Mejores combinaciones de hiper-parámetros

De todas las combinaciones de hiper-parámetros ejecutadas en la validación cruzada, presentamos a continuación dos tablas con cinco de las combinaciones que obtuvieron los mejores resultados en la validación. Tenemos una tabla para cada variante implementada, esto sería tanto para la variante que utiliza *features* extraídos del grafo (Tabla 12), como para la variante que utiliza coeficientes absolutos de *Pearson* (Tabla 13).

Una observación importante, es que los mejores resultados se dieron todos con el *atlas* denominado *Harvard-Oxford cortical*. Además, los resultados obtenidos varían desde 53 % de *accuracy* hasta **73 %** de *accuracy*.

Cuadro 12: Tabla con la combinación de hiper-párametros que obtuvo los cinco mejores resultados en la validación cruzada, para la variante con *features* extraídos del grafo más atributos demográficos.

ID Instancia	Accuracy	Precision	Recall	F1-Score
669	73 %	73 %	73 %	73 %
611	73 %	72 %	74 %	73 %
278	73 %	74 %	71 %	72 %
614	73 %	72 %	74 %	73 %
375	73 %	72 %	73 %	73 %

* *Hiper-parametro Atlas: Harvard-Oxford cortical*
* *Hiper-parametro Coeficiente de binarización:* 0.7
* *Variante: features extraídas del grafo resultante de cada sujeto más atributos demográficos*

Cuadro 13: Tabla con la combinación de hiperparámetros que obtuvo los cinco mejores resultados en la validación cruzada, para la variante que utiliza como *features* los coeficientes absolutos de *Pearson* más atributos demográficos.

ID Instancia	Accuracy	Precision	Recall	F1-Score
626	73 %	73 %	72 %	72 %
628	73 %	72 %	72 %	72 %
640	73 %	72 %	73 %	73 %
625	73 %	71 %	75 %	73 %

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7
* *Variante: Coeficientes absolutos de Pearson más atributos demográficos*

Para que ésta información sea útil y reproducible, las Tablas 14 y 15 presentan

los valores de los hiper-parámetros correspondientes a cada instancia identificada en los mejores resultados de cada variante.

Cuadro 14: Tabla con detalle de hiper-parámetros que obtuvieron los cinco mejores resultados en la validación cruzada, para la variante con *features* extraídos del grafo más atributos demográficos.

IDInstancia	n_estimators	criterion	warm_start	max_features	bootstrap	max_depth
669	100	entropy	false	log2	false	40
611	100	entropy	true	sqrt	false	20
278	20	entropy	true	sqrt	false	50
614	100	entropy	true	sqrt	false	50
375	50	gini	true	log2	false	40

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Cuadro 15: Tabla con detalle sobre los hiper-parámetros que obtuvieron los cinco mejores resultados en la validación cruzada, para la variante que utiliza como *features* los coeficientes absolutos de *Pearson* más atributos demográficos

ID Instancia	n_estimators	criterion	warm_start	max_features	bootstrap	max_depth
626	100	entropy	true	log2	false	30
628	100	entropy	true	log2	false	50
640	100	entropy	false	auto	false	30
625	100	entropy	true	log2	false	20
668	100	entropy	false	log2	false	30

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Luego de ejecutar la búsqueda de hiper-parámetros logramos obtener un *accuracy* similar a la vista en los artículos [11, 12, 13], los cuales fueron nuestro punto de partida desde el inicio del trabajo. No obstante terminamos implementando una solución distinta a la planteada inicialmente, ya que utilizando el total de los *features* disponibles no logramos conseguir buenos resultados. Las dos variantes de modelos construídos lograron muy buenos resultados, siempre utilizando el conjunto de datos reducido obtenido luego del filtrado (recordemos que con el *dataset* completo no logramos llegar al 70 % de exactitud en los resultados).

