Universidad de Valparaíso
CHILE

Facultad de Ciencias
Instituto de Estadística
Ingeniería en Estadística

Metodología de Radiómica basada en Inteligencia Artificial para la identificación de biomarcadores genéticos de tumores cerebrales en Imágenes de Resonancia Magnética.

Sebastián Ponce Olguín
March 28, 2024

Profesor Guía
Ph.D. Rodrigo Salas Fuentes
Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso

Profesor Co-Guía
Ph.D. Steren Chabert
Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso

Proyecto de titulación para optar a
Grado académico de *Magister en Estadística*

Resumen

Una característica de los tumores cerebrales es la presencia del biomarcador genético O6-Metilguanina-DNA-Metiltransferasa (MGMT), el cual es un indicador que, en el caso de estar presente apunta a un eventual mejor pronóstico, debido a su asociación con una respuesta favorable a la quimioterapia. Actualmente, la única forma de detectar su presencia es mediante un procedimiento invasivo de biopsia cerebral. El objetivo de este estudio es desarrollar una metodología de radiómica basada en minería de datos, para la identificación y evaluación de biomarcadores de tumores cerebrales en imágenes de resonancia magnética. Para lograr este objetivo, se utilizaron imágenes de resonancia magnética a las cuales se les aplicó una máscara de segmentación, para luego a partir de éstas, extraer variables. Una vez realizada esta etapa, se emplearon técnicas de selección de variables como LASSO, backward selection, forward selection, stepwise, Fisher-score y recursive feature elimination, para finalmente utilizar seis modelos de machine learning: Logistic Regression, Random Forest, XGBoost, Support Vector Machine, Gradient Boost y CatBoost. El modelo de clasificación que obtuvo los mejores resultados fue Logistic Regression, utilizando la técnica de selección LASSO. En el análisis de shap values, se identificaron tres variables de mayor importancia, una variable de secuencia T1CE relacionada a la textura, una variable de la secuencia FLAIR del tipo de estadísticas de primer orden y una variable de la secuencia T1 del tipo de estadísticas de primer orden.

Contents

Resumen **2**

1 Introducción **5**
1.1 Objetivos . . . 6

2 Marco Conceptual **7**
2.1 Cáncer Cerebral . . . 7
2.1.1 Gen O6-metilguanina-ADN-metiltransferasa (MGMT) . . . 7
2.2 Diagnóstico por neuroimagen del glioblastoma . . . 8
2.3 Radiómica . . . 9
2.4 Adquisición y reconstrucción de imágenes. . . . 10
2.4.1 Segmentación de imágenes. . . . 14
2.4.2 Extracción de variables o características. . . . 14
2.5 Selección de variables. . . . 21
2.6 Construcción de un modelo. . . . 21
2.7 Predicción y prognosis. . . . 21
2.8 Aprendizaje Automático (Machine Learning) . . . 21
2.9 Logistic Regression . . . 22
2.10 Random Forest . . . 23
2.11 CatBoost . . . 24
2.12 Gradient Boost . . . 24
2.13 XGBoost . . . 25
2.14 Support Vector Machine (SVM). . . . 25
2.15 t-Distributed Stochastic Neighbor Embedding (t-SNE). . . . 26
2.16 Interpretabilidad y explicabilidad. . . . 26
2.17 Explicabilidad por medio de shap values . . . 29

3 Estado del arte **31**

4 Metodología **34**
4.1 Software . . . 35
4.2 Conjunto de datos . . . 36
4.3 Análisis exploratorio . . . 36
4.4 Armonización de imágenes . . . 36
4.5 Extracción de variables. . . . 37
4.6 Selección de variables. . . . 37
4.7 Modelos y evaluación de desempeño . . . 40
4.8 Shap values para explicabilidad . . . 42

5 Resultados **43**
5.1 Histogramas 43
5.2 Armonización por coincidencia de histogramas 44
5.3 Base de datos generada con variables extraídas 47
5.4 Análisis exploratorio con PCA y t-SNE 48
5.5 Métricas de los modelos 52
5.6 Shap values 54

6 Discusión **55**

7 Conclusión **57**

Agradecimiento **65**

Chapter 1

Introducción

En Chile la mortalidad causada por presencia de "tumor maligno de las meninges, encéfalo y de otras partes del sistema nervioso central" fue de 2.6 por cada 100.000 habitantes durante el año 2013, siendo la segunda causa de mortalidad infantil por cáncer, luego de la leucemia, con una mortalidad de 0.6 por cada 100.000 habitantes (Contreras, 2017). En USA son diagnosticados entre 30.000 a 35.000 tumores cerebrales primarios cada año, siendo el más común el glioblastoma multiforme. El pronóstico general y la supervivencia de los pacientes con este diagnóstico es desfavorable, con una tasa de supervivencia global a 5 años del 20% (Newton, 2016).

Los tumores cerebrales pueden ser divididos en 2 grandes grupos:

- Lesiones primarias, que se originan a partir de las células del sistema nervioso central, y
- Lesiones secundarias, que se originan en otros lugares del cuerpo y producen metástasis en el cerebro.

Dentro de los tumores primarios más comunes están el meningioma y el glioblastoma, y los tumores primarios a su vez se pueden clasificar en I y II como "bajo grado", y en III y IV como de "alto grado" (Contreras, 2017).

Existe un biomarcador genético, que es la metilación del promotor de O6-Metilguanina-DNA-Metiltransferasa (MGMT), que al estar metilado es signo de mejor pronóstico, ya que se asocia a un indicador de buena respuesta a la quimioterapia (Suh et al., 2018). Una desventaja que presenta el diagnóstico, para confirmarlo es necesaria realizar una biopsia, procedimiento invasivo no exento de complicaciones. Por otro lado, el diagnóstico puede ser influenciado por la heterogeneidad temporal y espacial del tumor, reduciendo el valor diagnóstico (Suh et al., 2018).

La rapidez de diagnóstico y su precisión dependen de la experiencia del radiólogo o neurorradiólogo (Soltaninejad et al., 2018), sin embargo también implica algunos inconvenientes para ellos, ya que se generan una gran cantidad de datos. Por otro lado, en la actualidad cada vez se realizan más exámenes imagenológicos, lo que se traduce en un aumento de su carga laboral.

En el último tiempo se ha desarrollado la radiómica, que consiste en describir un amplio conjunto de métodos computacionales que extraen características cuantitativas a partir de las imágenes médicas. Estas características se pueden utilizar para mejorar el diagnóstico por imágenes, el pronóstico y la respuesta a la terapia oncológica (Zhou et al., 2018).

El Instituto Nacional de la Salud de Estados Unidos (*National Institutes of Health (NIH)*) estableció la definición de biomarcador como aquellas características biológicas, bioquímicas, fisiológicas, antropométricas, etc., objetivamente medibles, capaces de identificar procesos fisiológicos o patológicos, o bien una respuesta farmacológica a una intervención terapéutica (Torres Courchoud & Pérez Calvo, 2016). Existen diferentes tipos de biomarcadores: diagnósticos, pronósticos, o terapéuticos. El biomarcador ideal debe ser específico, sensible, predictivo, rápido y económico, estable *in vivo* e *in vitro*, no invasivo, y que tenga suficiente relevancia preclínica y clínica como para modificar las decisiones relativas al proceso patológico en que se aplica (Torres Courchoud & Pérez Calvo, 2016).

