En este documento, se describen los pasos llevados a cabo en las diferentes secciones del Jupyter Notebook *Aguilar\_Alba\_TFM.ipynb*

A lo largo del documento, se encontrarán las siguientes notas:

* Pendiente – Acciones pendientes que me gustaría llevar a cabo
* Duda – Dudas que necesitaría aclarar

**1. Importación de datos**

En esta sección, procedemos a importar los diferentes ficheros que son necesarios para el presente estudio. Estos son:

|  |  |
| --- | --- |
| Fichero | Descripción |
| mindboggle\_ROIs.txt | Nombre de las diferentes regiones del cerebro y su correspondiente orden de aparición en las matrices de conectividad |
| clinic.csv | Datos clínicos de los sujetos que participan en el estudio |
| XXXX\_FA\_factor.csv | Matriz de conectividad estructural del sujeto XXX |
| XXXX\_GM\_matrix.csv | Matriz de conectividad morfológica del sujeto XXXX |
| XXXX\_r\_matrix.csv | Matriz de conectividad funcional del sujeto XXXX |

1. **Preprocesamiento**

Dadas *m* matrices de conectividad (estructural, morfológica o funcional) de tamaño *nxn*, en este estudio – dada la naturaleza simétrica de estas matrices – únicamente consideraremos los valores de la diagonal superior de cada una de ellas. Esto resulta en un total de n(n-1)/2 variables de longitud *m.*

Para cada variable *y,* con el fin de minimizar los factores de confusión, llevaremos a cabo los ajustes descritos en los siguientes apartados.

* 1. **Corrección por edad y sexo**

Realizaremos una corrección por edad y sexo usando una regresión lineal para tener en cuenta la variabilidad debido a estos factores. Aplica a todas las matrices (estructurales, morfológicas y funcionales)

Para ello, para cada variable *y*:

1. Construimos un modelo de regresión y – edad + género
2. Guardamos los residuos (representan la información no explicada por las variables independientes)
3. Sumamos a estos residuos la media de *y* obtenida únicamente considerando los sujetos control
4. Reemplazamos los valores actuales por los obtenidos en el paso anterior

* Pendiente. Mostrar visualmente, para alguna de las variables, la influencia de la edad y el género
* Duda. ¿Corresponde el valor 0 al sexo femenino y 1 al sexo masculino?
  1. **Valores nulos**
* Pendiente. Hasta ahora no he utilizado la variable zAttention, por lo que no he necesitado tratar ningún valor nulo todavía
  1. **Normalización de datos**

Normalizamos los datos para tenerlos en un rango de [0,1]*.* Aplica a todas las matrices (estructurales, morfológicas y funcionales)

Para ello, para cada una de las variables *y,*

* + 1. Obtenemos una lista con los valores actualesde la variable
    2. Normalizamos la lista con MinMaxScaler(feature\_range=(0, 1))
    3. Reemplazamos los valores actuales por los valores normalizados

* 1. **Harmonización de pesos**

Las matrices de conectividad estructural podrían sufrir variabilidad entre sitios debido a la heterogeneidad de ambos protocolos de adquisición (MVIS y FIS). Por ello, será necesario harmonizar los pesos. Solo aplica a matrices de conectividad estructural.

Para ello, para cada una de las variables *y*:

* + 1. Obtenemos un listado *vals* de los valores actuales de la variable
    2. Generamos un listado *batch* que distinga (con los valores 1 y 2) los sites que se han empleado para la obtención de cada uno de los valores de la lista anterior
    3. Calculamos los valores harmonizados de *vals* con la función *pycombat* utilizando la distinción *batch*
    4. Reemplazamos los valores actuales por los valores harmonizados.
* Pendiente. Mostrar visualmente, para alguna de las variables, la influencia del tipo de adquisición
  1. **Aplicación de umbrales y filtrado de conexiones**

Con el objetivo de minimizar la presencia de falsas conexiones, solo contemplaremos en el análisis aquellas conexiones que, tras todo el preprocesamiento anterior,

* tengan un peso mínimo de 0.1
* estén presentes en más del 60% de los sujetos sanos

Las conexiones que no cumplen estas condiciones se reemplazan por el valor 0 para ser posteriormente descartadas.

1. **Estudio de los datos clínicos y demográficos**

En este apartado llevamos a cabo un pequeño estudio de los datos demográficos de los sujetos para verificar si existen diferencias significativas entre los dos grupos (pacientes/controles).

