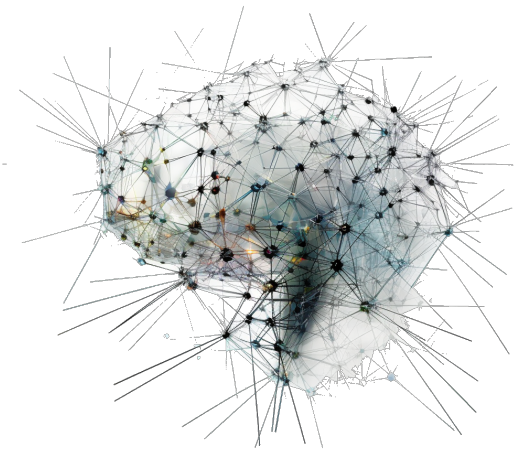
Análisis estadístico y clasificación automática de datos de Tomografía de Emisión por Positrones en Enfermedad de Alzheimer

Juan A. Arias

Manuel Oviedo de la Fuente

Rubén Fernández Casal



19 de Junio, 2023

La investigación conducente a este Trabajo de Fin de Máster se enmarca dentro del proyecto de tesis doctoral "*Desarrollo de métodos estadísticos para el análisis de datos de neuroimagen de cara al diagnóstico precoz de enfermedades neurodegenerativas*" del PD en Neurociencia y Psicología Clínica de la USC.

Índice de Contenidos

La Enfermedad de Alzheimer

Contexto y motivación

- **Desafío creciente**: Alzheimer y otras enfermedades neurodegenerativas en la sociedad contemporánea.
- Impacto en la calidad de vida de pacientes, familias, sistema de salud.
- Diagnóstico temprano: crucial, pero limitado por técnicas de análisis de datos disponibles.
- **Enfoque innovador**: combinar técnicas de aprendizaje estadístico (SL) avanzadas (aprendizaje profundo) con resultados de análisis de datos funcionales (FDA) (Arias-Lopez, Cadarso-Suarez, and Aguiar-Fernandez, 2021; Arias-Lopez, Cadarso-Suarez, and Aguiar-Fernandez, 2022).

La Enfermedad de Alzheimer

Sintomatología, etiología, diagnóstico, y tratamiento

- **Síntomas**: pérdida progresiva de memoria, desorientación, deterioro cognitivo, cambios de personalidad, dependencia creciente.
- **Etiología**: múltiples teorías no excluyentes, sin causa específica identificada.
- **Tratamiento**: paliativo, no curativo.
- **Diagnóstico**:
  - → Basado en pruebas neuropsicológicas y técnicas de neuroimagen, a menudo tardío.
  - → **Diagnóstico temprano**: esencial para la mitigación de una creciente crisis de salud pública.

Técnicas de aprendizaje estadístico

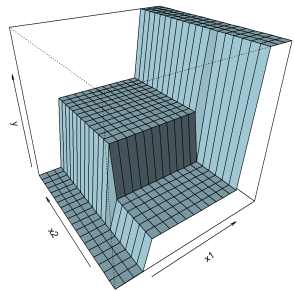
Introducción y avances

- **Técnicas de SL** son métodos que usan teoría de la probabilidad y teoría de funciones para inferir y predecir en base a datos previos.
- Desarrollado inicialmente por científicos como **Alan Turing** y **Frank Rosenblatt**. Ahora incluye técnicas avanzadas como SVM y aprendizaje profundo.
- SL tiene aplicaciones en gran variedad de campos y sigue expandiéndose.
- **En este estudio**, se utilizarán varias técnicas de SL para comparar la eficiencia del nuevo modelo propuesto con otros en la literatura.

Árboles de decisión

Resumen

- Los árboles de decisión predicen el valor de una variable objetivo basándose en reglas de decisión que segmentan el espacio predictor dividiéndolo en regiones.
- El proceso de creación de un árbol de decisión consta de dos fases: *crecimiento* y *poda*.
- Un equilibrio clave: complejidad del árbol *vs* pureza de los nodos.
- Para controlar el tamaño del árbol, se seleccionan hiperparámetros mediante validación cruzada.