Algunos ejemplos de biomarcadores son la proteína C reactiva como marcador de isquemia intestinal, el antígeno prostático específico para cáncer de próstata, ca19-9 para cáncer de páncreas y vesícula biliar, receptores de somatostatína para tumores neuroendocrinos de páncreas o gástricos, y el MGMT como marcador de metilación de tumor cerebral (del Cancer, 2022).

Existe entonces la necesidad de mejorar los procedimientos de obtención de biomarcadores no invasivos, como en este caso la detección de la metilación del promotor de MGMT, ya que se obtiene de manera de invasiva, por lo cual se podría hacer uso de imágenes de resonancia magnética y de esta forma obtener nuevos biomarcadores, para así utilizarse para la predicción del biomarcador de la metilación del promotor de MGMT.

Por lo tanto, a partir de lo anteriormente mencionado establecemos la siguiente pregunta de investigación: ¿es posible desarrollar una metodología de radiómica basada en minería de datos para identificar y evaluar biomarcadores de tumores cerebrales en resonancias magnéticas?.

1.1 Objetivos

El objetivo general de esta Tesis sigue a continuación.

Objetivo General: Identificar el biomarcador genético O6-Metilguanina-DNA-Metiltransferasa (MGMT) de tumor cerebral a partir de imágenes de resonancia magnética mediante una metodología que involucra radiómica y aprendizaje automático explicable.

Los objetivos específicos de la Tesis son los siguientes:

Objetivos específicos:

1. Extraer características a partir de imágenes de resonancia magnética del cerebro, mediante la utilización de métodos de procesamiento de imágenes.

2. Identificar biomarcadores genéticos de tumores cerebrales que se encuentren presentes en imágenes de resonancia magnética, mediante la aplicación de técnicas de aprendizaje automático.

3. Investigar e incorporar elementos de explicabilidad en modelo de aprendizaje automático, para la identificación de biomarcadores genéticos en tumores cerebrales, mediante el uso de imágenes de resonancia magnética.

Chapter 2

Marco Conceptual

2.1 Cáncer Cerebral

El glioblastoma, también conocido como glioblastoma multiforme (GBM), se origina en las células gliales y ependimarias, que son el tipo de células de soporte del tejido cerebrales. Estas células gliales son responsables de mantener la estructura y el funcionamiento normal del cerebro. El GBM corresponde al subtipo histológico de mayor malignidad (grado IV) y representa más del 80 % de ellos (Dolecek et al., 2012).

El GBM tiene, por características histológicas, el presentar una necrosis, proliferación vascular o ambos, también con un incremento en la celularidad y pleomorfismo (Louis et al., 2007). Uno de los principales problemas del GBM, es que su tasa de incidencia es varía entre 0,59 y 3,69 por 100.000 personas y la sobrevida a 5 años se mantiene en 5%, a pesar de los avances en su tratamiento durante las últimas décadas (Ostrom et al., 2015).

2.1.1 Gen O6-metilguanina-ADN-metiltransferasa (MGMT)

Es importante considerar en el GBM, la metilación del promotor del gen O6-metilguanina-ADN-metiltransferasa (MGMT). Este gen se encuentra en la banda cromosómica 10q26 y es responsable de la producción de una enzima reparativa del ADN que elimina los grupos alquilos de la posición O6 de la guanina, evitando el emparejamiento incorrecto de bases y la muerte celular. La presencia de la enzima MGMT está relacionada con la protección de las células normales contra la carcinogénesis y también contra el efecto letal de la quimioterapia con agentes alquilantes (Pegg, 2000)

La expresión de MGMT está mediada por la metilación de su promotor, cuando existe una metilación aberrante de las citocinas de estos dinucleótidos, se puede dar un silenciamiento de manera parcial o total del gen, por lo cual se desarrolla una disminución o ausencia de la proteína MGMT, esto podría explicar la mayor susceptibilidad a la quimioterapia que presentará el GBM (Hegi et al., 2005). Los tumores de GBM metilados se encuentran entre 33% y 60% de los pacientes, dependiendo del estudio y del método utilizado (Quillien et al., 2012).

Para lograr una identificación del estado de MGMT, se requiere realizar en primera instancia, una biopsia invasiva del GBM, para posteriormente determinar por medio del nivel de ADN la metilación del promotor. También por medio del nivel de ARN mediante PCR de transcripción inversa (RT-PCR), o con el nivel de proteína mediante inmunohistoquímica (IHC) (Quillien et al.,

2012). Existe otra técnica, llamada pirosecuenciación (PSQ), que ha demostrado tener una buena predicción con una alta sensibilidad y reproducibilidad (Xie et al., 2015).

En cuanto al tratamiento de los pacientes con GBM, este se compone generalmente por una resección máxima segura del tumor, quimioterapia con temozolomida (TMZ), seguida de la terapia adyuvante con TMZ, con o sin campos eléctricos para el tratamiento de tumores "Optune" (TTField). La acción terapéutica de TMZ es provocar daños en el ADN al desencadenar la vía de reparación de errores de emparejamiento del ADN y la muerte de las células tumorales (Nguyen et al., 2021). Sin embargo, algunas células tumorales pueden revertir el efecto del daño en el ADN expresando y sobreexpresando la enzima reparadora de ADN MGMT. Esta proteína elimina la alquilación de la posición O6 de la guanina, revirtiendo y neutralizando así el efecto citotóxico de TMZ (Esteller et al., 2000). Por lo mencionado anteriormente, el estado de metilación del promotor de MGMT es un biomarcador importante, con el cual se puede predecir, si el efecto farmacológico de TMZ en el tratamiento de GBM, y por consecuencia, tener una aproximación pronóstica.

2.2 Diagnóstico por neuroimagen del glioblastoma

Para el estudio del sistema nervioso central, la técnica imagenológica que entrega una gran información, es la resonancia magnética (RM). La RM se caracteriza por la detallada visualización de las estructuras anatómicas de la línea media cerebral, como son la hipófisis, el hipotálamo, la glándula pineal y el tallo cerebral (Atlas, 2016). Además, es una prueba segura e indolora que utiliza, un campo magnético y ondas de radiofrecuencia para obtener las imágenes detalladas del cerebro (Medelo Ballesteros et al., 2019).

La guía NICE (National Institute for Health and Care Excellence) recomienda de manera estándar las siguientes secuencias como protocolo de exploración de glioma como prueba de diagnóstico inicial para la sospecha de gliomas: T2, FLAIR, serie DWI y volumen T1 antes y después del contraste (for Health & Care Excellence, 2018).

