Aplicaciones de ML al estudio de la metástasis cerebral

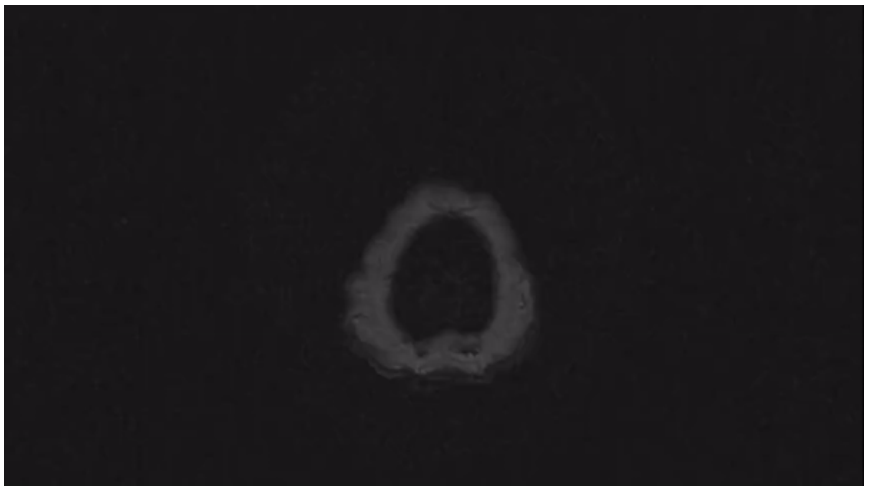
¿Quiénes somos?

- Adrián Aguado García
- Arturo Linares Muñiz
- Mateo Rodríguez Suárez
- Olumayowa Onabanjo
- Raquel Martínez Martínez

¿Qué es la metástasis cerebral?

El cáncer es una enfermedad en la que las células anormales se multiplican y crecen sin control, invadiendo tejidos y órganos. Este crecimiento puede formar tumores y diseminarse por el cuerpo. Cuando estas células cancerosas llegan al cerebro, se produce una **metástasis cerebral**. La detección temprana y tratamientos avanzados son esenciales para mejorar la calidad de vida.

Aunque suene algo triste, por tratarse de un cáncer terminal, merece la pena conocer el tiempo que les queda a estos pacientes para disfrutar de la compañía de los suyos, así como ayudar a prolongar ese tiempo y la calidad de este con tratamientos acertados.



¿Cuál es nuestro objetivo?

Uno de nuestros objetivos es predecir con la mayor precisión cuánto tiempo queda para disfrutar la vida. En segundo lugar, y posiblemente más importante aún, es ayudar a los médicos a predecir cuándo es probable la recaída de un paciente y métodos de tratamiento más eficaces

¿Cuál es nuestro Dataset?

Trabajamos con MOLAB Brain Metastasis Datase. MOLAB (*Mathematical Oncology Laboratory*) es un grupo de investigación multidisciplinar que desarrolla estudios matemáticos sobre problemas relacionados con el cáncer. El dataset está formado por un conjunto de datos estructurado y anonimizado que cubre escaneos de resonancia magnética (MRI) segmentados, datos clínicos y mediciones morfológicas para 75 pacientes.
Tenemos 3 grupos de archivos con datos estructurados y no estructurados.

- Un archivo Excel con los **datos clínicos** del paciente y su historial de tratamiento.
- Un archivo Excel con las **medidas morfológicas**, las dimensiones de los tumores identificados a partir de la resonancia magnética.
- Las **resonancias magnéticas** (MRI) contenidas en un archivo zip en varios formatos, como *nifti* y *DICOM*

