

## VERSIÓN PRELIMINAR 2

### Objeto del estudio:

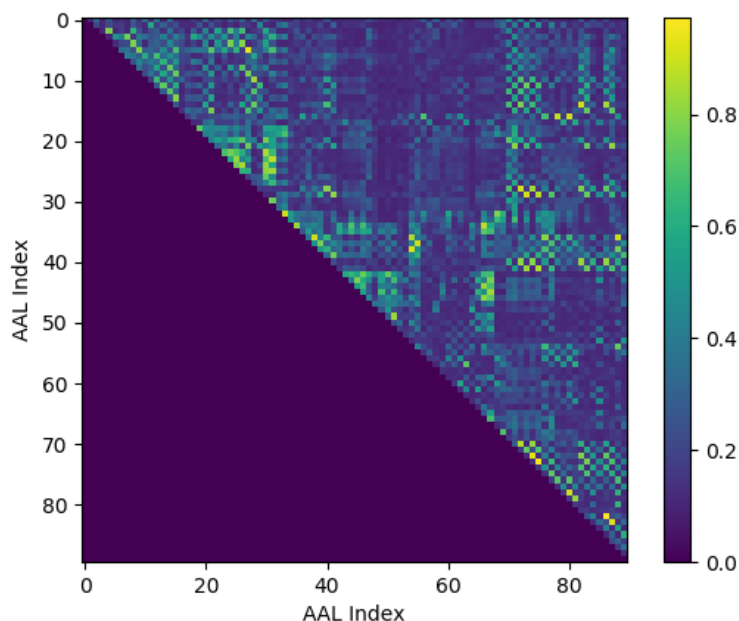
PLV de entre las 90 zonas del atlas AAL, suponiendo cada zona representada por su primera componente del PCA (de las ~25 que habría en cada zona) de pacientes con DCL, quejas subjetivas de memoria y el grupo de control.

### Objetivo del estudio:

Implementar un método de clasificación supervisado robusto.

### Descripción de los pacientes:

Cada paciente presenta 9 matrices de 90x90, cada una de ellas perteneciente a las bandas Delta, Theta, Alpha, Beta1, Beta2, Beta, Gamma, HighGamma1 y HighGamma2; con bordes en frecuencia correspondientes a [2, 4] Hz, [4, 8] Hz, [8, 12] Hz, [12, 20] Hz, [20, 30] Hz, [12, 30] Hz, [30, 45] Hz, [55, 75] Hz y [75,95] Hz, respectivamente.



*Ilustración 1: Matriz de PLV en banda Gamma para el paciente UMEC - 002*

### Construcción del dataset:

Los métodos de aprendizaje supervisado consisten, *grosso modo*, en “interpolan” las categorías de cada paciente (0 ó 1 si es clasificación binaria) en función de puntos en espacio n-dimensional. En nuestro caso, estos puntos serán la concatenación de los elementos de la diagonal superior de la Ilustración 1.

### Estrategia de clasificación:

Todos los métodos que vamos a utilizar (tanto pseudoinversa como regresión logística) son métodos lineales, consistentes en encontrar un hiperplano que separe los dos grupos de forma que se minimice el error de training.

Una vez encontrado ese plano, se puede desplazar en la dirección de su vector normal hasta ajustar el ratio de verdaderos positivos y falsas alarmas donde interese.

Se hará primero una separación lineal sobre los datos directamente para evaluar un valor base de separación y posteriormente se realizará una transformación no lineal de los datos, previo mapeo a distinta dimensión, para mejorar la clasificación basándonos en el comportamiento descrito en [1].