Las imágenes de RM en GBM presentan varias características distintivas, entre ellas, se encuentran la necrosis, el realce al medio de contraste, el efecto de masa, el tejido peritumoral y el desplazamiento de la línea media. Estas características resaltan la complejidad macroscópica del glioblastoma. En general, se suele observar una única lesión con realce en la periferia. Sin embargo, también pueden aparecer múltiples áreas de realce dentro de una región más extensa de señal anormal en la secuencia ponderada en T2 (glioblastoma multifocal). Además, pueden presentarse regiones discretas de realce sin evidencia de conexión tumoral (glioblastoma multicéntrico) (Shukla et al., 2017). En la figura 2.1, se puede ver una imagen de secuencia ponderada T1 posadministración de contraste basado en gadolinio, en el cual se puede observar realce periférico irregular heterogéneo asociado con la masa del lóbulo frontal derecho con área central sin realce, que es signo de necrosis (Shukla et al., 2017).

En un metanálisis relacionado con GBM, se demostró que la resección total macroscópica se asoció con una mejor supervivencia en comparación con la resección subtotal (Brown et al., 2016). Después de la resección segura máxima, que se evalúa en la resonancia magnética posoperatoria inmediata, el tratamiento es la quimiorradiación con TMZ (Stupp et al., 2009). La planificación de la radioterapia incluye el registro de la RM posoperatoria, principalmente las secuencias T1CE y T2-FLAIR, con la simulación de planificación de la tomografía computada, lo que permite delimi-

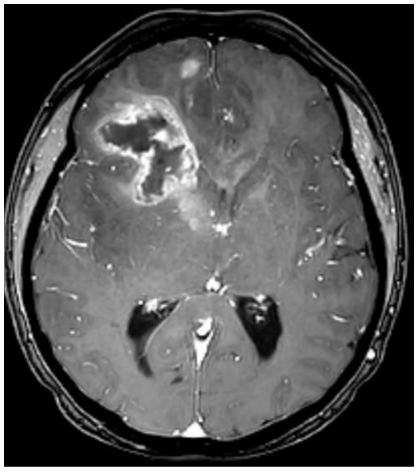

Figura 2.1: Imagen axial potenciada en T1 (T1CE)

tar las anomalías T2-FLAIR y el realce residual, es por esto que la RM posoperatoria tiene como objetivos definir los volúmenes objetivos y los órganos en riesgo.

En la figura 2.2, la imagen T1CE postoperatoria, presenta un leve realce residual, especialmente a lo largo de los aspectos mediales del sitio quirúrgico, lo cual podría indicar la presencia de un tumor residual mínimo. La mayor parte de la señal hiperintensa en la región parieto-occipital derecha se atribuye a la presencia de hemoderivados postoperatorios (Shukla et al., 2017).

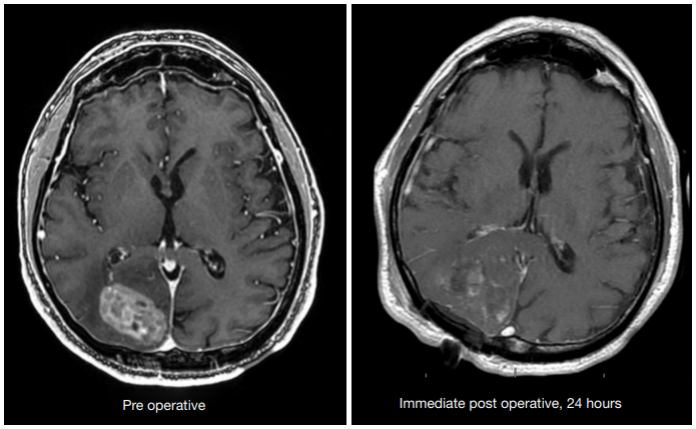

Figura 2.2: Imagen axial potenciada en T1CE, pre operatoria y post operatoria

2.3 Radiómica

La primera aparición publicada del término radiómica fue en 2012 (Lambin et al., 2012). Radiómica consiste en extraer información cuantitativa de alguna modalidad de imagen médica, con la intención de crear bases de datos para realizar análisis (Avanzo et al., 2020; Kumar et al., 2012). Una hipótesis, es que la información que refleja la genómica y los patrones proteómicos, se encuentran y se pueden identificar por medio del análisis de imágenes macroscópicas (Lambin et al., 2012), por lo tanto, uno de los beneficios de un biomarcador de imágenes es su no invasividad.

El concepto de radiómica, que se ha aplicado más ampliamente en el campo de la oncología, aunque puede ser aplicada a otro tipo de enfermedades, por ejemplo cardiovasculares, se basa en la suposición de que las imágenes biomédicas contienen información de procesos específicos de la enfermedad y que son accesibles mediante análisis cuantitativos de imágenes (Neisius et al., 2019). Aunque muchos de los conceptos de extracción de características de imagen han existido durante décadas, la producción de investigación en el campo ha crecido exponencialmente, con más de 1500 publicaciones en 2020 que contienen el término radiómica, como se representa en la figura 2.3

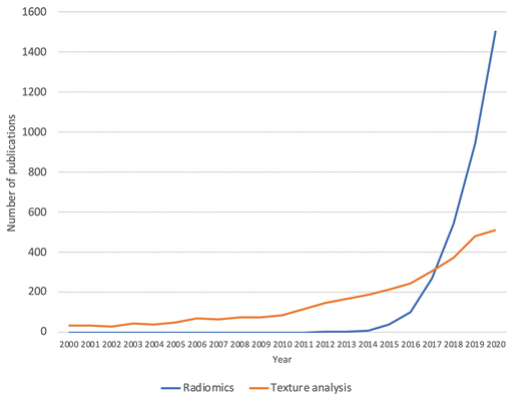

Figura 2.3: Producción de investigación en radiómica

La radiómica permite la traducción de imágenes médicas en puntos de datos multidimensionales, por ejemplo, por medio del análisis de textura, que cuantifica objetivamente la textura de las imágenes radiológicas al explotar las relaciones entre píxeles. Estas métricas pueden usarse para diagnosticar anomalías en imágenes radiológicas que no pueden ser visto solo por los ojos de los radiólogos (Mannil et al., 2018). Es importante resaltar, que la radiómica proporciona datos adicionales a los existentes, y que no implica ninguna automatización de los procesos de diagnóstico.

Junto con la inteligencia artificial, estas características sirven de biomarcador y pueden usarse para construir modelos predictivos, que puedan proporcionar una evaluación más allá de lo que es logrado mediante la evaluación cualitativas de las imágenes. Cada vez hay más evidencia que sugiere que estos enfoques se pueden utilizar para mejorar la caracterización tumoral no invasiva, incluida la predicción de ciertas características moleculares del tumor, asociación con la diseminación tumoral, mejor predicción de la respuesta al tratamiento y pronóstico (Forghani et al., 2019).

El proceso de radiómica se puede dividir en las siguientes etapas: Adquisición y reconstrucción de imagénes, segmentación, extracción de variables o características, selección de variables, construcción de un modelo, y finalmente la predicción y prognosis.