📁Datos clínicos

Tabla 1: *Clinical_data*

Variables	**Significado**
Patient Data (Datos del paciente)	Esta sección incluye datos como la edad en la 1ª exploración y el sexo
Lesion (Lesión)	Los tumores se enumeran en el orden en que se descubren
Primary Tumor (Tumor primario)	Los tumores primarios se clasifican por número (1 - 12) y subtipo para las clases más comunes (cáncer de mama y pulmón)
Whole Brain Radiation Therapy (WBRT) (Radioterapia de todo el cerebro)	Como su nombre indica (incluye dosis y periodo administrado)
Stereotactic Radiosurgery (SRS)	Una radioterapia no quirúrgica
Radiation Necrosis (Necrosis por radiación)	Descomposición tisular debida a la radiación
Surgery (Cirugía)	Describe el tipo (total, parcial, desconocida) y el momento en que se llevó a cabo
Systemic Treatment (Tratamiento sistemático)	Fármacos administrados y periodo de tiempo
Death (Muerte)	El tiempo, medido en días y la causa

MRI Follow up dates	Mínimo 4 citas, máximo 19

🗒️ *Medidas morfológicas*

Tabla 2: *Morphological_measures*

Variables	**Significado**
General data	Número de identificación del paciente, información sobre la imagen y el equipo utilizado para capturarla.
Other measurement data (Otros datos de medición) (No dar demasiada importancia a esto. No nos hemos convertido en expertos en resonancia magnética)	• SBS (Space Between Slices, espacio entre secciones) en milímetros • SCS (Slice Thickness, espesor de la sección) en mimíletros • REPTIME (Tiempo de repetición) en milisegundos. • CEVOLUME (Contrast-Enhancing Volume, Volumen con contraste de aumento) V_{CE} • NECVOLUME (Necrotic (Non-Enhancing) Volume, Volumen necrótico, sin aumento) V_N • TOTALVOLUME $V = V_{CE} + V_N$ • CERIMWIDTH: Contrast-Enhancing (CE) Spherical Rim Width, función del volumen que calcula la anchura media de las áreas CE. • SURFACEREGULARITY: cuanto más próximo a 1 se acerca a un tumor más esférico • MAXDIAMETER3D: máxima medida longitudinal del tumor

🧠 *Resonancias Magnéticas*

La base de datos (*MRI_scans*) incluía imágenes ya procesadas (Figura 1). Una de las ventajas de trabajar con esta base de datos fue todas las dimensiones de las diferentes imágenes claramente extraídas.

PATIENT N10020 FEMALE AGE 44 TIME OF DEATH 1625 DAYS

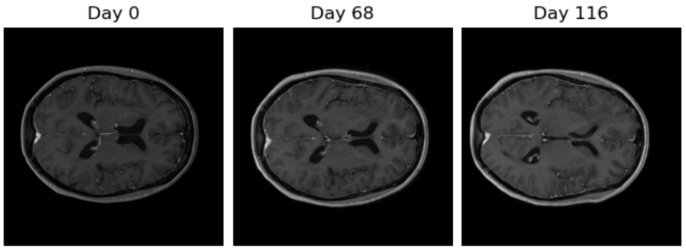


Figura 1: Ejemfdfdfddfplo de resonancia magnética a lo largo del tiempo.

Nuestra estrategia

Mientras trabajábamos con este conjunto de datos, éramos conscientes de varios factores que podrían afectar a la precisión de nuestro modelo, como el tamaño, el desequilibrio de clases, la alta dimensionalidad y los valores atípicos. Ahora explicaremos cómo exploramos nuestros datos, así cómo la decisión por un enfoque de machine learning (ML) y los resultados logrados.

Preparación de los datos (preprocesamiento y limpieza)

Los datos clínicos y las medidas morfológicas estaban estructurados de forma diferente en archivos separados. Organizar los datos en un único conjunto nos facilitó la comprensión de nuestros modelos. Además, había celdas vacías y problemas de formato de datos que debíamos resolver.

Análisis de datos

Hizo falta un montón de diagramas, gráficos y código para entender la relación entre nuestras variables y saber qué podíamos predecir con mayor certeza. Mientras trabajamos con nuestros datos, fuimos siendo conscientes de los factores que podían afectar a la precisión de nuestro modelo, como el tamaño, el desequilibrio de clases y los valores atípicos (Figuras 2 y 3).