A continuación veremos dos sub-experimentos que realizamos con motivaciones

bien diferentes. Una fue poder ver de algún modo que tan sensible es nuestro modelo al conjunto de datos utilizado. Con esto nos referimos por ejemplo a que en varios grupos de investigación se utilizaron escáneres diferentes a la hora de realizar la imagen *fMRI* para cada paciente, entonces, ¿Qué sucederá si ejecutamos el modelo dividiendo los datos por grupos de investigación?
La segunda motivación es buscar un algoritmo de clasificación que funcione mejor con una gran cantidad de *features*, y remover de nuestro *pipline* el método de selección de atributos, que en principio no parece brindar muchos dividendos, con el objetivo de clasificar a un individuo utilizando todo el conjunto de *features* disponible (tanto coeficientes absolutos de *Pearson*, cómo las métricas obtenidas a partir de los grafos y los datos demográficos).

4.4. Impacto de centros de investigación en calidad de resultados

Como se comentó anteriormente, aquí trabajamos con los mismos *features*, pero esta vez trabajando con todos los grupos de investigación por separado (particionando el *dataset*). Este experimento intenta identificar como se comporta cada grupo de investigación de manera independiente, sin combinar con pacientes de otras clínicas, con el objetivo de obtener mejores resultados. Contamos con un total de 25 grupos de investigación, de los cuales podemos utilizar 24, puesto que los datos del grupo de investigación identificado con el número 4 se encuentran corruptos, y no fueron utilizados para ningún estudio.

Presentaremos los resultados obtenidos en esta situación planteada trabajando únicamente con el clasificador de *Random Forest*, y al igual que en los experimentos anteriores, detallamos los mismos trabajando con el *dataset* completo y también con el *dataset* reducido.

Trabajando con el *dataset* completo, ver Tabla 16, notamos que en varios grupos de investigación se obtienen resultados muy buenos que superan el **80 %** en todas las métricas estudiadas, aunque cada uno de ellos presenta como máximo 100 individuos. Existen algunos pocos grupos de investigación que presentan resultados muy malos sin importar que la cantidad de individuos sea acotada. En la mayoría de los grupos de investigación, se obtienen resultados cercanos al **60 %**, **70 %** en todas las métricas.

Al trabajar con el *dataset* reducido, ver Tabla 17, la calidad de los resultados se incrementa considerablemente, obteniendo resultados positivos en todos los centros de investigación. El filtrado del *dataset* justamente deja de lado los grupos de investigación con muy pocos individuos, y también filtra las resonancias de mala calidad, por lo que maximizamos la exactitud de nuestros modelos. Notamos que en varios grupos de investigación se obtienen resultados muy buenos que superan el **80 %** en todas las métricas estudiadas, y estos resultados no ocurren solamente en los centros de investigación con pocos individuos, también se manifiestan en los centros con más cantidad de sujetos, ejemplo en centro número 20, con 479 individuos, y centro número 8, con 116 (recordemos que estamos trabajando con el *dataset* reducido, originalmente estos dos grupos de investigación estaban formados por 533 y 150 individuos respectivamente).

Estos resultados muestran que el objetivo de estudio es complejo desde el momento que se recabaron los datos, dado que si acotamos el universo de los datos a grupos de estudio, los resultados son realmente altos. Esto puede deberse a diversos factores, como la utilización de la misma maquinaria para extraer las imágenes de resonancia, así como definir el mismo tamaño de *voxel* para todos los sujetos. Este tipo de detalles escapan un poco de nuestro conocimiento, pero hacen a la extracción de imágenes *fMRI*.

4.5. Evaluación de técnicas adicionales, completitud de features

Luego de analizar los resultados obtenidos en los experimentos previos, nos parece interesante implementar algún modelo de aprendizaje automático que utilice la totalidad de los coeficientes obtenidos directamente luego de aplicar la correlación de *Pearson*.