2.4 Adquisición y reconstrucción de imágenes.

Las modalidades de imágenes médicas que se puede utilizar en este proceso son variadas y pueden obtenerse a partir de diferentes modalidades, pero principalmente son: la tomografía computada,

y la resonancia nuclear magnética.

La adquisición en la tomografía computada consiste en un tubo de rayos X que gira alrededor del paciente y que emite un haz de radiación, el cual lo atraviesa, que a continuación emerge a través del paciente, asi llegando a un detector el haz atenuado. La información recogida en los detectores es analizada por un ordenador, que reconstruye la imagen para poder visualizarla en un monitor. Cada uno de estos cortes tomográficos, son secciones pequeñas del paciente, compuestos de un número determinado de elementos volumétricos. La imagen que se visualiza se encuentra representada por píxels, en caso de una imagen bidimensional, y por voxels en caso de una imagen tridimensional (Hernández Muñiz & Mitjavila Casanovas, 2006).

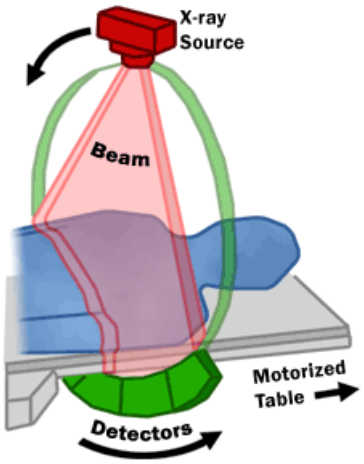

Figura 2.4: Proceso de adquisición de imagen en tomografía computada (food and drug administration)

Las diferentes densidades representaba en la imagen, tienen realación con el grado de atenuación del haz al atravesar las estructuras. Para esta representación se utilizan las unidades Hounsfield (UH). A estas unidades se les asigna un valor arbitrario, siendo 0 la densidad del agua, la grasa posee valores negativos (–70 o –90 UH) y los tejidos blandos valores positivos (+ 30, + 70 UH). Existen dos valores extremos en la escala, la densidad de hueso (+ 500 UH) y del aire (–1.000 UH) (Hernández Muñiz & Mitjavila Casanovas, 2006).

En la tomografía computada, la reconstrucción de la imagen puede llevarse a cabo por medio de dos tipos de reconstrucciones: Retroproyección filtrada, y la reconstrucción iterativa. La retroproyección filtrada consiste en un algoritmo que retroproyecta las diferentes proyecciones a través de un espacio de imagen, sumando las contribuciones de todas las proyecciones en cada punto del espacio de la imagen. También se utilizan filtros que corrigen las distorsiones y mejora la calidad de la imagen. Por otra parte, los métodos iterativos comienzan con una estimación Los algoritmos iterativos comienzan por proponer un objeto, que puede ser uno con distribución homogénea de atenuaciones. A continuación, se calculan las proyecciones, correspondientes al objeto propuesto y se comparan con las proyecciones originales medidas en el detector, este objeto propuesto se actualiza, con base en la diferencia de las proyecciones, hasta converger a un valor óptimo. Los métodos iterativos pueden ofrecer una calidad de imagen mejorada y una reducción de ruido en comparación con FBP, especialmente en situaciones de baja dosis de radiación (Ramírez Giraldo

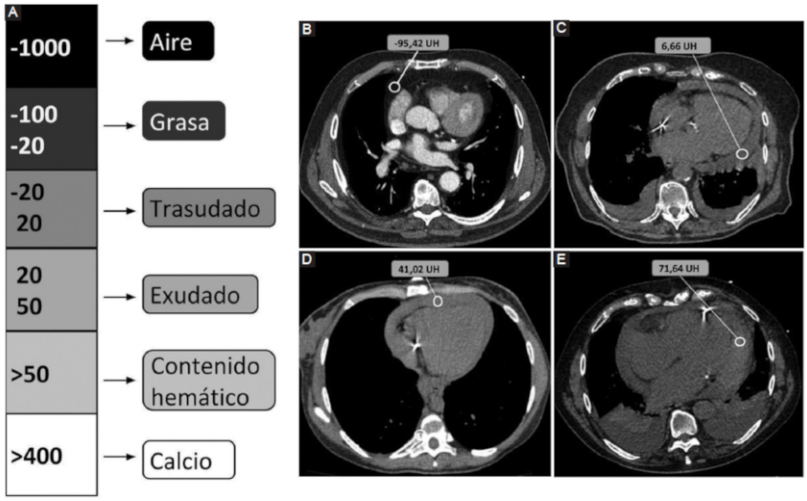

Figura 2.5: Densidades tomográgicas según sus Unidades Hounsfield (Saidman et al., 2022)

et al., 2008).

En el contexto de la resonancia magnética, el procedimiento consiste situar al paciente dentro de un imán de gran tamaño que genera un potente y uniforme campo magnético. Esto provoca que los núcleos de numerosos átomos en el cuerpo, siendo el hidrógeno el más útil, se alineen con dicho campo magnético conforme a su momento magnético. Posteriormente, se aplica una señal de radiofrecuencia, instante en el que se induce la resonancia (Lévano Loayza et al., 2020).

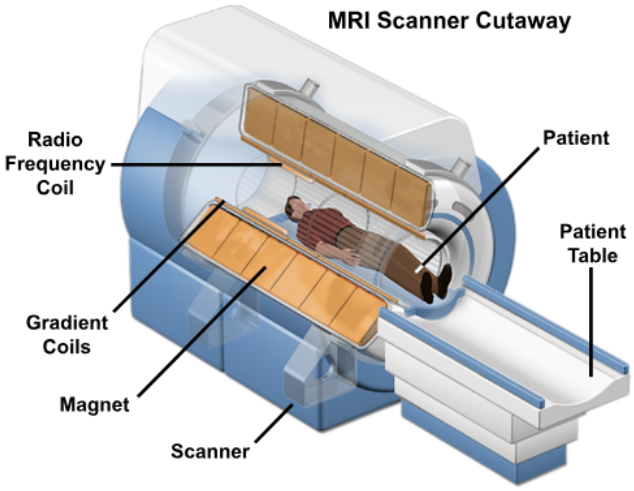

Figura 2.6: Paciente dentro de un equipo de resonancia magnética (nationalmaglab)

Durante la adquisición de la imagen, los gradientes de campo magnético se utilizan para codificar espacialmente las señales emitidas por los núcleos atómicos. Estos gradientes cambian la frecuencia de resonancia de los núcleos en función de su ubicación, lo que permite que las señales de diferentes ubicaciones se distingan entre sí. La secuencia en que estos gradientes se aplican y las señales que se adquieren son conocidas como secuencias de pulsos, y estas se utilizan para generar diferentes tipos de contraste en las imágenes de resonancia magnética (Katti et al., 2011).