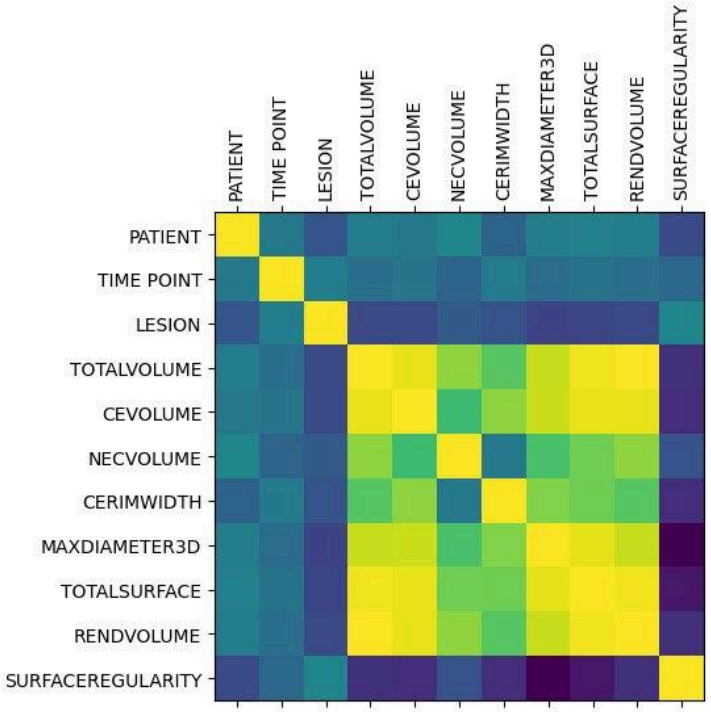


Figura 2: Matriz de correlación para las medidas morfológicas. Alta correlación entre variables derivadas entre sí.