Dado que la magnitud de estos coeficientes suele ser grande dependiendo del *atlas* utilizado (en el orden de centenas, incluso miles) debemos seleccionar algún modelo que se comporte de buena manera trabajando con cantidades de *features* elevadas. Planteamos entonces la utilización de dos nuevos clasificadores, *xgboost* [51] y *lightgbm* [52], y construimos los modelos respectivos trabajando con todos los coeficientes de *Pearson*, sin incluir los *features* globales a la red, ni tampoco los atributos demográficos. Para este escenario también detallamos los

Cuadro 16: Tabla de resultados particionando el *dataset* por grupo de investigación, trabajando con el *dataset* completo.

Clínica	Individuos	Accuracy	Precision	Recall	F1-Score
ID1	148	57 %	56 %	66 %	60 %
ID2	60	58 %	66 %	57 %	62 %
ID3	64	69 %	67 %	40 %	50 %
ID4	-	-	-	-	-
ID5	24	40 %	10 %	40 %	57 %
ID6	30	50 %	50 %	33 %	40 %
ID7	87	27 %	20 %	10 %	13 %
ID8	150	67 %	78 %	47 %	58 %
ID9	**100**	**80 %**	**80 %**	**80 %**	**80 %**
ID10	**83**	**65 %**	**73 %**	**73 %**	**73 %**
ID11	61	53 %	50 %	66 %	57 %
ID12	**38**	**100 %**	**100 %**	**100 %**	**100 %**
ID13	42	30 %	25 %	100 %	400 %
ID14	**96**	**80 %**	**90 %**	**77 %**	**83 %**
ID15	100	60 %	25 %	16 %	20 %
ID16	62	54 %	60 %	43 %	50 %
ID17	91	32 %	28 %	57 %	40 %
ID18	**41**	**89 %**	**80 %**	**100 %**	**89 %**
ID19	**87**	**78 %**	**100 %**	**64 %**	**78 %**
ID20	**533**	**66 %**	**67 %**	**64 %**	**65 %**
ID21	156	50 %	60 %	50 %	55 %
ID22	50	50 %	60 %	55 %	53 %
ID23	62	54 %	50 %	50 %	50 %
ID24	63	54 %	57 %	57 %	57 %
ID25	**152**	**67 %**	**66 %**	**75 %**	**70 %**

∗ *Atlas: Harvard-Oxford cortical*

∗ *Coeficiente de binarización:* 0.7

Cuadro 17: Tabla de resultados particionando el *dataset* por grupo de investigación, trabajando con el *dataset* reducido.

Clínica	Individuos	Accuracy	Precision	Recall	F1-Score
ID1	**146**	**67 %**	**55 %**	**83 %**	**67 %**
ID2	**30**	**100 %**	**100 %**	**100 %**	**100 %**
ID3	-	-	-	-	-
ID4	-	-	-	-	-
ID5	-	-	-	-	-
ID6	-	-	-	-	-
ID7	71	63 %	50 %	66 %	57 %
ID8	**116**	**92 %**	**100 %**	**80 %**	**89 %**
ID9	**96**	**70 %**	**67 %**	**50 %**	**57 %**
ID10	**71**	**63 %**	**67 %**	**80 %**	**73 %**
ID11	**37**	**100 %**	**100 %**	**100 %**	**100 %**
ID12	-	-	-	-	-
ID13	**36**	**75 %**	**100 %**	**50 %**	**67 %**
ID14	**93**	**80 %**	**86 %**	**86 %**	**86 %**
ID15	**67**	**71 %**	**67 %**	**67 %**	**67 %**
ID16	-	-	-	-	-
ID17	82	56 %	75 %	50 %	60 %
ID18	-	-	-	-	-
ID19	49	60 %	33 %	100 %	50 %
ID20	**479**	**81 %**	**81 %**	**84 %**	**82 %**
ID21	**144**	**67 %**	**64 %**	**88 %**	**74 %**
ID22	**38**	**75 %**	**67 %**	**100 %**	**80 %**
ID23	45	60 %	50 %	50 %	50 %
ID24	-	-	-	-	-
ID25	-	-	-	-	-

∗ *Atlas: Harvard-Oxford cortical*

∗ *Coeficiente de binarización:* 0.7

resultados primero trabajando con el *dataset* completo (ver Tabla 18), y luego utilizando el *dataset* reducido (ver Tabla 19).