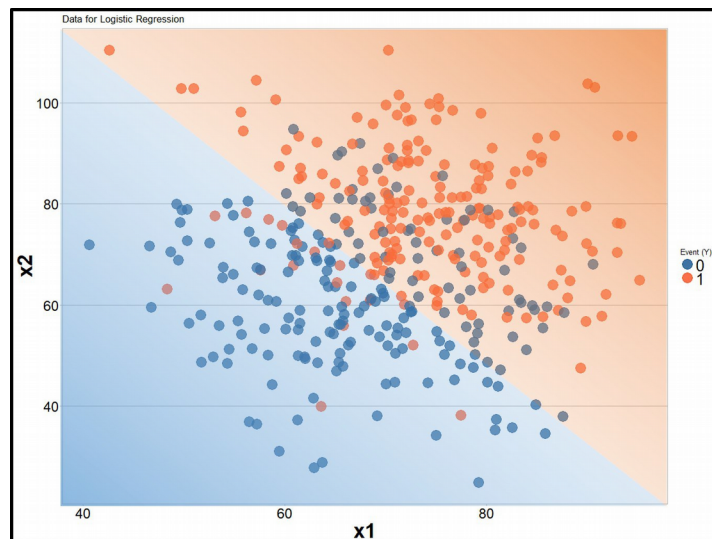
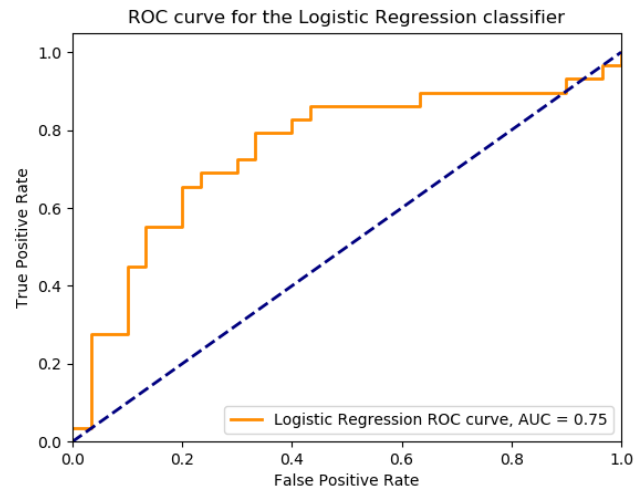


Ilustración 2: Ejemplo de separación lineal de puntos

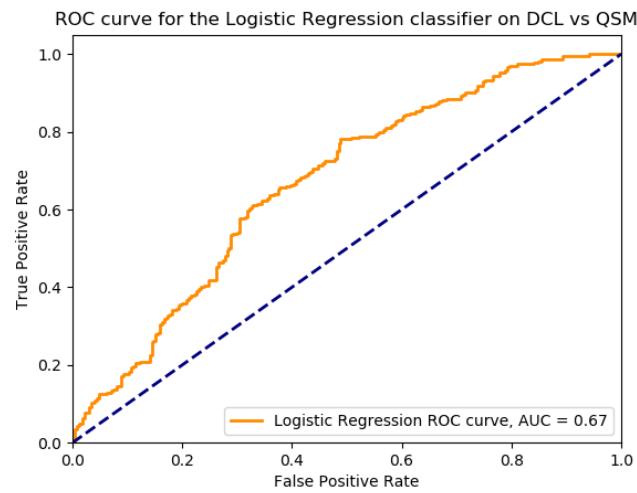
### Primera clasificación:

Esta se realiza mediante regresión logística de los datos. El entrenamiento se puede realizar con todas las bandas o sólo con una. Fuimos probando para cada caso y nos quedamos con la que nos daba la mejor separación de pacientes. Para el caso DCL vs Control, limitar el entrenamiento a la banda Gamma mejoró mucho la clasificación. Sin embargo, la elección final de hacer el entrenamiento en todas las bandas para el caso DCL vs QSM o elegir HighGamma2 para el caso QSM vs Control no fue tan clara pues el área bajo la curva no variaba tanto de prueba a prueba.