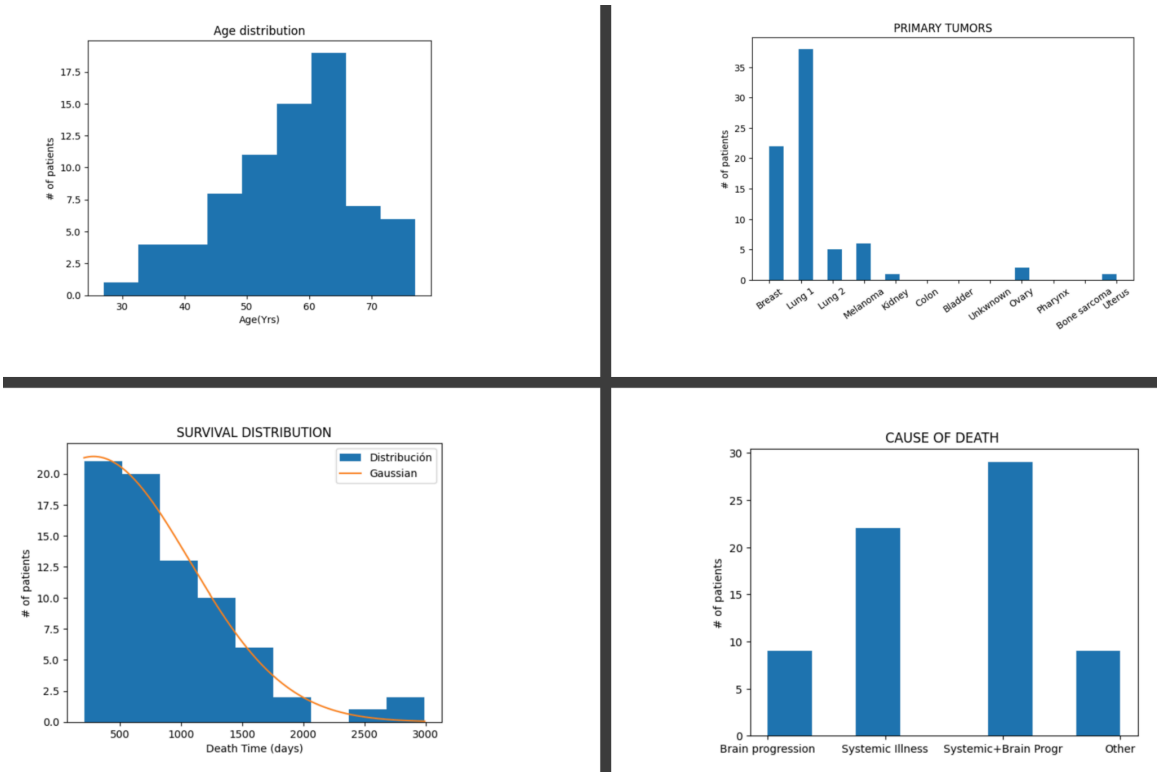


Figura 3: Distribución de datos clínicos (edad, tumor primario, tiempo de supervivencia y causa de muerte).

Enfoque de predicción

Ha sido todo un reto seleccionar a qué nivel aplicar nuestro modelo (paciente, tumor, evolución en el tiempo o una combinación de ambos). La Figura 4 es un claro ejemplo del volumen de la lesión. Vimos que sería difícil hacer una predicción de series temporales, ya que la evolución del tumor variaba mucho.

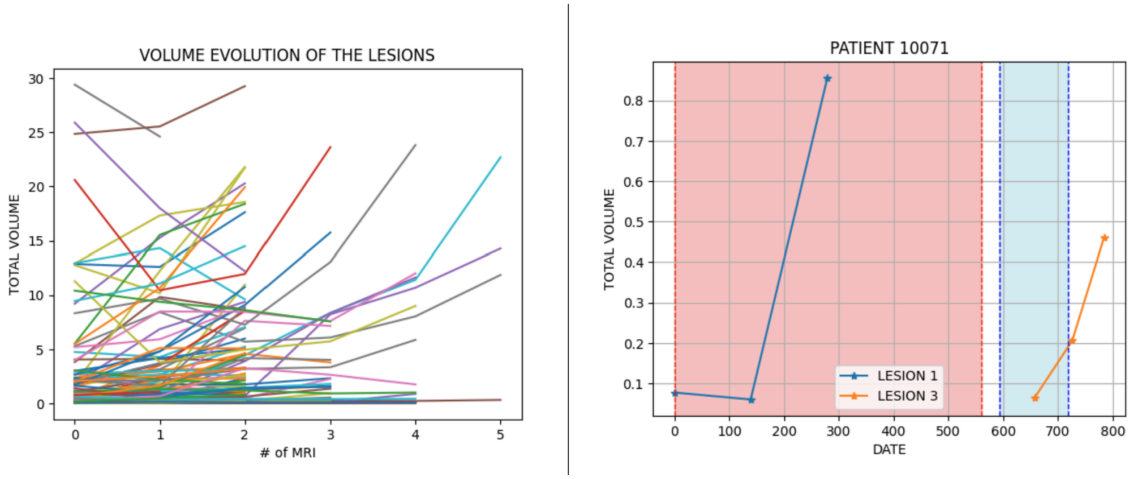


Figura 4: Evolución temporal del volumen total del tumor visto por tumor y por paciente.

Selección de modelos

Los datos de solo 75 pacientes nos limitó la selección de modelos. La Figura 5 muestra los principales grupos de métodos de ML con los que hemos utilizado resaltados en rojo (Decision trees, Logistic Regression, Neural Networks, Random Forests, and XGBoost).