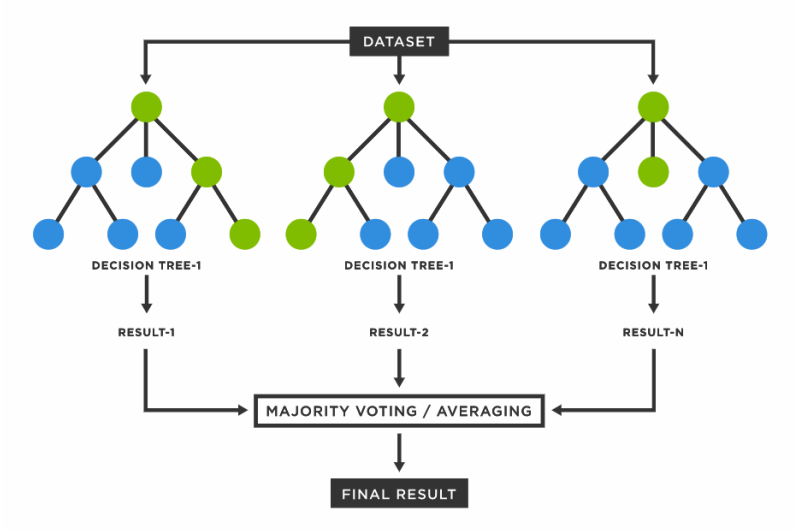
Bosques Aleatorios

Resumen

- Método de SL que combina árboles de decisión usando *bagging* para crear un modelo predictivo más robusto.
- El *bagging* disminuye la varianza al fusionar varios modelos.
- Cada modelo genera una predicción. En regresión, se toma la media de las predicciones; en clasificación, se utiliza el voto mayoritario.
- Aumentar el número de árboles no siempre mejora las predicciones.
- Aunque se pierde interpretabilidad con *bagging*, existen métodos para calcular la importancia de los predictores.

Bosques Aleatorios

Funcionamiento básico de un bosque aleatorio.



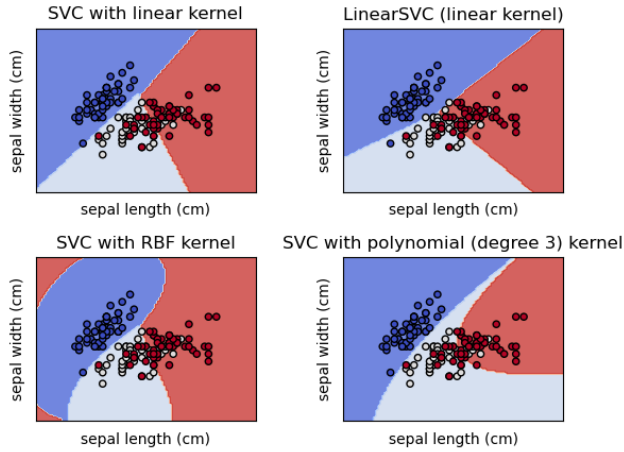
Máquinas de Vectores de Soporte (SVM)

Resumen

- Los clasificadores de soporte vectorial determinan un hiperplano óptimo de separación entre clases de datos.
- Maximizan el margen entre las instancias más cercanas al hiperplano, llamadas vectores de soporte.
- El entrenamiento se puede representar como un problema de optimización, donde se busca maximizar el margen y minimizar la suma de las variables de holgura, con un límite tolerable de error.
- Para datos no linealmente separables, se aplican funciones kernel pasando a ser Máquinas de Soporte Vectorial.
- Ejemplos de kernels: lineal, polinómico, radial, tangente hiperbólica.

Máquinas de Vectores de Soporte

Ejemplo de uso sobre datos Iris.