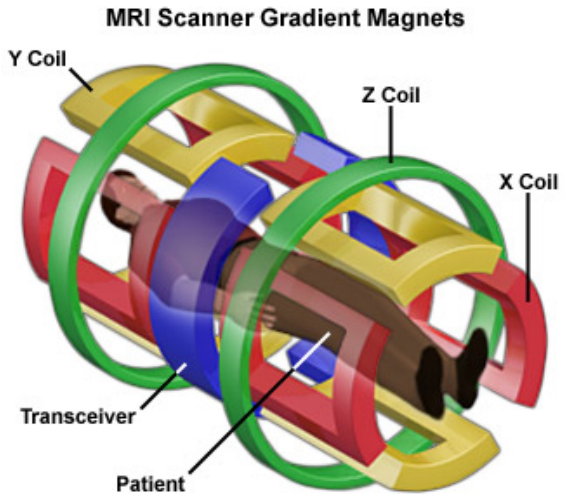

Figura 2.7: Disposición de las gradientes en un equipo de resonancia magnética (Filos, 2008))

Seguidamente, los núcleos atómicos regresan a su estado de equilibrio mediante un proceso conocido como relajación, y al hacerlo, emiten energía que es captada por una bobina receptora. Las señales recibidas están en el dominio de la frecuencia y se almacenan en un espacio conocido como espacio k, que es esencialmente un gráfico de datos en dicho dominio de frecuencia. Una vez que las señales son leídas, se aplica una transformación matemática denominada transformada

de Fourier, utilizada para convertir las señales del dominio de la frecuencia (espacio k) al dominio espacial (Katti et al., 2011).

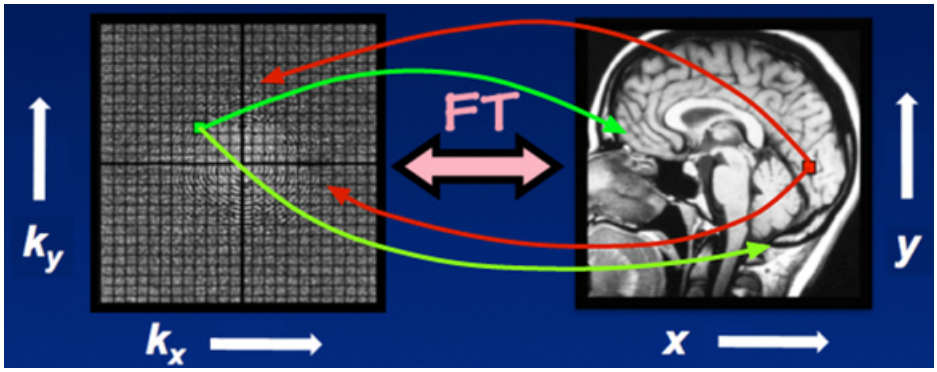

Figura 2.8: Representación del espacio k y dominio espacial (imagen) (mriquestions)

La calidad y la resolución de la imagen reconstruida pueden ser influenciadas por varios parámetros, como la matriz de adquisición, el campo de visión, y el grosor de corto. Además, diversas técnicas y algoritmos de reconstrucción pueden ser empleados para mejorar la calidad de la imagen o para reducir el tiempo de adquisición (Costa & Jerez, 2021).

Finalmente, la imagen se representa en tonalidades de grises en una matriz, por ejemplo, una matriz de 256x256. Las intensidades de la imagen representan las diferentes densidades de los tejidos en el cuerpo y están relacionadas con la cantidad de núcleos de hidrógeno y las propiedades de relajación en cada ubicación, produciendo los diversos contrastes visualizados en la imagen, para ser interpretadas posteriormente (Costa & Jerez, 2021).

2.4.1 Segmentación de imágenes.

El primer paso luego de adquirir las imágenes, es demarcar la región de interés (ROI), en el caso de imágenes bidimensionales (2D), o limitar volumen de interés (VOI) en el caso de imágenes tridimensionales (3D). Los ROI/VOI definen la región en la que se calculan las características radiómicas (van Timmeren et al., 2020).

Existen diversos métodos de segmentación, entre los cuales se incluye la segmentación manual llevada a cabo por personal experto. No obstante, este enfoque presenta un desafío en términos de variabilidad, ya que depende del operador y puede haber diferencias significativas entre cada experto. Además, se emplean métodos semiautomáticos y automáticos, donde los modelos de machine learning y deep learning desempeñan un papel fundamental. Estos últimos enfoques permiten una segmentación más precisa y consistente, aunque es dependiente de la eficacia del algoritmo.

2.4.2 Extracción de variables o características.

La librería o extensión pyradiomics (van Griethuysen et al., 2017) permite extraer las siguientes clases de variables con sus respectivas subdivisiones de features: De primer orden estadístico, de

forma y de textura.

Las estadísticas de primer orden (first order), se utilizan para describir la distribución de los pixels/voxels dentro de la segmentación, siendo:

Sea X un conjunto de de N_p incluidos en el ROI $P_(i))$, es el histograma de primer orden con $N_(g)$ niveles de intensidad discretos, donde $N_(g)$ es el número de contenedores distintos de cero. $p_(i))$ sea el histograma de primer orden normalizado e igual a $\frac{P_{(i)}}{N_p}$

Energy.

La energía es una medida de la magnitud de los valores de vóxel en una imagen. Un valor mayor implica una mayor suma de los cuadrados de estos valores.

$$energy = \sum_{i=1}^{N_p} (X(i) + c)^2 \tag{2.1}$$

Donde c es un valor opcional, que cambia las intensidades para evitar valores negativos en X. Esto asegura que los vóxeles con los valores de gris más bajos contribuyan menos a la energía, en lugar de los vóxeles con una intensidad de nivel de gris más cercana a 0.

Total Energy

$$TotalEnergy = V_{voxel} \sum_{i=1}^{N_p} (X(i) + c)^2 \tag{2.2}$$

La energía total es el valor de Energy Feature escalado por el volumen del voxel en mm^3.

Entropy

La entropía especifica la aleatoriedad en los valores de la imagen, midiendo la cantidad promedio de información requerida para codificar los valores de la imagen.

$$entropy = -\sum_{i=1}^{N_p} p(i) \log(X(i) + c)^2 \tag{2.3}$$

ϵ es un número positivo pequeño ($\approx 2.2 * 10^{-16}$)

Minimum

La intensidad mínima del nivel de gris dentro del ROI.

$$minimum = min(X) \tag{2.4}$$

10th percentile

El percentil 10 de X

90th percentile

El percentil 90 de X

Maximum

Es la intensidad máxima del nivel de gris dentro del ROI.

$$maximum = max(X) \tag{2.5}$$

Mean

Es la intensidad promedio del nivel de gris dentro del ROI.

$$mean = \frac{1}{N_p} \sum_{i=1}^{N_p} X(i)) \tag{2.6}$$

Median

La mediana de la intensidad del nivel de gris dentro del ROI.

Interquartile Range

$$interquantilerange = P_{75} - P_{25} \tag{2.7}$$

Siendo P_{25} y P_{75} son los percentiles 25 y 75 de la matriz de imágenes.