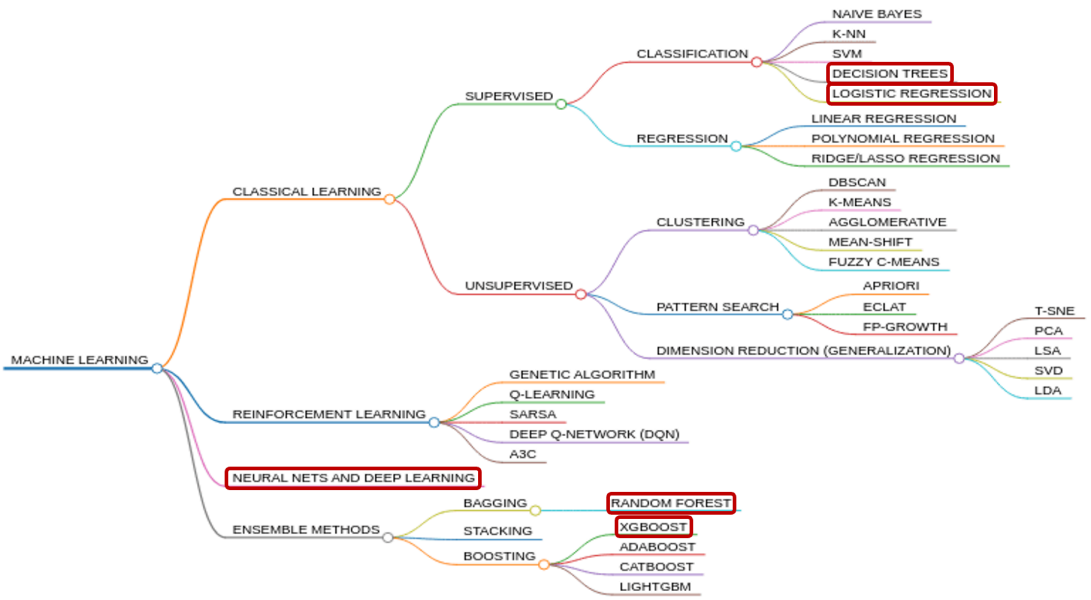


Figura 5: Mapa conceptual de métodos de ML que muestra modelos seleccionados de los principales grupos de ML (Classical, Reinforcement, Neural Networks/Deep Learning & Ensemble methods), resaltando los utilizados en nuestro proyecto.

Entrenamiento de modelos

Los principales objetivos se tradujeron en variables que nuestros modelos debían predecir.

- **Objetivo 1**: Predecir el tiempo de supervivencia.
- **Objetivo 2**: Predecir la recaída de la lesión.
    - (a) Para todos los pacientes.
    - (b) Para los pacientes con cirugía.

Se eligieron dos enfoques diferentes, binario y multiclase, para comparar el rendimiento del modelo:

Clases binarias

Grupos de clasificación binaria, **Objetivo 1** (Supervivencia)

- Supervivencia <500 días
- Supervivencia ≥500 días

Grupos de clasificación binaria, **Objetivo 2, a y b**. (Recaída)

- Recaída: crecimiento del tumor > 30%
- Sin recaída: crecimiento del tumor < 30%

Multiclases

Grupos de clasificación binaria, **Objetivo 1** (Supervivencia)

- Supervivencia <500 días
- Supervivencia ≥500 días o <1500 días
- Supervivencia ≥1500 días

Grupos de clasificación binaria, **Objetivo 2, a y b**. (Recaída)

- Recaída: crecimiento del tumor > 30%
- Sin recaída: crecimiento del tumor 0 - 30%
- Mejora: crecimiento del tumor < 0%

Parámetros del modelo

Para llegar a unos resultados óptimos se utilizaron varios parámetros.

- **Profundidades del árbol:** 3 - 5
- **Reparto de los datos de entrenamiento / test:** 80 / 20 %

Resumen de los datos de entrada

Tabla 3: *Input Data Summary*

Objetivo	Tamaño	Entradas	Variable a predecir
1	75 filas 5 columnas	sex (categorical) age (numerical discrete) gpa (numerical discrete) nlesions (categorical ordinal) type (categorical ordinal)	daystodeath (numerical discrete)
2 (a y b)	345 filas (a) 105 filas (b) 13 columnas	sex (categorical) age (numerical discrete) gpa (numerical discrete) cermwidth (numerical continuous) surfaceregularity (numerical continuous) totalvolume (numerical continuous) srs (numerical continuous) days_since_srs (numerical discrete) wbrt (numerical continuous) days_since_wbrt (numerical discrete)	recaída/relapse (categorical boolean)

Resultados

Objetivo 1: Predecir el tiempo de supervivencia

Tabla 4: Resultados del entrenamiento del modelo (binario)