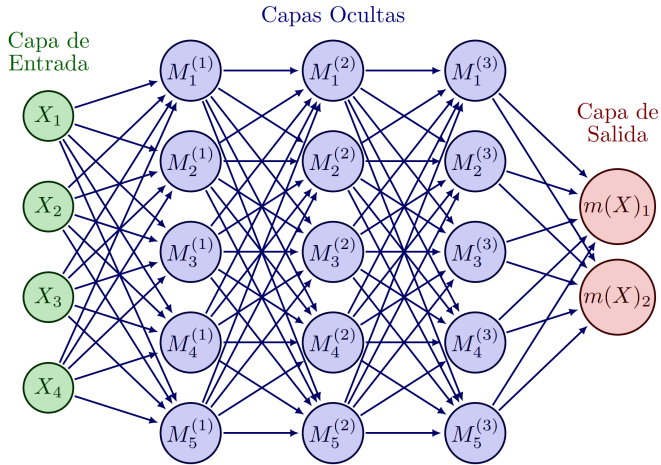
Aprendizaje Profundo y Redes Neuronales Artificiales

Resumen

- El aprendizaje profundo, una rama del SL, utiliza redes neuronales artificiales (ANN) con múltiples capas ocultas para aprender representaciones jerárquicas de los datos.
- ANNs pueden manejar un gran número de parámetros, lo que las hace adecuadas para problemas con estructuras subyacentes muy complejas.
- Las ANNs actúan como aproximadores universales de funciones. Buscan aprender una función $f(x;\theta)$ que minimice la discrepancia entre la función objetivo $f^*(x)$ y $f(x;\theta)$.
- Un modelo de aprendizaje profundo implica una ANN con múltiples capas: una de entrada, un gran número de capas ocultas y una capa de salida.

Aprendizaje Profundo y Redes Neuronales Artificiales

Estructura de una red neuronal con tres capas ocultas.



Análisis de Datos Funcionales (FDA)

Generalidades

- El FDA es un campo emergente de la estadística que analiza datos representados como funciones.
- Útil cuando los datos son contínuos y existen dependencias entre observaciones cercanas.
- Opera en espacios de Hilbert, lo que permite la manipulación de los datos funcionales.
- Una representación común de los datos en FDA es la expansión en series de bases. Siendo el número de funciones base a considerar es crucial para la calidad de la representación.
- Aplicaciones de FDA abarcan campos como la biología, la medicina, la economía...

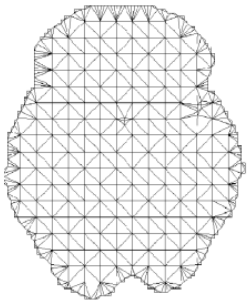
Análisis de Datos Funcionales (FDA)

Aplicación a Imagen Médica

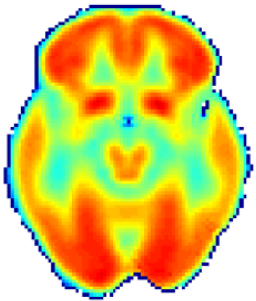
- Las técnicas de FDA se pueden aplicar al campo de la imagen. Desarrollos recientes permiten obtener el valor medio de grupos de imágenes y sus intervalos de confianza simultáneos (SCCs).
- Este procedimiento se puede extender al caso de dos muestras para comparar las funciones medias de dos poblaciones de datos de imágenes.
- El objetivo es hacer inferencias **sobre las diferencias entre las funciones medias** de las dos poblaciones.
- Útil para recalcar áreas con cambios en actividad significativos al comparar grupos (Arias-Lopez, Cadarso-Suarez, and Aguiar-Fernandez, 2021).
- La eficiencia de la técnica hace que sea una opción atractiva para complementar el algoritmo de aprendizaje profundo.

