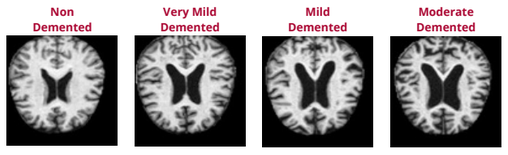
CLASIFICACIÓN SUPERVISADA DE IMÁGENES DE RESONANCIA MAGNÉTICA PARA LA DETECCIÓN DE ALZHEIMER

Julia Martins Guardia, Javier Álvarez Liébana
Máster de Bioestadística de la Fac. de Estudios Estadísticos de la Universidad Complutense de Madrid

Objetivos

Las **imágenes de resonancia magnética (MRI)** son una herramienta fundamental para observar la reducción del hipocampo, una característica clave de la **Enfermedad de Alzheimer (EA)**. Estas imágenes permiten analizar las distintas densidades cerebrales, facilitando la identificación de la pérdida de volumen cerebral asociada a la muerte neuronal.
Este trabajo da continuidad a un estudio previo centrado en la clasificación de imágenes de este tipo con el objetivo de desarrollar técnicas de clasificación binaria en el diagnóstico de la enfermedad (presencia o ausencia de la enfermedad), desarrollado en el Trabajo de Fin de Máster (TFM) del año pasado por Elisa Caballero Testón, bajo la tutoría de los profesores Javier Álvarez Liébana y Aida Calviño Martínez.
En esta fase, el objetivo será extenderlo hacia una **clasificación multiclase** de las diferentes etapas del Alzheimer, abordando tres desafíos principales: a) el desbalanceo de las categorías en el caos multiclase; b) la correcta validación y evaluación de los modelos; c) técnicas de remuestreo que nos ayuden a solventar los dos anteriores problemas.

Datos

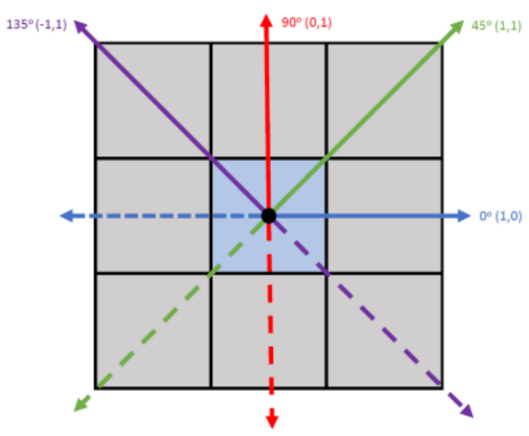


- Non Demented: 3200 imágenes
- Very Mild Demented: 2240 imágenes
- Mild Demented: 896 imágenes
- Moderate Demented: 64 imágenes

Los datos utilizados provienen principalmente de la base ADNI (Alzheimer's Disease Neuroimaging Initiative) y están disponibles en Kaggle. Este conjunto de datos incluye 6400 imágenes de resonancia magnética (MRI) en 2D de diferentes secciones del cerebro, representando varias etapas de la Enfermedad de Alzheimer (EA).
Para extraer información relevante de estas imágenes, en el trabajo previo se empleó la matriz GLCM, una técnica utilizada para analizar la textura de las imágenes en función de las intensidades de grises. En este proceso, se contabiliza cuántas veces aparecen juntos dos píxeles adyacentes con diferentes intensidades en una determinada dirección (shift), generando una matriz simétrica y cuadrada. Posteriormente, esta matriz se normaliza para representar la frecuencia relativa de cada combinación de intensidades adyacentes.
De esta manera, a partir de las matrices GLCM construidas, se extrajeron nueve métricas utilizadas para la caracterización de las imágenes MRI. Estas características constituyen la base sobre la que se desarrolla el presente estudio.

Metodologia

Balanceo



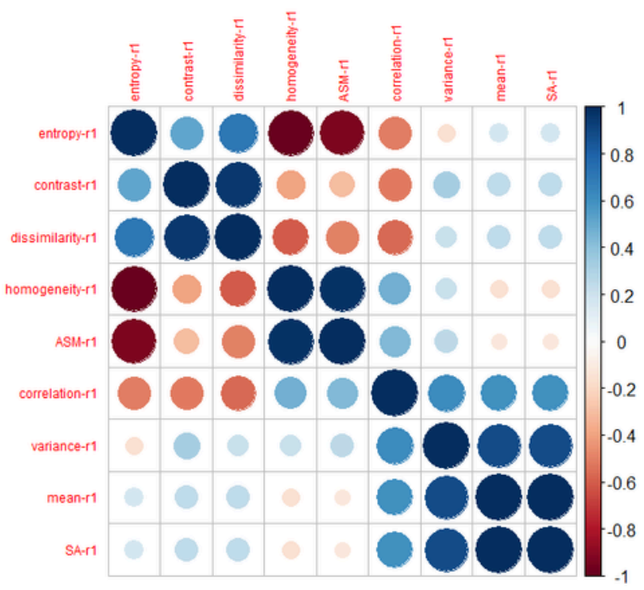
Es importante considerar que, además de las direcciones tradicionales de shift, se introdujo la dirección radial, que contabiliza los píxeles vecinos en todas las direcciones dentro de un radio unitario alrededor del píxel central. Este enfoque permite un análisis más completo de las texturas, integrando la información de múltiples direcciones en una única representación.
Por esta razón, el balanceo de clases se llevó a cabo utilizando las características extraídas de esta dirección. Para ello, se evaluaron diferentes estrategias con el objetivo de garantizar una distribución equitativa de las muestras. Inicialmente se consideró el uso exclusivo del método **Bootstrap**. Sin embargo, esta opción fue descartada debido al alto desbalance en las muestras. El principal inconveniente de utilizar solo Bootstrap en este contexto es que, al generar nuevas muestras con reposición, podría ocurrir una repetición excesiva de los mismos datos de la clase minoritaria. Esto podría provocar un sobreajuste del modelo a esa clase, ignorando o dando menos peso a las clases mayoritarias, lo que resultaría en una representación inadecuada de la distribución real de los datos y generaría sesgo, afectando negativamente la capacidad de generalización del modelo.
Por esta razón, se optó por no proceder exclusivamente con Bootstrap. En su lugar, se utilizó el método **SMOTE**, que genera muestras sintéticas para la clase minoritaria, equilibrando de manera más efectiva la distribución de las clases. No obstante, dado que este método por sí solo no fue suficiente para equilibrar completamente las clases, se complementó con **Bootstrap con reposición**, utilizando como tamaño máximo de la muestra el de la clase mayoritaria. Esto permitió asegurar que el modelo tuviera datos representativos de todas las clases sin inducir sesgo por la sobreabundancia de la clase minoritaria.