Adicionalmente, agregamos un nuevo experimento, y en este caso además de utilizar todos los coeficientes absolutos de *Pearson*, construimos nuevamente un grafo, y sumamos a estos coeficientes las métricas globales calculadas sobre los grafos, y también agregamos los atributos demográficos brindados por el *dataset*. Estamos trabajando entonces con todos los *features* disponibles, sin la necesidad de escoger los mejores, ya que estos modelos suelen performar de buena manera con muchos *features*. Para esta variante también detallamos los resultados trabajando con el *dataset* completo (ver Tabla 20), y luego con el *dataset* reducido (ver Tabla 21). Podemos resaltar que para este último experimento, en la cual utilizamos todos los *features* disponibles, alcanzamos muy buenos resultados, con métricas muy similares a las obtenidas por nuestros modelos base detallados en la sección de la solución. Además, alcanzamos estos resultados sin implementar validación de hiper-parámetros sobre nuestros modelos de *xgboost* y *lightgbm*, por lo que la utilización de los mismos parece acompañar de buena manera a nuestro estudio, y puede ser un buen punto de partida para seguir experimentando sobre la temática, ya que los mismos permiten trabajar con cantidades de *features* grandes, y seguir desarrollando nuevas variantes.

Cuadro 18: Tabla de resultados obtenidos al utilizar las herramientas *xgboost* y *lightgbm*, trabajando con todos los coeficientes absolutos de *Pearson* como *features*, y con el *dataset* completo.

Clasificador	Accuracy	Precision	Recall	F1-Score
xgboost	58 %	56 %	71 %	63 %
lightgbm	57 %	57 %	74 %	64 %

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Cuadro 19: Tabla de resultados obtenidos al utilizar las herramientas *xgboost* y *lightgbm*, trabajando con todos los coeficientes absolutos de *Pearson* como *features*, y con el *dataset* reducido.

Clasificador	Accuracy	Precision	Recall	F1-Score
xgboost	60 %	58 %	59 %	59 %
lightgbm	62 %	59 %	68 %	62 %

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Cuadro 20: Tabla de resultados obtenidos al utilizar las herramientas *xgboost* y *lightgbm*, trabajando con todos los *features* disponibles, y con el *dataset* completo.

Clasificador	Accuracy	Precision	Recall	F1-Score
xgboost	63 %	65 %	73 %	69 %
lightgbm	59 %	60 %	72 %	66 %

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

Cuadro 21: Tabla de resultados obtenidos al utilizar las herramientas *xgboost* y *lightgbm*, trabajando con todos los *features* disponibles, y con el *dataset* reducido.

Clasificador	Accuracy	Precision	Recall	F1-Score
xgboost	68 %	66 %	70 %	68 %
lightgbm	68 %	65 %	73 %	69 %

* *Atlas: Harvard-Oxford cortical*
* *Coeficiente de binarización:* 0.7

5. Conclusiones

Se experimentaron exhaustivamente distintos métodos y variantes para intentar aproximar una mejor solución al problema de predecir *MDD* a partir de imágenes *fMRI*, trabajando con una cantidad elevada de pacientes.

Logramos validar que al trabajar con magnitudes grandes de individuos (grandes para esta temática), se pierde bastante la exactitud de los resultados, ratificando lo que obtienen la mayoría de los artículos que estudian este difícil problema. También visualizamos que una de las razones principales por la cual perdemos exactitud, es la heterogeneidad con la que se recolectan las imágenes *fMRI*, y la variedad de los escáners utilizados en la resonancia. Por estos motivos, uno de los puntos que marcamos como posible mejora sería protocolizar la obtención de las imágenes, la calibración y variabilidad de los escáneres, y todo aquello que diferencia las medidas entre clínicas (punto que escapa bastante de nuestro alcance).