Análisis de Datos Funcionales (FDA)

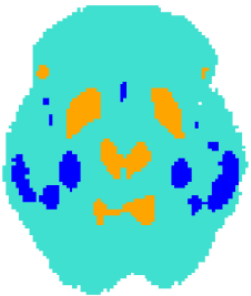
Visualización de Resultados de la metodología SCC



(a) Triangulaciones de Delaunay



(b) Imagen media del grupo AD



(c) Regiones fuera de los SCC

Problema y Objetivos del Estudio

Importancia:

- La importancia de este estudio radica en su potencial para explorar nuevas vías de mejora en la detección temprana de AD.
- Este trabajo también contribuye al crecimiento y expansión de las técnicas de SL y FDA, explorando nuevas aplicaciones de estas disciplinas al ámbito de la neurociencia.

Objetivos:

1. Diseñar un algoritmo de aprendizaje profundo que incorpore resultados de FDA para mejorar la precisión en el diagnóstico de AD usando escáneres PET.
2. Comparar el rendimiento del algoritmo propuesto con enfoques más convencionales: árboles de decisión, bosques aleatorios, SVM, y aprendizaje profundo sin FDA.

Índice de Contenidos

Descripción de los Datos

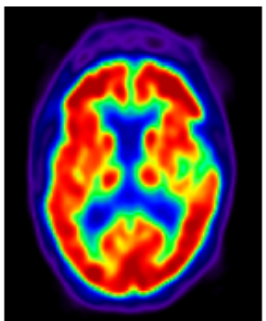
- 126 pacientes (51 AD, 75 CN) con datos demográficos (edad y sexo) y de neuroimagen.
- Edad media de 75.18 años con una desviación estándar de 5.99 años.
- 52 mujeres y 74 hombres distribuidos entre los grupos de AD y CN.
- Imágenes PET en formato Analyze, con dimensiones de $79 \times 95 \times 79$.
- Imágenes bidimensionales extraídas en el corte $Z = 30$.
- Total de 7508 variables predictoras por paciente: 7505 píxeles de las imágenes y 3 variables demográficas.

Preprocesamiento de los Datos de Neuroimagen

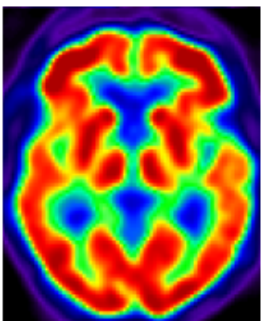
- Utilizamos el software SPM para el preprocesamiento de las imágenes PET:
  1. *Realineamiento*: Corrige las diferencias de movimiento entre las imágenes adquiridas.
  2. *Normalización espacial*: Registra las imágenes a un espacio estandarizado de referencia, el *cerebro estereotáctico*.
  3. *Normalización de intensidad*: Corrige las diferencias en la intensidad de las imágenes debido a variaciones en la adquisición o características anatómicas individuales.
  4. *Enmascaramiento*: Selecciona regiones de interés, excluyendo áreas irrelevantes.
- El preprocesamiento garantiza la comparabilidad pixel-a-pixel entre pacientes a costa de una cierta pérdida de información.

Preprocesamiento de Datos de Neuroimagen

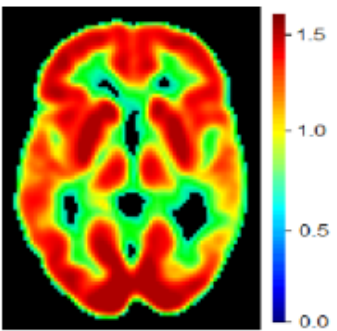
Visualización por etapas:



(a) Imagen original



(b) Imagen procesada



(c) Imagen enmascarada

Índice de Contenidos

Métricas de Evaluación

- Métricas derivadas de la matriz de confusión para evaluar nuestros modelos:
  - → *Sensibilidad* (tasa de verdaderos positivos)
  - → *Especificidad* (tasa de verdaderos negativos)
  - → *Precisión* (proporción de identificaciones positivas correctas)
  - → *Precisión balanceada* (media aritmética de la sensibilidad y especificidad)
  - → *Coeficiente Kappa de Cohen* (medida de acuerdo entre dos raters)
- Visión completa del rendimiento del modelo.
- Especial atención a la precisión balanceada y al coeficiente de Kappa debido al ligero desbalance en las clases.
- En todos los modelos se realizó una búsqueda de hiperparámetros en rejilla.