Modelo	Exactitud	AUC*	Comentarios
Logistic regression	71	-	Semillas: 10000
Decision Tree	72	67	La profundidad de árbol más eficaz es 4
Random Forest	73	-	Número de factores de estimación más eficaz: 15, Iteraciones: 5000
XGB	73	75	Número de factores de estimación más eficaces: 4, profundidad: 4, iteraciones: 10000
Neural networks	75	-	Periodos: 5, Semillas: 1000 (mayor incertidumbre que en los ensayos anteriores)

*AUC (**A**rea **U**nder ROC **C**urve, Área por debajo de la curva ROC)

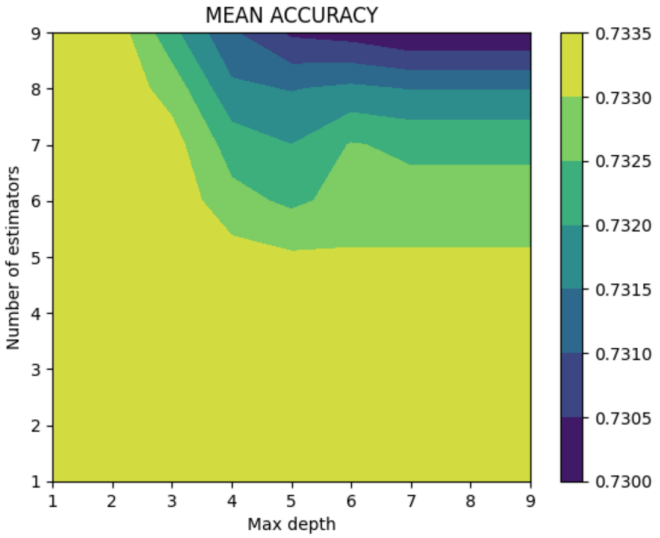


Figura 6: Selección de los parámetros del modelo más eficaz (XGBoost).

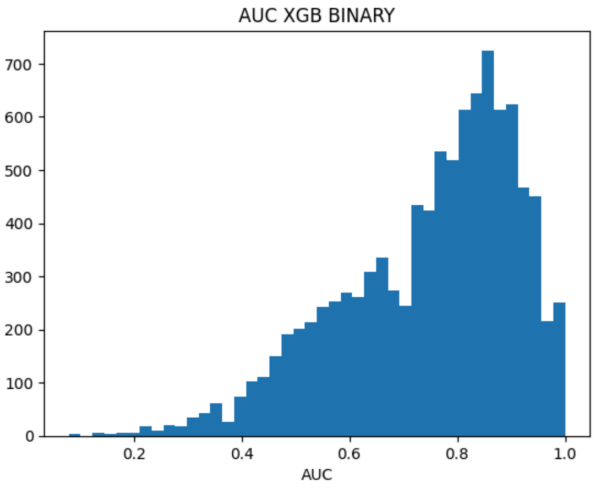


Figura 7: Comportamiento AUC (XGBoost).

Tabla 5: Resultados del entrenamiento del modelo (multiclase)

Entrenamiento del modelo		**Resultados**
Objetivo	**Clase**	
1	Multiclase	La precisión de los modelos osciló entre el **35** y el **46 %**. Mejor rendimiento XGBoost.

Objetivo 2: Predecir la recaída de la lesión

- (a) Para todos los pacientes.
- (b) Para los pacientes con cirugía.

Tabla 6: Resultados del entrenamiento del modelo (binario y multiclase)

Entrenamiento del modelo		**Resultados**
Objetivo	**Clase**	
2 (a)	Binario	La precisión de los modelos osciló entre el **47** y el **58 %**. Los árboles de decisión mostraron el mejor rendimiento.
	Multiclase	La precisión de los modelos osciló entre el **43** y el **50 %**. Mejor rendimiento XGBoost.
2 (b)	Binario	La precisión de los modelos osciló entre el **52** y el **59 %**. Mejor rendimiento con la regresión logística.
	Multiclase	La precisión de los modelos osciló entre el **33** y el **41 %**. Mejor rendimiento XGBoost.