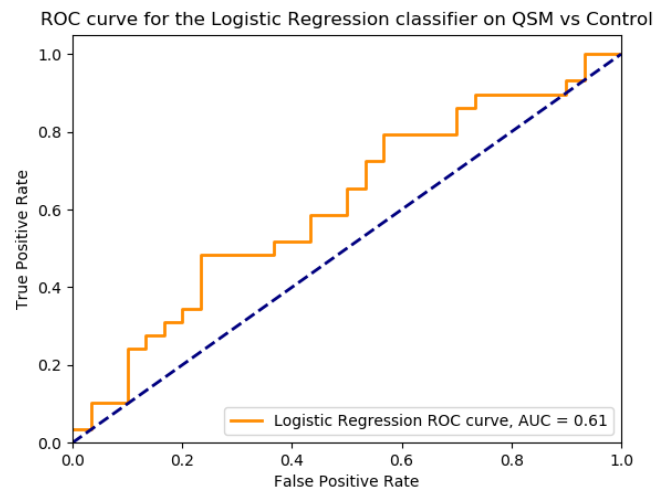
ÁREA BAJO ROC EN FUNCIÓN DE BANDAS ( DCL vs Control )			
BANDA	AUC	BANDA	AUC
Todas	0.64	Beta2	0.64
Delta	0.70	Beta	0.57
Theta	0.66	Gamma	0.75
Alpha	0.60	HighGamma1	0.70
Beta1	0.58	HighGamma2	0.72



*Ilustración 3: Curva ROC caso DCL vs Control*



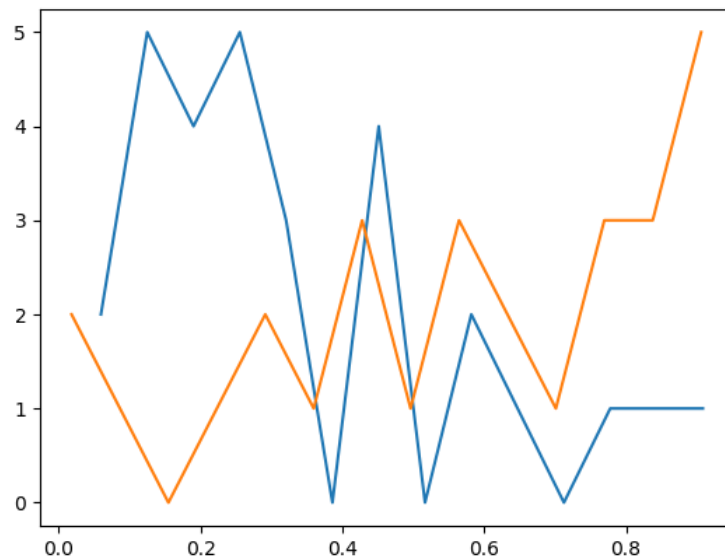
*Ilustración 4: Curva ROC caso DCL vs QSM*



*Ilustración 5: Curva ROC caso QSM vs Control*

En el caso DCL vs Control, el área bajo la curva de la Ilustración 3 se traduce en una distribución de

las distancias con respecto al hiperplano de separación dado por el siguiente histograma



*Ilustración 6: Histograma de distancias al plano separador para cada clase. DCL en azul y Control en naranja.*

donde el histograma naranja pertenece al control (etiqueta 1) y el histograma azul perteneciente a los pacientes con DCL (etiqueta 0). Un threshold puesto en 0.5 muestra una exactitud (*accuracy*) del clasificador del 73%.

Realizamos la clasificación con la selección de features que usted nos indicó, esto es, el menor de la matriz original formado por las filas y las columnas 25-26-31-32-33-34-37-38-61-62-67-68-85-86, y, lamentablemente, no encontramos una mejor división de los grupos.

### **Segunda clasificación:**

Una vez obtenida la banda que mejor separa en el clasificador base, se calcula un clasificador entrenado con todos los PCA de las bandas mapeadas a otra dimensionalidad y transformadas no linealmente.