Range

Es el rango de valores de grises en el ROI

$$range = max(X) - min(X) \tag{2.8}$$

Mean Absolute Deviation (MAD)

Mean Absolute Deviation, es la distancia media de todos los valores de intensidad desde el valor medio de la matriz de imágenes.

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_{10-90}} \left| X(i) - \bar{X}_{10-90} \right| \tag{2.9}$$

Robust Mean Absolute Deviation (rMAD)

La Robust Mean Absolute Deviation es la distancia media de todos los valores de intensidad desde el valor medio calculado en el subconjunto de la matriz de imágenes con niveles de gris intermedios, o igual al percentil 10 y 90.

$$rMAD = \frac{1}{N_{10-90}} \sum_{i=1}^{N_{10-90}} \left| X_{10-90}(i) - \bar{X}_{10-90} \right| \tag{2.10}$$

Root Mean Squared (RMS)

RMS es la raíz cuadrada de la media de todos los valores de intensidad al cuadrado.

$$RMS = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - c)^2} \tag{2.11}$$

Standard Deviation

La desviación estándar mide la cantidad de variación o dispersión del valor medio.

$$RMS = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^2} \tag{2.12}$$

Skewness

Skewness mide la asimetría de la distribución de valores sobre el valor medio. Dependiendo de dónde se alarga la cola y se concentra la masa de la distribución, este valor puede ser positivo o negativo

$$Skewness = \frac{\sigma_3}{\mu_3} = \frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^3}{\sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^2}^3} \tag{2.13}$$

Kurtosis

La kurtosis es una medida de la "concentración" de la distribución de valores en el ROI de la imagen. Una curtosis más alta implica que la masa de la distribución se concentra hacia la(s) cola(s) en lugar de hacia la media. Una curtosis más baja implica lo contrario: que la masa de la distribución se concentra hacia un pico cerca del valor medio.

$$kurtosis = \frac{\mu_4}{\sigma_4} = \frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^4}{\sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^2}^2} \tag{2.14}$$

Variance

La varianza es la media de las distancias al cuadrado de cada valor de intensidad desde el valor medio. Esta es una medida de la dispersión de la distribución sobre la media.

$$variance = \frac{1}{N_p} \sum_{i=1}^{N_p} (X(i) - \bar{X})^2 \tag{2.15}$$

Uniformity

La uniformidad es una medida de la suma de los cuadrados de cada valor de intensidad. Esta es una medida de la homogeneidad de la matriz de imágenes, donde una mayor uniformidad implica una mayor homogeneidad o un menor rango de valores discretos de intensidad.

$$uniformity = \sum_{i=1}^{N_g} p(i)^2 \tag{2.16}$$

Además, se pueden extraer características del tipo de forma y de textura. Las característica de forma o morfológicas (shape features), reflejan características físicas en términos de forma y volumen, e incluyen descriptores bidimensionales o tridimensionales, por ejemplo como diámetro, volumen, área superficial, relación superficie-volumen, esfericidad, compacidad, convexidad, entre otras que se resumen en el cuadro 4.1.

Junto a las características de primer orden y las de forma, se pueden encontrar las características de segundo orden o relacionadas a la textura, las cuales entregan información sobre la interrelación entre vóxeles. Estas características de textura que describen variaciones en escala de grises, esto quiere decir, la distribución de intensidad entre vóxeles vecinos. Las características de segundo orden se extraen mediante la aplicación de diferentes técnicas, como la matriz de co-ocurrencia de nivel de gris (Gray Level Co-occurrence Matrix(GLCM)), siendo este el mas común, o la matriz de longitud de ejecución de nivel de gris (Gray Level Run Length Matrix (GLLRM)), que corresponde a la longitud de vóxeles consecutivos que tienen la misma intensidad en una dirección preestablecida en el ROI. También encuentra el método de la matriz de diferencia de tono de gris del vecindario (Neighbouring Gray Tone Difference Matrix(NGTDM)), que es una suma de las diferencias entre todos los píxeles con tono de gris y el valor medio de su vecino circundante (Mayerhoefer et al., 2020). Las características que se pueden extraer de los métodos mencionados anteriormente, se encuentran en el cuadro 2.1

Table 2.1: Clase y features.

Clase	**Features**
First Order	Energy, Total Energy, Entropy, Minimum, 10th percentile, 90th percentile, Maximum, Mean, Median, Interquartile Range, Range, Mean Absolute Deviation (MAD), Robust Mean Absolute Deviation (rMAD), Root Mean Squared (RMS), Standard Deviation, Skewness, Kurtosis, Variance, Uniformity
Shape Features (3D)	Mesh Volume, Voxel Volume, Surface Area, Surface Area to Volume ratio, Sphericity, Compactness 1, Compactness 2, Spherical Disproportion, Maximum 3D diameter, Maximum 2D diameter (Slice), Maximum 2D diameter (Column), Maximum 2D diameter (Row), Major Axis Length, Minor Axis Length, Least Axis Length, Elongation, Flatness,
Shape Features (2D)	Mesh Surface, Pixel Surface, Perimeter, Perimeter to Surface ratio, Sphericity, Spherical Disproportion, Maximum 2D diameter, Major Axis Length, Minor Axis Length, Elongation
Gray Level Co-occurrence Matrix (GLCM) Features	Autocorrelation, Joint Average, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Joint Energy, Joint Entropy, Informational Measure of Correlation (IMC) 1, Informational Measure of Correlation (IMC) 2 Inverse Difference Moment (IDM), Maximal Correlation Coefficient (MCC), Inverse Difference Moment Normalized (IDMN), Inverse Difference (ID), Inverse Difference Normalized (IDN), Inverse Variance, Maximum Probability, Sum Average, Sum Entropy, Sum of Squares

Clase	Features
Gray Level Co-occurrence Matrix (GLCM) Features	Small Area Emphasis (SAE), Large Area Emphasis (LAE), Gray Level Non-Uniformity (GLN), Gray Level Non-Uniformity Normalized (GLNN), Size-Zone Non-Uniformity (SZN), Size-Zone Non-Uniformity Normalized (SZNN), Zone Percentage (ZP), Gray Level Variance (GLV), Zone Variance (ZV), Zone Entropy (ZE), Low Gray Level Zone Emphasis (LGLZE), High Gray Level Zone Emphasis (HGLZE), Small Area Low Gray Level Emphasis (SALGLE), Small Area High Gray Level Emphasis (SAHGLE), Large Area Low Gray Level Emphasis (LALGLE), Large Area High Gray Level Emphasis (LAHGLE)
Gray Level Run Length Matrix (GLRLM) Features	Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray Level Non-Uniformity (GLN), Gray Level Non-Uniformity Normalized (GLNN), Run Length Non-Uniformity (RLN), Run Length Non-Uniformity Normalized (RLNN), Run Percentage (RP), Gray Level Variance (GLV), Run Variance (RV), Run Entropy (RE) , Low Gray Level Run Emphasis (LGLRE), High Gray Level Run Emphasis (HGLRE), Short Run Low Gray Level Emphasis (SRLGLE) , Short Run High Gray Level Emphasis (SRHGLE), Long Run Low Gray, Level Emphasis (LRLGLE), Long Run High Gray Level Emphasis (LRHGLE)
Neighbouring Gray Tone Difference Matrix (NGTDM) Features	Coarseness, Contrast, Busyness, Complexity, Strength,
Gray Level Dependence Matrix (GLDM) Features	Small Dependence Emphasis (SDE), Large Dependence Emphasis (LDE), Gray Level Non-Uniformity (GLN), Dependence Non-Uniformity (DN), Dependence Non-Uniformity Normalized (DNN), Gray Level Variance (GLV), Dependence Variance (DV), Dependence Entropy (DE), Low Gray Level Emphasis (LGLE), High Gray Level Emphasis (HGLE), Small Dependence Low Gray Level Emphasis (SDLGLE), Small Dependence High Gray Level Emphasis (SDHGLE), Large Dependence Low Gray Level Emphasis (LDLGLE), Large Dependence High Gray Level Emphasis (LDHGLE)