Logramos replicar de buena manera el método que nos definimos como punto base de trabajo, experimentando también distintas variantes que se desprendieron de la tarea inicial, y que nos ayudaron a tener una mejor perspectiva a la hora de construir una solución final, tratando de explotar al máximo los conocimientos adquiridos en cada etapa.

Los resultados obtenidos por nuestros modelos parecen ser alentadores, incluso superando los resultados obtenidos por otros artículos que trabajan sobre el mismo *dataset*, y aplicando métodos de aprendizaje automático similares. De todos los clasificadores que hemos utilizado en la experimentación, concluimos que los clasificadores basados en árboles de decisión pueden ser los más adecuados para tratar esta temática de las *fMRI*, apoyándonos también en los artículos estudiados. La experimentación que utiliza como predictores métricas extraídas de los grafos construidos para cada paciente, combinando atributos demográficos, fue el modelo que obtuvo resultados más confiables luego de la validación de hiper-parámetros. Si bien obtuvimos resultados muy similares con la experimentación que selecciona los mejores coeficientes absolutos de *Pearson* como *features*, no identificamos resultados concluyentes en relación a la cantidad de coeficientes a seleccionar, parámetro que impacta directamente al modelo implementado.

Entrando en detalle de la solución final implementada, podemos afirmar que agregar *features* demográficos a los atributos calculados de las redes (o directa-

mente sobre los coeficientes de *Pearson*), implica un mejor rendimiento en general del algoritmo. Esta es una diferencia grande en términos del método implementado en comparación con los artículos que utilizamos como base de nuestro trabajo.

También cabe destacar que cada parámetro de entrada y cada *feature* a seleccionar, está totalmente atado al *atlas* a utilizar para la experimentación, ya que para el dataset utilizado se pueden seleccionar distintos *atlas*, donde las regiones de interés definidas pueden variar considerablemente. Un *atlas* con pocas *ROI's* tendrá por consiguiente menos nodos en el grafo, y si al momento de definir el coeficiente de binarización, utilizamos un valor muy alto, esto podría desembocar en un grafo poco conexo y pequeño. Pasaría lo contrario si utilizamos un *atlas* con gran cantidad de regiones de interés y un coeficiente de binarización bajo. Finalmente, llegada la hora de extraer *features* de estos grafos, los mismos pueden identificar comportamientos muy diferentes en base a las decisiones tomadas previamente. Esto revela que el universo de posibles grafos para representar la actividad cerebral de un paciente es infinito.

Consideramos que nuestra solución está totalmente alineada con los experimentos realizados por diversos científicos que abordan la temática de las *fMRI*, siendo un problema muy desafiante, y que todavía no se ha encontrado un esquema adecuado para atacar la solución del mismo.

Pensando en lo que sería el trabajo futuro y cómo seguir investigando en esta temática, podemos decir que establecimos un nuevo punto de partida, del cual se pueden realizar diversos estudios, trabajando por ejemplo con nuevos *features*, nuevas formas de preprocesar el *dataset* y sacar información lo menos ruidosa posible, experimentar con otros *atlas*, incluso con nuevos modelos de aprendizaje automático (al menos probando dentro de las distintas variantes que proveen los árboles de decisión). Experimentar con nuevos *features* de acuerdo a los *atlas* utilizados puede generar un gran impacto en cuanto a los resultados obtenidos.

En nuestro estudio reducimos la selección de *features* solamente a trabajar con funciones de correlación y a construir grafos a partir de las mismas, pero creemos que aportaría valor incursionar en otros enfoques a partir de las *fMRI*, sin dejar de lado los atributos demográficos de los pacientes, que sin lugar a dudas acompañan la precisión de las predicciones.