Consideraciones Adicionales

- Para cada modelo, una vez ajustado con hiperparámetros óptimos, **se realizó una única predicción de la categoría de los datos de prueba**.
- **No utilizamos la técnica de remuestreo *bootstrap*** debido al reducido tamaño de la muestra de prueba, ya que puede producir sobreestimación del rendimiento del modelo y sesgos en los intervalos de confianza.

Árboles de Decisión

- **Hiperparámetros óptimos seleccionados:**
  - → Parámetro de complejidad (cp): 0.089
  - → Profundidad máxima del árbol (Max.Depth): 1

Predicción/Observado	Positivo	Negativo
Positivo	1	1
Negativo	4	6

Sensibilidad	Especificidad	Precisión	Precisión Balanceada	Kappa Coef.
0.2	0.8571	0.5833	0.5285	0.0625

Bosques Aleatorios

- **Hiperparámetros óptimos seleccionados:**
  - → Número de árboles (N.Tree): 100
  - → Número de variables disponibles para la división en cada nodo (M.Try): 1920

Predicción/Observado	Positivo	Negativo
Positivo	3	2
Negativo	2	5

Sensibilidad	Especificidad	Precisión	Precisión Balanceada	Kappa Coef.
0.6	0.7142	0.6666	0.6571	0.3142

La eficiencia ha sido superior a la del árbol de decisión, aún así no son capaces de capturar la correlación espacial ni de modelar relaciones no lineales propias de estos datos.

Máquinas de Vectores de Soporte (SVM)

- **Hiperparámetros óptimos seleccionados:**
  - $\rightarrow$ $coste = 2^{-1}$
  - $\rightarrow$ $gamma = 0.5$

Predicción/Observado	Positivo	Negativo
Positivo	0	0
Negativo	5	7

Sensibilidad	Especificidad	Precisión	Precisión Balanceada	Kappa Coef.
0	1	0.5833	0.5	0

A pesar de las fortalezas de las SVM en espacios de alta dimensión, la eficacia de este modelo fue pobre en nuestros datos.

Aprendizaje Profundo

Hiperparámetros Óptimos

Utilizamos la biblioteca `keras` en Python y `caret` en R para implementar el aprendizaje profundo. Se seleccionan hiperparámetros óptimos mediante búsqueda en cuadrícula.

Hiperparámetro	**Valores óptimos**
Número de capas ocultas	5
Número de nodos en cada capa	256
Función de activación	'relu'
Tasa de aprendizaje	0.001
Número de épocas de entrenamiento	10

Aprendizaje Profundo

Matriz de Confusión y Métricas de Precisión

Predicción/Observado	Positivo	Negativo
Positivo	4	1
Negativo	2	5

Sensibilidad	Especificidad	Precisión	Precisión Balanceada	Kappa Coef.
0.6667	0.8333	0.75	0.75	0.5

Hasta el momento esta es la técnica con los mejores resultados.

Aprendizaje Profundo con Ingeniería de Características

Diseño de Características

- **Objetivo:** Mejorar el rendimiento de los modelos de aprendizaje profundo al incorporar información de regiones identificadas como más relevantes según técnicas de FDA.
- **Método:** Crear nuevas características basadas en estas regiones para proporcionar información adicional al algoritmo (i.e., conocimiento de campo).
- Las características generadas son la media, mediana, máximo, mínimo, desviación estándar, varianza, asimetría, kurtosis y diferencia de vecindario de los píxeles de mayor relevancia.