2.5 Selección de variables.

Aunque la radiómica extrae una amplia variedad de características, es importante tener en cuenta que algunas de estas características pueden resultar redundantes para el objetivo deseado. Si todas las características radiómicas se incluyen en el modelo, existe el riesgo de sobreajuste, lo que puede conducir a un modelo con una capacidad de generalización limitada (Lee et al., 2020).

Con la intención de eliminar las características redundantes y disminuir el sobreajuste, existen técnicas de selección características, que pueden ser clasificadas en tres métodos: métodos de filtro (filter), envoltura (wrapper) e incrustación (embedding)(Avanzo et al., 2020). Los métodos de filtrado se basan en correlaciones estadísticas, como la prueba de chi-cuadrado, para determinar qué características son relevantes. Por otro lado, los métodos de envoltura utilizan algoritmos de aprendizaje automático, como el método de eliminación de características recursivas, para encontrar la combinación óptima de características que maximice la precisión del modelo. Por último, los métodos de incrustación se pueden implementar en conjunto con la selección de características y la creación de modelos, como el operador de selección y contracción mínima absoluta (Least Absolute Shrinkage and Selection Operator (LASSO))(Satake et al., 2022). En la sección de metodología, se explicaran los métodos utilizados en este estudio.

2.6 Construcción de un modelo.

Una vez que se han aplicado los métodos de selección de características, y se han obtenido las características informativas y no redundantes, el siguiente paso consiste en construir un modelo para abordar el objetivo de la tarea. En el caso de un problema de clasificación, se pueden utilizar diversos modelos, como la regresión logística, random forest, support vector machine (SVM) y redes neuronales, entre otros (Satake et al., 2022). Estos modelos son ampliamente utilizados debido a su capacidad para aprender patrones y relaciones en los datos, permitiendo así realizar predicciones y clasificaciones precisas.

2.7 Predicción y prognosis.

El objetivo final de la radiómica es construir modelos predictivos para el resultado del tratamiento y para la evaluación del riesgo, basados en características fenotípicas del tumor calculadas a partir de imágenes médicas (Larue et al., 2017).

2.8 Aprendizaje Automático (Machine Learning)

Después de seleccionar las características radiómicas, se aplicará un modelo con la intención de predecir una patología, o cierto evento. La aplicación de un modelo de machine learning es el enfoque más común para la predicción basada en radiómica, y a su vez, este subconjunto de la investigación de inteligencia artificial (IA) ha crecido rápidamente, junto con una mayor disponibilidad de datos (Avery et al., 2022).

Un cálculo de machine learning es un algoritmo el cual puede aprender a partir de un conjunto entregado, se dice que "un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y medida de desempeño P, si su desempeño en las tareas en T, medido

por P, mejora con la experiencia E" (Mitchell, 1997), 1997).

Un modelo de machine learning puede definirse como algoritmos que construyen clasificadores basados en el análisis de datos de entrenamiento, pudiendo inferir en una hipótesis (o función) y predecir las etiquetas de observaciones no vistas, en el caso de este estudio el resultado del paciente o fenotipo del tumor (Avanzo et al., 2020).

Los modelos de machine learning se pueden dividir en 2 grandes tipos: Aprendizaje supervisado y aprendizaje no supervisado. El aprendizaje supervisado es el enfoque más frecuente en diferentes áreas de aplicación de la IA, incluida la neurorradiología (Wagner et al., 2021). El aprendizaje supervisado se basa en conjuntos de datos etiquetados, la cual puede ser una clase, una puntuación continua o incluso una imagen completa, en cambio el aprendizaje no supervisado implica el reconocimiento de patrones sin la participación de un atributo objetivo, es decir, todas las variables utilizadas en el análisis se utilizan como entradas y debido al enfoque, las técnicas son adecuadas para las técnicas de agrupamiento y minería de asociación (Alloghani et al., 2020).

Para la evaluación de los modelos de machine learning, se usan los índices o métricas de desempeño, para caso de los modelos de clasificación, se consideran la sensibilidad, la especificidad, la precisión y el área bajo la curva característica operativa del receptor (ROC-AUC). Cuando el objetivo es la evaluación de la capacidad de supervivencia, se utiliza comúnmente el índice de concordancia de Harrell para evaluar la capacidad de discriminación del modelo. También se utilizan las curvas de Kaplan-Meier, para estimar la probabilidad de supervivencia a partir de datos de tiempo de vida, o se pueden utilizar para comparar la capacidad de supervivencia entre diferentes grupos de riesgo (Satake et al., 2022).

La precisión de la predicción de los modelos basados en radiómica se evalúa más comúnmente por la ROC-AUC. Este enfoque incorpora predicciones verdaderas positivas, verdaderas negativas, falsas positivas y falsas negativas realizadas por el modelo de radiómica. También es importante considerar un análisis del rendimiento del modelo con una validación independiente, por lo cual el modelo basado en radiómica se aplica a conjuntos de datos en los que no se entrenó. Esto es fundamental para demostrar la generalización, la solidez y el potencial de un modelo para una utilidad clínica más amplia (Avery et al., 2022). A continuación, se presentarán algunos de los modelos de machine learning más utilizados.

2.9 Logistic Regression

La regresión logística no es una regresión, sino un algoritmo de aprendizaje de clasificación. Este nombre proviene de la estadística, esto se debe a que la formulación matemática de la regresión logística es similar a la de la regresión lineal (Burkov, 2019).