La transformación realizada es la siguiente:

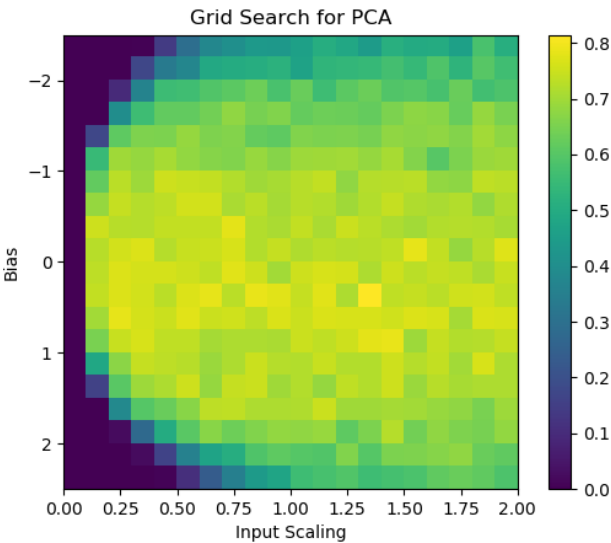
$$\tilde{x} = a \cdot W_{input} \cdot x + b$$

$$X = \tanh(\tilde{x})$$

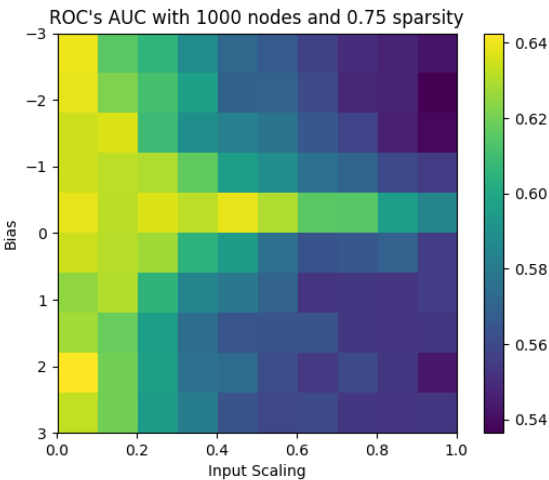
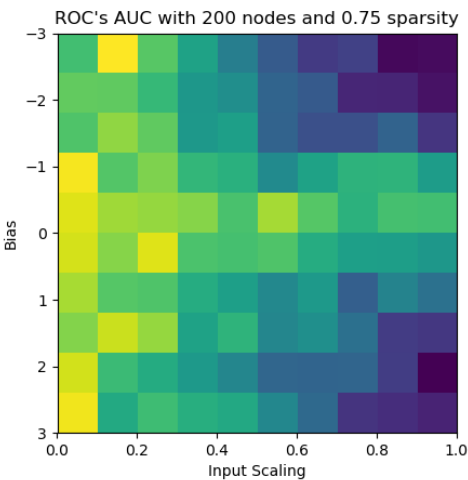
donde  $a$  es el input scaling,  $b$  es el bias,  $x$  son los coeficientes de cada componente principal de la banda de un paciente,  $W_{input}$  son los pesos de entrada aleatorios y  $X$  es la activación de los nodos.

La optimización de los parámetros resulta en las áreas bajo la curva ROC que se muestran a continuación. Una vez encontrados los valores óptimos, incrementamos el número de nodos hasta encontrar el mejor resultado.

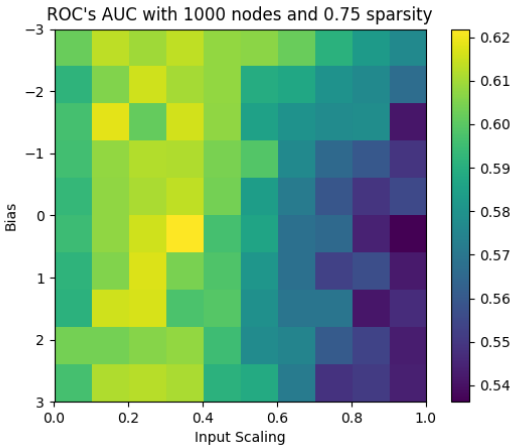
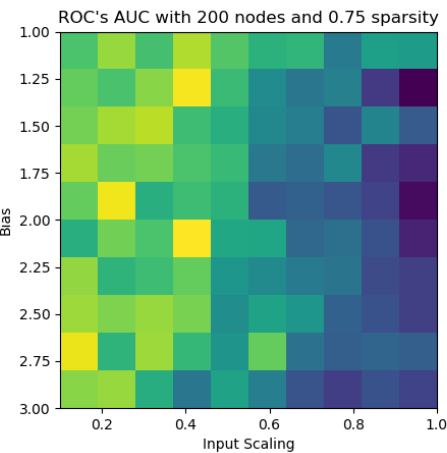
DCL vs Control