Aprendizaje Profundo con Ingeniería de Características

Matriz de Confusión y Métricas de Precisión

Predicción/Observado	Positivo	Negativo
Positivo	4	1
Negativo	1	6

Sensibilidad	Especificidad	Precisión	Precisión Balanceada	Coef. Kappa
0.8	0.86	0.8	0.83	0.6571

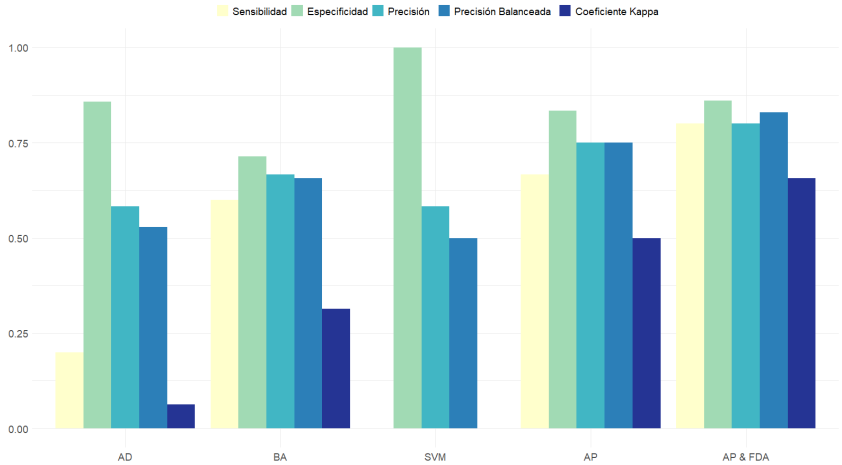
Resultados Generales

Tabla de Métricas de Eficiencia

Modelo	Sensibilidad	Especificidad	Precisión	P. Balanceada	Kappa Coef.
Árboles	0.2	0.8571	0.5833	0.5285	0.0625
Bosques	0.6	0.7142	0.6666	0.6571	0.3142
SVM	0	1	0.5833	0.5	0
AP	0.6667	0.8333	0.75	0.75	0.5
AP & FDA	0.8	0.86	0.8	0.83	0.6571

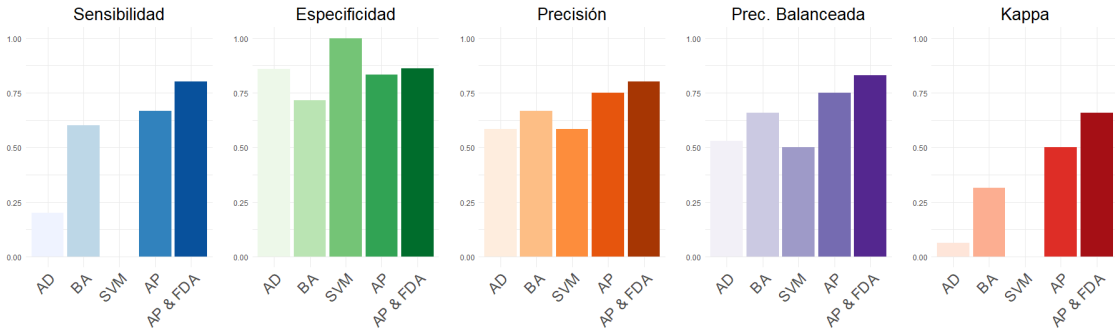
Visualización de Resultados

Agrupados según la metodología de SL analizada



Visualización de Resultados

Agrupados según la métrica analizada



Índice de Contenidos

Interpretación de Resultados

Parte I

- **Árboles de Decisión:** Producen resultados insatisfactorios debido a la decisión de diagnóstico basada en un número reducido de píxeles individuales.
- **Bosques Aleatorios:** Mejor rendimiento que los árboles de decisión, debido a su naturaleza de combinación de múltiples árboles y consideración de múltiples píxeles de diferentes regiones.
- **Máquinas de Vectores de Soporte (SVM):** Resultados pobres posiblemente debido a la alta correlación espacial en los datos de neuroimagen PET que tratan cada característica de forma independiente.