Conclusiones

El modelo obtuvo los mejores resultados en la predicción de los días de supervivencia; se pueden mejorar los resultados con más datos. Los resultados de la predicción de recaída fueron peores debido a la mayor variación. En general, el rendimiento del modelo fue más bajo en las clasificaciones multiclase que en las binarias. El rendimiento de los modelos depende de la función, lo que destaca la importancia de experimentar con varios modelos para comparar los resultados.

Hay factores que influyen en ambos resultados y que no se recogen en nuestro conjunto de datos. Aún así, representa un buen punto de partida para estudios comparables sobre enfermedades con un tamaño de muestra de pacientes relativamente reducido.

Debemos destacar lo prometedores que son estos resultados. Con tan solo 75 pacientes somos capaces de tener un acierto de casi el 75% al predecir si un paciente va a sobrevivir más o menos de 500 días. En el caso de la predicción de recaídas, es evidente que tan solo estamos en el comienzo: se necesitan mejores datos, más cantidad de los mismos, y utilizar un mayor número de enfoques a la hora de ajustar los modelos.

Diseño del ML

El propósito de nuestro modelo es hacer predicciones fundamentales para la vida de las personas y sus decisiones de futuro. Así que también nos hemos asegurado de que sea transparente tanto para nosotros como para los pacientes. Para ello, hemos añadido los valores SHAP (*SHapely Additive exPlanations*), representados en la Figura 6, enlaces a los datos y a nuestro código. ¿Por qué? Porque sabemos que es más difícil confiar en la IA cuando está rodeada de misterio.

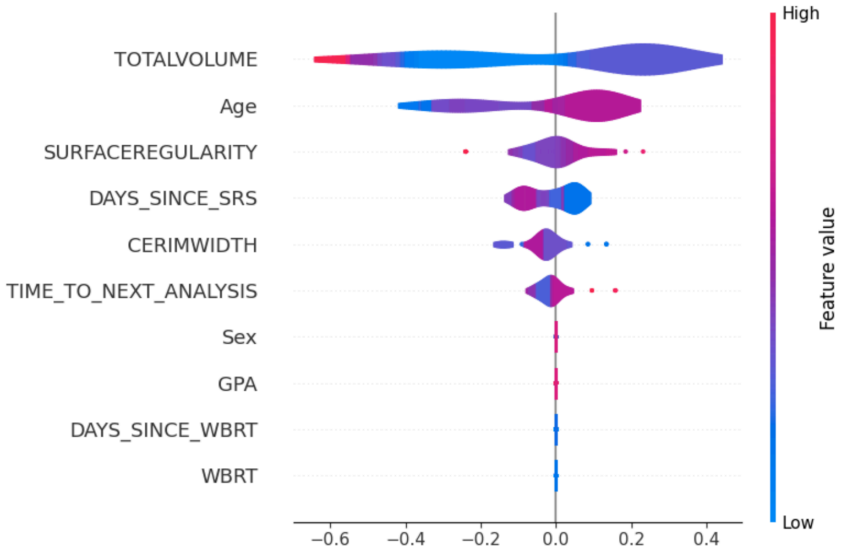
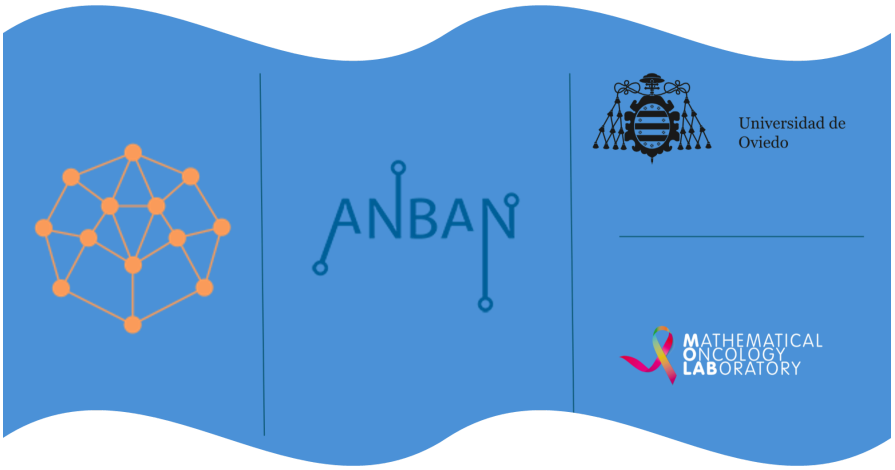