Seleción de variables

La **multicolinealidad** se refiere a la alta correlación entre las variables independientes de un modelo. Cuando hay multicolinealidad, esto puede resultar en estimaciones imprecisas e inestables, afectando la confiabilidad del modelo. Para abordar este problema, se realizó una selección previa de variables basada en tres métodos:

- **Matriz de correlaciones:** analiza la relación lineal entre las variables, identificando aquellas que están altamente correlacionadas a través del coeficiente de Pearson.
- **VIF (Factor de Inflación de la Varianza):** evalúa la multicolinealidad entre las variables, indicando si una variable puede ser explicada por otras. Un valor alto de VIF sugiere una alta correlación entre las variables.
- **PCA (Análisis de Componentes Principales):** reduce la dimensionalidad de los datos, creando nuevas variables a partir de combinaciones lineales de las variables originales, que conservan la mayor parte de la información y facilitan tanto la interpretación como el rendimiento del modelo.

El análisis de correlación se realizó considerando todas las variables del conjunto de datos en diversas direcciones. Se observó que las correlaciones eran similares en todas las direcciones, por lo que se eligió la **dirección radial**, ya que considera los píxeles vecinos en todas las direcciones dentro de un radio unitario, proporcionando un análisis más completo de las texturas en las imágenes.
A continuación, se muestra el gráfico de correlaciones de las variables en la dirección seleccionada.



Clasificación supervisada

Para la clasificación supervisada, primero se realizó una optimización de los parámetros utilizando **validación cruzada k-fold** (con tres repeticiones y cinco particiones) para cada modelo, excepto la regresión logística. Posteriormente, se entrenaron los modelos utilizando la **parametrización óptima**. Los modelos utilizados fueron:

- **KNN (K-Nearest Neighbors):** clasifica los datos en función de la proximidad a los puntos más cercanos. Sin embargo, a medida que aumenta la complejidad (número de clases) y disminuye el volumen de datos, este modelo pierde efectividad y su capacidad de visualización se ve limitada.
- **Árboles de Decisión:** dividen los datos en función de sus características, facilitando la interpretación y permitiendo el manejo eficiente de variables categóricas y continuas.
- **Random Forest:** consiste en un conjunto de múltiples árboles de decisión, lo que genera un modelo más robusto. Al combinar varios árboles, reduce el riesgo de sobreajuste (overfitting), mejorando la precisión en comparación con un árbol de decisión individual.
- **XGBoost:** algoritmo basado en árboles de decisión que ajusta sucesivamente los modelos para mejorar su precisión. Gracias a su capacidad para aplicar regularización (Lasso y Ridge) de manera simultánea, es un modelo altamente robusto, especialmente adecuado para grandes volúmenes de datos.
- **Regresión Logística:** modelo lineal utilizado para predecir la probabilidad de eventos en clasificación. Se destaca por su simplicidad, eficiencia y facilidad de interpretación, lo que lo hace ideal en escenarios donde la interpretabilidad es clave.

Matriz de confusión					
		Predición			
		Non	Very Mild	Mild	Moderate
Real	Non	True	False	False	False
	Very Mild	False	True	False	False
	Mild	False	False	True	False
	Moderate	False	False	False	True

$$\text{Precision} = \frac{VP}{VP+FP}$$

$$\text{Recal} = \frac{VP}{VP+FN}$$

$$\text{Acurácia} = \frac{\text{Total de previsones correctas}}{\text{Total de previsiones}}$$

Pasos seguintes

1. Finalización de los modelos utilizando la selección previa de variables.
2. Realización de modelos utilizando las variables creadas en el PCA.
3. Validación de los modelos mediante el uso de las metricas elegidas.
4. Aplicación de las técnicas de balanceo directamente en imágenes.

Bibliografía

1. Kumar S y Shastri S. Alzheimer MRI Preprocessed Dataset. 2022. doi: 10.34740/KAGGLE/DSV/3364939. Available from: https://www.kaggle.com/dsv/3364939
2. Caballero Testón, E., Álvarez Liébana, J., & Calviño Martínez, A. (2024). Trabajo fin de máster en bioestadística: Clasificación supervisada de imágenes de resonancia magnética para la detección de enfermedad de Alzheimer. Julio 2024.
3. Fernández A, García S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863-905. Submitted 06/17; published 04/18.