También se abre un abanico grande de posibilidades para seguir trabajando utilizando clasificadores que funcionan bien con una gran cantidad de *features*

como *xgboost* y *lightgbm*, con los cuales obtuvimos muy buenos resultados (al nivel de las variantes implementadas en nuestra solución final), aunque no abordamos el proceso completo con la validación de hiper-parámetros pertinente. Estos modelos nos pueden permitir combinar una gran cantidad *features*, un ejemplo concreto podría ser incluir métricas locales a cada nodo si estamos trabajando con grafos (además de las métricas globales utilizadas).

Como conclusión final de todo el trabajo realizado, quedamos muy satisfechos con el estudio que logramos llevar adelante, en un campo sumamente complejo y desconocido para nosotros, que implica una curva de aprendizaje realmente desafiante. Logramos comprender conceptos básicos de como funcionan las *fMRI*, qué intentan medir y cómo lo hacen, y a partir de las mismas entrar en un campo en el que quizás nos sentimos más cómodos, cómo son los grafos, extraer métricas relevantes de los mismos e involucrar conceptos matemáticos y de optimización para lograr caracterizar a cierto individuo. Si bien conocemos un poco más del mundo de aprendizaje automático, y tenemos presente ciertos problemas habituales en este campo como puede ser el sobreajuste, intentamos construir un modelo clasificador lo más acertado posible, tratando de validar distintos hiper-parámetros que hacen a la precisión del modelo. Podemos decir que intentamos completar todas las etapas que requiere un proyecto de investigación con esta temática, aprendimos sobre el universo del problema, extraímos distintos tipos de *features*, elegimos un modelo clasificador y validamos sus correspondientes hiper-parámetros. Además, dejamos la puerta abierta para seguir investigando distintas variantes que parecen ser alentadoras.

6. Referencias

[1] C Bohm, T Greitz, R Seitz, and L Eriksson. Specification and selection of regions of interest (rois) in a computerized brain atlas. *Journal of Cerebral Blood Flow & Metabolism*, 11(1_suppl):A64–A68, 1991.

[2] David J Heeger and David Ress. What does fmri tell us about neuronal activity? *Nature reviews neuroscience*, 3(2):142–151, 2002.

[3] Christian Otte, Stefan M Gold, Brenda W Penninx, Carmine M Pariante, Amit Etkin, Maurizio Fava, David C Mohr, and Alan F Schatzberg. Major depressive disorder. *Nature reviews Disease primers*, 2(1):1–20, 2016.

[4] Robert M Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Physics in Medicine & Biology*, 37(3):705, 1992.

[5] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[6] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

[7] Max Roser Saloni Dattani, Hannah Ritchie. Mental health. 2019.

[8] Calhoun V. D. Sui J Gao, S. Machine learning in major depression: From classification to treatment outcome prediction. 24:1037–1052, 2018.

[9] Vinod Menon Gary H Glover Hugh B Solvason Heather Kenna Allan L Reiss Alan F Schatzberg Michael D Greicius, Benjamin H Flores. Resting-state functionalconnectivity in major depression: abnormally increased contributions from subgen ual cingulate cortexand thalamus. vol. 62:429–437, 2007.

[10] I B Hickie J Lagopoulos L Wang, D F Hermens. A systematic review of resting state functional-mri studies in major depression. vol. 142:6–12, 2012.

[11] Bonnie Klimes-Dougan Kathryn Cullen Bhaskar Sen, Bryon Mueller and Keshab K. Parhi1. Classification of major depressive disorder from resting-state fmri. 2019.

[12] Lara C. Foland-Ross Paul M. Thompson Matthew D. Sacchet, Gautam Prasad and Ian H. Gotlib. Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. 2015.

[13] Baolin Liu Yonggang Shi Xiangfei Geng, Junhai Xu. Multivariate classification of major depressive disorder using the effective connectivity and functional connectivity. 2018.