Figura 8: Valor SHAP (impacto en el resultado del modelo)

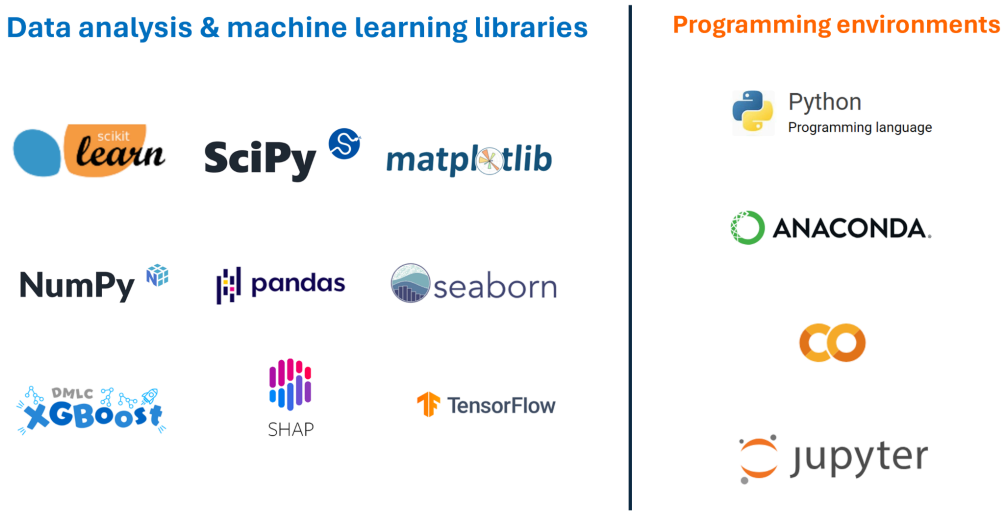
Disclaimer: los propios modelos no son perfectos y NUNCA deben utilizarse sin consultar a un profesional.

¿Cómo lo conseguimos?

Con trabajo en equipo y ayuda de nuestro coordinador, instructores, mentores, los organizadores de SaturdaysAI Asturias, nuestro anfitrión, y el maravilloso equipo de MOLAB.



Algunos apenas sabíamos nada de Python y en sólo 12 semanas juntos, entrenamos nuestro modelo ML. Esperamos que te sirva de ayuda en tu camino para mejorar la calidad de vida con inteligencia artificial. Estaremos encantados de recibir sus opiniones y comentarios sobre nuestro proyecto de ML.



Agradecemos el apoyo de la IA Generativa (ChatGPT, Bing Copilot y DALL-E).

Líneas futuras

Para quien desee unirse a nuestra lucha contra la metástasis cerebral, a continuación se indican algunas vías que nos hubiera gustado trabajar.

- Descubrimos que la causa de la muerte era una variable que se podía predecir durante nuestra exploración de datos. Sin embargo, no pudimos trabajar esta opción por falta de tiempo.
- Ajustar el umbral del porcentaje de aumento del tumor para definirlo como recaída. Utilizamos un crecimiento del 30% como criterio.
- Incrementar el tamaño del conjunto de datos con redes generativas de tipo adversarial.
- No se obtuvieron mejores resultados separando a los pacientes operados del resto. No tiene sentido separar a estos pacientes, aunque merece la pena estudiar si la división del conjunto de datos por otros criterios dará mejores resultados.

Recursos

- MOLAB
- MOLAB brain metastasis dataset
- Brainlab.org
- Brain metastases: A documentary