Interpretación de Resultados

Parte II

- **Aprendizaje Profundo:** Eficiente, proporciona mejores resultados que los métodos anteriores debido a su capacidad para modelar relaciones altamente no lineales, que son comunes en los datos de neuroimagen PET.
- **Aprendizaje Profundo con Datos Funcionales:** Produce los mejores resultados en este estudio. Usa características adicionales para el diagnóstico de AD, lo que proporciona información adicional que puede mejorar la eficacia del modelo.
- **Conclusión:** SL que puedan manejar la complejidad y tengan en cuenta la relación espacial entre los píxeles proporcionan mejores resultados. FDA y SL obtienen los mejores resultados.

Limitaciones del Estudio

- **Tamaño de Muestra:** Grande para un estudio de AD, pero podría no ser adecuado para técnicas de SL, en particular, aprendizaje profundo.
- **Diversidad de Características:** Aunque se incorporaron varias características relevantes, podrían existir otras que mejorarían la eficacia del modelo.
- **Interpretabilidad de los Modelos:** Los modelos de aprendizaje profundo incurren en el efecto caja negra, no proporcionando información sobre el porqué de sus decisiones.
- **Generalización de Modelos:** A pesar del uso de la validación cruzada, los modelos pueden no generalizarse bien a nuevos datos.

Conclusiones

- **Contribuciones:** Se ha demostrado que la combinación de técnicas de SL y FDA puede mejorar la precisión del diagnóstico de AD, específicamente combinar la complejidad del aprendizaje profundo y las métricas basadas en FDA.
- **Implicaciones Sociales:** El envejecimiento de la población y la falta de una cura para la AD hacen de su diagnóstico una prioridad. Un diagnóstico temprano permite la aplicación de tratamientos que retrasen la progresión de la enfermedad y su prevalencia.
- **Resumen:** Este estudio ha mostrado que el uso de técnicas de SL, en combinación con técnicas de FDA, puede mejorar significativamente la precisión del diagnóstico de AD, con importantes implicaciones para la práctica médica y la sociedad en general.

Futura Investigación

- **Aumentar el tamaño de la muestra:** mejorando la capacidad de los modelos de aprendizaje profundo para manejar relaciones complejas y no lineales.
- **Mejorar la interpretabilidad de los modelos de aprendizaje profundo:** tratar de evitar la caja negra mediante técnicas como el Brain-Inspired Modular Training.
- **Explorar el uso de bosques aleatorios con XGBoost:** A pesar de su eficiencia limitada en este estudio, estas técnicas pueden ser útiles para identificar las regiones de mayor relevancia para el diagnóstico de AD.
- **Aplicación tridimensional:** Extender las metodologías FDA a 3D podría permitir al algoritmo acceder a datos de correlación espacial en todas las direcciones.

Bibliografía Seleccionada I

Arias-Lopez, JA, C Cadarso-Suarez, and P Aguiar-Fernandez (2021). “Simultaneous Confidence Corridors for neuroimaging data analysis: applications to Alzheimer’s Disease diagnosis”. *Arxiv*.

— (2022). “Functional Data Analysis for Imaging Mean Function Estimation: Computing Times and Parameter Selection”. *Computers* 11, p. 91.

Breijyeh, Z and R Karaman (2020). “Comprehensive Review on Alzheimer’s Disease: Causes and Treatment”. *Molecules* 25.24, p. 5789.

Breiman, L et al. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chernick, Michael R (2011). *Bootstrap methods: A guide for practitioners and researchers*. John Wiley & Sons.

Friston, K (2007). *Statistical Parametric Mapping*. Elsevier.

Bibliografía Seleccionada II

Hastie, T, R Tibshirani, and J Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Ramsay, J and BW Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Wang, Y. et al. (2020). "Simultaneous confidence corridors for mean functions in functional data analysis of imaging data". *Biometrics* 76, pp. 427–437.

Gracias por su atención.