[14] Le Li Yu-Feng Zang Chao-Gan Ya, Xiao Chen. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. vol. 4, 2019.

[15] Owen J. Arthurs and Simon Boniface. How well do we understand the neural origins of the fmri bold signal? 25, 2002.

[16] Mark D'Esposito3-4 Daniel A. Handwerker1, Javier Gonzalez-Castillo1 and Peter A. Bandettini. The continuing challenge of understanding and modeling hemodynamic variation in fmri. 62, 2002.

[17] b-[U+204E] Liana Portugal-a c Tim Hahn d Andreas J. Fallgatter e Marta I. Garrido f g h John Shawe-Taylor a Maria J. Rosa, a and Janaina Mourao-Mirandaa. Sparse network-based models for patient classification using fmri. 105, 2015.

[18] Ling-Li Zeng, Hui Shen, Li Liu, and Dewen Hu. Unsupervised classification of major depression using functional connectivity mri. *Human brain mapping*, 35(4):1630–1641, 2014.

[19] P. Thomas Fletcher Andrew L. Alexander Nicholas Lange7 8 Erin D. Bigler Janet E. Lainhar5 Jared A. Nielsen, Brandon A. Zielinski and Jeffrey S. Anderson1. Multisite functional connectivity mri classification of autism: Abide results. 2013.

[20] Kosuke Yoshida, Yu Shimizu, Junichiro Yoshimoto, Masahiro Takamura, Go Okada, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya. Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional mri data with partial least squares regression. *PloS one*, 12(7):e0179638, 2017.

[21] Xue Zhong, Huqing Shi, Qingsen Ming, Daifeng Dong, Xiaocui Zhang, Ling-Li Zeng, and Shuqiao Yao. Whole-brain resting-state functional connectivity identified major depressive disorder: a multivariate pattern analysis in two independent samples. *Journal of Affective Disorders*, 218:346–352, 2017.

[22] Xin Wang, Yanshuang Ren, and Wensheng Zhang. Depression disorder classification of fmri data using sparse low-rank functional brain network and graph-based features. *Computational and mathematical methods in medicine*, 2017, 2017.

[23] Benedikt Sundermann, Stephan Feder, Heike Wersching, Anja Teuber, Wolfram Schwindt, Harald Kugel, Walter Heindel, Volker Arolt, Klaus Berger, and Bettina Pfleiderer. Diagnostic classification of unipolar depression based on resting-state functional connectivity mri: effects of generalization to a diverse sample. *Journal of Neural Transmission*, 124(5):589–605, 2017.

[24] Runa Bhaumik, Lisanne M Jenkins, Jennifer R Gowins, Rachel H Jacobs, Alyssa Barba, Dulal K Bhaumik, and Scott A Langenecker. Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity. *NeuroImage: Clinical*, 16:390–398, 2017.

[25] Hao Guo, Chen Cheng, Xiaohua Cao, Jie Xiang, Junjie Chen, and Kerang Zhang. Resting-state functional connectivity abnormalities in first-onset unmedicated depression. *Neural regeneration research*, 9(2):153, 2014.

[26] Anton Lord, Dorothea Horn, Michael Breakspear, and Martin Walter. Changes in community structure of resting state functional connectivity in unipolar depression. 2012.

[27] Longlong Cao, Shuixia Guo, Zhimin Xue, Yong Hu, Haihong Liu, Tumbwene E Mwansisya, Weidan Pu, Bo Yang, Chang Liu, Jianfeng Feng, et al. Aberrant functional connectivity for diagnosis of major depressive disorder: a discriminant analysis. *Psychiatry and clinical neurosciences*, 68(2):110–119, 2014.

[28] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[29] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[30] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Spiking neural networks. *International journal of neural systems*, 19(04):295–308, 2009.

[31] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. *Advances in neural information processing systems*, 17, 2004.

[32] Submitted by YAN Chao-Gan. Data sharing of the rest-meta-mdd project from the direct consortium. 2019.