* Pendiente. Ampliar el estudio para diferenciar también entre los diferentes fenotipos/duración de la enfermedad
* Duda. Los fenotipos se obtienen de la variable *mstype*, que puede tomar valores 0, 1 y 2. ¿Me podrías proporcionar los descriptivos correspondientes?

1. **Cálculo de métricas de grafos**
   1. **Transformación a grafos**

En esta sección, se transforman las diferentes matrices de conectividad en objetos de tipo grafo donde

* cada nodo se corresponde con una región cerebral
* se genera un enlace entre dos nodos si la correspondiente conectividad es diferente de 0
* el peso del enlace se corresponde con el valor de la conectividad
  1. **Métricas globales**

En este apartado, se calcula para cada uno de los grafos (tanto estructurales, morfológicos, como funcionales) las siguientes métricas:

|  |  |
| --- | --- |
| Métrica | Descripción |
| Strength | Suma de los grados (ponderados por peso) |
| Transitivity | Proporción de triángulos |
| Global efficiency | Media de la inversa de los caminos más cortos entre pares de nodos |
| Assortativity | Coeficiente de correlación entre los grados de todos los nodos en dos extremos opuestos de un enlace |
| Path length | Media de las longitudes de los caminos más cortos entre pares de nodos |
| Betwenness centrality | Media de la suma de ratios suma de la fracción de todos los pares de caminos más cortos que pasan a través de un nodo |

Una vez calculadas las métricas, se procede a comparar cada una de ellas entre los sujetos pacientes y los sujetos control, para ver si existen diferencias significativas entre ambos grupos. Para ello, empleamos el test U de Mann-Whitney con corrección de Bonferroni.

* Pendiente. Diferenciar entre hipótesis alternativa ‘less’ y ‘greater’
* Pendiente. Añadir comprobación de normalidad y, en base al resultado, utilizar el test U de Mann-Whitney o el test-t de Student
* Pendiente. Realizar este análisis de forma combinada (multicapa)
  1. **Métricas locales**
     1. **Nodos**

En este apartado, se calcula para cada uno de los nodos (tanto para las conectividades estructurales, morfológicas, como funcionales) sus grados medios ponderados por peso en cada uno de los grafos.

Posteriormente, se comparan los listados obtenidos para cada uno de los grupos (pacientes y controles) para ver si existen diferencias significativas, y se obtiene el número de nodos que muestran estas diferencias.

Finalmente, se estudia qué nodos pueden ser considerados hubs. Para ello, los nodos que muestran una mayor strength media (> mean + 1std) se consideran hubs.

* Duda. ¿Podrías confirmarme si los cálculos empleados para la obtención de strengths y de hubs en el código son correctos? Hago medias de medias y no me queda muy claro si debería ser así.
* Pendiente. Realizar este análisis de forma combinada (multicapa)
  + 1. **Enlaces**

En este apartado, se compara el lsitado de pesos FA de cada uno de los enlaces de entre grupos de sujetos. Esto nos devuelve un número de enlaces que presentan diferencias significativas.

Se aprovecha este proceso para etiquetas cada uno de estos enlaces como ‘red’/’blue’ en función de si los pacientes presentan una disminución/aumento significativo de los pesos con respecto a los sujetos sanos. Esto se utilizará más adelante para la representación y visualización del conectograma.

* Pendiente. Realizar este análisis para la capa morfológica, la capa funcional y la capa combinada (multicapa)
* Pendiente. Realizar estudio de correlación entre Edge-FA y zAttention

1. **Clasificación**

En este apartado, se lleva a cabo la creación de un modelo de predicción SVM que diferencie sujetos sanos de sujetos con EM basándose en las diferentes métricas que hemos calculado a lo largo del estudio.

* Pendiente. Realizar este análisis para la capa morfológica, la capa funcional y la capa combinada (multicapa)
* Pendiente. Realizar este análisis combinando todas las métricas en el mismo modelo

1. **Visualización**

En este apartado, se ha construido un conectograma que permite hacer una comparación del FA de los enlaces y la strength de los nodos. La disminución del FA de los enlaces se muestra en rojo; el aumento del FA de los enlaces se muestra en azul; los nodos con strength significativa se pintan de rojo.

* Pendiente. Conectograma para visualizar la correlación entre Edge FA y zattention