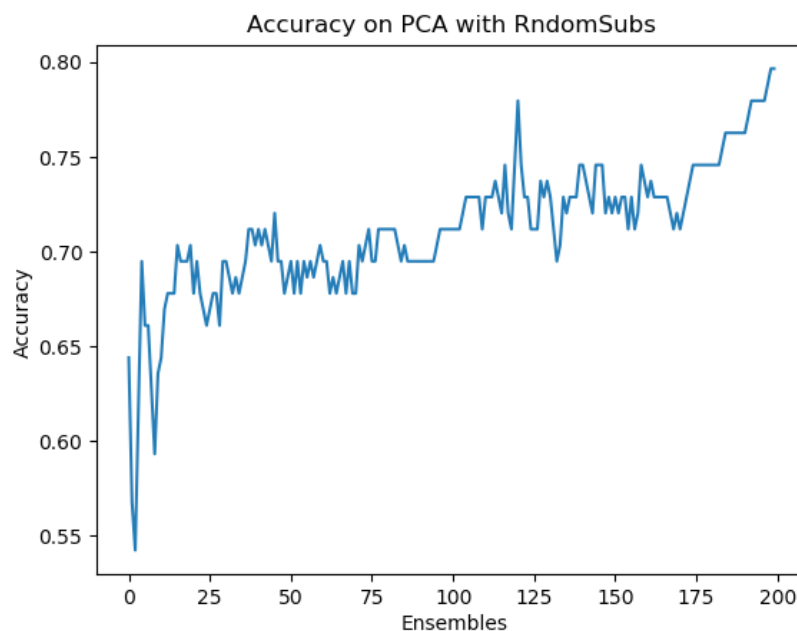
DCL vs QSM



QSM vs Control



Una vez hecha la optimización, hacemos una clasificación basada en mezcla de realizaciones (*ensembles*), cuyo resultado, en el caso de DCL vs Control, da una exactitud en función del número de ensembles que varía de la siguiente forma



cada *ensemble* consiste en la clasificación de los pacientes habiendo hecho el entramiento con una selección aleatoria de *features* de la mitad del número de *features* total.

La matriz de confusión para un threshold colocado en 0.5 (en el sentido expuesto en la Ilustración 6) para el algoritmo de separación es la siguiente

Clasificación		Valor Predicho	
		DCL	Control
Valor Real	DCL	23	7
	Control	5	24

Como nota final, implementamos clasificaciones basadas en varios algoritmos clasificadores entre los que se encuentran Naive-Bayes, AdaBoost, Nearest Neighbours y Random Forest, para valorar el desempeño de estos en comparación con la ELM usada en este informe. Todos ellos dieron peores resultados a excepción de los Random Forests, que consiguieron la misma calidad de clasificación. Es posible que esto se deba a que los *ensembles* usados en nuestra ELM se han construido con subespacios aleatorios del espacio de *features*, exactamente igual que cada árbol de decisión se construye dentro del random forest.

### Conclusiones:

Las pruebas realizadas dan a entender que hay diferencias claras entre los sujetos Control y los sujetos con deterioro cognitivo leve (sobre todo en la banda Gamma). Por otra parte, los sujetos con quejas subjetivas de memoria están en un estado intermedio ya que no es posible separarlos de los sujetos Control ni de los sujetos con deterioro cognitivo leve.

---

## **Bibliografía**

1: Appeltant, Lennert; Reservoir Computing Based On Delay-Dynamical Systems; Introduction, 2012