[33] Yachen Shi, Linhai Zhang, Zan Wang, Xiang Lu, Tao Wang, Deyu Zhou, and Zhijun Zhang. Multivariate machine learning analyses in identification of major depressive disorder using resting-state functional connectivity: a multicentral study. *ACS Chemical Neuroscience*, 12(15):2878–2886, 2021.

[34] Abubakar Ado, Mustafa Mat Deris, Noor Azah Samsudin, and Ahmed Aliyu. A new feature filtering approach by integrating ig and t-test evaluation metrics for text classification. *International Journal of Advanced Computer Science and Applications*, 12(6), 2021.

[35] Jose Manuel Calabuig, Lluis Miquel Garcia Raffi, and Enrique Alfonso Sánchez-Perez. Á´lgebra lineal y descomposición en valores singulares. *Modelling in Science Education and Learning*, 8(2):133–144, 2015.

[36] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[37] John Ashburner, Gareth Barnes, Chun-Chuan Chen, Jean Daunizeau, Guillaume Flandin, Karl Friston, Stefan Kiebel, James Kilner, Vladimir Litvak, Rosalyn Moran, et al. Spm12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, 2464:4, 2014.

[38] Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[39] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[40] Chaogan Yan and Yufeng Zang. Dparsf: a matlab toolbox for"pipeline"data analysis of resting-state fmri. *Frontiers in systems neuroscience*, page 13, 2010.

[41] Yuanwei Xie Philip Moore Jiaxiang Zheng Zhijun Yao, Bin Hu. A review of structural and functional brain networks: small world and atlas. 2015.

[42] Jorge Camacho-Sandoval. Asociación entre variables cuantitativas: análisis de correlación. *Acta Médica Costarricense*, 50(2):94–96, 2008.

[43] Kenneth J Berry and Paul W Mielke Jr. A monte carlo investigation of the fisher z transformation for normal and nonnormal distributions. *Psychological Reports*, 87(3_suppl):1101–1114, 2000.

[44] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer New York, 2009.

[45] F. Long H. Peng and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. vol. 27 no. 8:1226–1238, 2005.

[46] Peter E Latham and Yasser Roudi. Mutual information. *Scholarpedia*, 4(1):1658, 2009.

[47] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62, 2017.

[48] Jan Salomon Cramer. The origins of logistic regression. 2002.

[49] Oliver Kramer. K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23. Springer, 2013.

[50] Harry Zhang. The optimality of naive bayes. *Aa*, 1(2):3, 2004.

[51] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[52] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.

[53] Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications*, pages 246–252. Springer, 2012.

[54] Rupali Bhardwaj and Sonia Vatta. Implementation of id3 algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 2013.

[55] Andrew Vince. A framework for the greedy algorithm. *Discrete Applied Mathematics*, 121(1-3):247–260, 2002.

[56] Hervé Abdi. Singular value decomposition (svd) and generalized singular value decomposition. *Encyclopedia of measurement and statistics*, pages 907–912, 2007.

[57] Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer, 2016.

[58] Stéphane Mussard, Françoise Seyte, and Michel Terraza. Decomposition of gini and the generalized entropy inequality measures. *Economics Bulletin*, 4(7):1–6, 2003.

[59] Aric Hagberg and Drew Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, 2020.

[60] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[61] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. volume 3, pages 523– 528, 09 2003.

[62] Mallory Wang Zhenyu Zhao, Radhika Anand. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. 2019.

[63] Sergio Nesmachnow and Santiago Iturriaga. Cluster-uy: Collaborative scientific high performance computing in uruguay. In Moisés Torres and Jaime Klapp, editors, *Supercomputing*, pages 188–202, Cham, 2019. Springer International Publishing.

[64] Andy B Yoo, Morris A Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*, pages 44–60. Springer, 2003.

[65] Thorsten Gressling. 11 python standard libraries and conda. In *Data Science in Chemistry*, pages 45–54. De Gruyter, 2020.