Las regresiones logísticas modelan una relación entre la variable de respuesta categórica y las covariables, en este caso se dice que existe una combinación lineal de variables independientes con probabilidades logarítmicas de la probabilidad de un evento. Las regresiones logísticas binarias estiman la probabilidad de que una característica de una variable binaria esté presente, dados los valores de las covariables. Supongamos que Y es una variable de respuesta binaria donde $Y_i = 1$ si la variable objetivo esta presente, y en caso contrario $Y_i = 0$, siendo los datos $(Y_1, Y_2, ..., Y_n)$

independientes. Considerar p como probabilidad de éxito. Además, considere $X = (x_1, x_2, ..., x_n)$ como un conjunto de variables explicativas que pueden ser discretas, continuas o una combinación de ambas discretas y continuas (Joshi & Dhakal, 2021). Según lo anterior, se elabora una función basada en el cálculo de la probabilidad de que la variable de interés toma el valor del evento anteriormente definido, de la manera siguiente (Alba Vega & Calle Jara, 2020):

$$p(x) = \frac{1}{1 + e^{-Z(x)}} \tag{2.17}$$

$$Z(x) = \beta_0 + \beta_1 x + \beta_2 x... \tag{2.18}$$

$$logit(p(x)) = ln(\frac{p(x)}{1 - p(x)}) = Z(x) \tag{2.19}$$

Donde,
x = Variable predictora
$p(x)$ = Probabilidad de la variable de interés o variable predictora
$Z(x) = \beta_0 + \beta_1 x + \beta_2 x...$: Son los coeficientes de las regresiones logísticas aplicadas

Ya que la nueva variable independiente que se desea estimar puede tomar cualquier valor, y no está sujeta a un rango de valores, se puede recurrir a los métodos de estimación de los modelos de regresión tradicionales para poder así elaborar el modelo de regresión logística. Tras realizar una serie de transformaciones se puede deducir que (Alba Vega & Calle Jara, 2020):

$$p(x) = \frac{1}{1 + e^{-(\beta_1 + \beta_1 x)}} \tag{2.20}$$

$$p(x) = \frac{e^{(\beta_1 + \beta_1 x)}}{1 + e^{-(\beta_1 + \beta_1 x)}} \tag{2.21}$$

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} \tag{2.22}$$

Donde,
$p(x)/(-p(x))$: Son los odds (posibilidades) a favor del éxito

2.10 Random Forest

Un algoritmo de random forest es una combinación de árboles predictores en el cual cada árbol depende de los valores del vector aleatorio muestreado de manera independiente y con la misma distribución para todos los árboles (Breiman, 2001).

Cada árbol de este modelo se obtiene del siguiente proceso (Espinosa-Zúñiga, 2020): se genera un numero de arboles de decisión a partir del conjunto de datos entregado, cada uno de estos árboles está compuesto por un subconjunto aleatorio de variables n, de forma que n < N, donde N es el número total de variables, luego cada árbol crece hasta su máxima extensión. Cada uno de estos árboles por medio de bootstrap elige un grupo de observaciones aleatorias, y las observaciones que no fueron incluidas se utilizan para la validación del modelo. Finalmente, con todas las salidas de los árboles se combinan en una salida Y.

2.11 CatBoost

El algoritmo de Catboost es utilizado en problemas de clasificación, es una implementación del gradient boosting, pero que hace uso de árboles de binarios como predictores (Prokhorenkova et al., 2018). En este problema de clasificación binario, el objetivo del aprendizaje es entrenar una función $H : \mathbb{R}^n \to \mathbb{R}$ que minimice la pérdida.

$$L(H) := EL(y, H(X)) \tag{2.23}$$

Donde $L(.,.)$ es una pérdida de función suavizada y (X, y) es una función de prueba de los datos del entrenamiento D.

El procedimiento para el gradient boosting (Friedman, 2001) es construido iterativamente a una secuencia de aproximaciones $H^t : mathbbR^m \to \mathbb{R}, t = 0, 1, 2...$ en un algoritmo voraz. Desde la aproximación previa H^{t-1}, H^t es obtenido en un proceso aditivo, como $H^t = H^{t-1} + \alpha g^t$, con un tamaño de paso α y función $g^t : \mathbb{R}^n \to \mathbb{R}$, el cual es un predictor base, y se selecciona a partir de un conjunto de funciones G para minimizar la pérdida:

$$g^t = \arg min_{g \epsilon G} L(H^{t-1} + g) \tag{2.24}$$

$$= \arg min_{g \epsilon G} L(y, H^{t-1}(X) + g(X)) \tag{2.25}$$

El problema de minimización se resuelve mediante el método de Newton, utilizando un segundo orden de aproximación de $L(H^{t-1} + g^t)$ a H^{t-1} (Ibrahim et al., 2020).

2.12 Gradient Boost

Suponiendo que tenemos M árboles de decisión, la predicción del conjunto de árboles se modela con una función sigmoide (Burkov, 2019):

$$PR(y = 1|x, f) = \frac{1}{1 + e^{-f(x)}} \tag{2.26}$$

donde $f(x) = \sum_{m=1}^{M} f_m(x)$ y f_m es un árbol de decisión

El algoritmo de gradient boosting comienza con un modelo inicial $f = f_0 = 1\frac{p}{1-p}$, donde $p = \frac{1}{N} \sum_{1=i}^{N} y_i$. Luego de cada iteración m, se agrega un nuevo nuevo árbol f_m al modelo.

A continuación se reemplaza y_i, con la derivada parcial de g_i, y se construye un nuevo árbol f_m. Entonces la actualización óptima de ρ_m es:

$$\rho_m = \arg\max L_{f+\rho f_m} \tag{2.27}$$

En la última iteración m, se actualiza el modelo con f agregando f_m:

$$f \leftarrow f + \alpha\rho_m f_m \tag{2.28}$$

Esta iteración se realiza hasta $m = M$, entonces se detiene y vuelve al modelo ensamblado f (Burkov, 2019).

2.13 XGBoost

Xgboost es una variante de un modelo de gradient tree boosting propuesta por Tianqi Chen en 2016 (Chen & Guestrin, 2016). El modelo de xgboost esta dado por la siguiente ecuación:

$$L^{(t)} = \sum l(y_i \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.29}$$

Donde l es la función de pérdida que mide la diferencia entre la predicción $\hat{(}y_i)$ y el objetivo y_i y Ω es el término de regularización. También $f_t(x_i)$ es un árbol que mejora el modelo actual propuesto en t-1 pasos.

Las principales características de este modelo es que es un ensamblado secuencial de árboles de decisión. La idea principal es que se agreguen árboles secuencialmente para que, a partir de los resultados, los árboles nuevos aprendan de los previos, para así corregir el error y poder reducirlo al mínimo. La característica que diferencia al modelo random forest es que al primero el usuario puede indicar la cantidad de árboles mientras que en el caso del xgboost crecen hasta un punto máximo (Espinosa-Zúñiga, 2020).

2.14 Support Vector Machine (SVM).

Este algoritmo consiste en modelos lineales para representar una frontera de decisión no lineal en el espacio origen. A este modelo se le entregan un conjunto de datos, los cuales se entrenarán para construir un clasificador binario no probabilístico. Esto se hace por medio de que a la muestra de datos se le determinará un hiperplano para que sean ordenados (Nieto, 2016).

Los modelos de support vector machine requieren la solución del siguiente problema de optimización (Cortes & Vapnik, 1995; Giovanetti, 2021):

$$\underset{\omega,b,\epsilon}{min} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{1}\epsilon_i^2 \tag{2.30}$$

Sujeto a:

$$yi(\omega^T \varnothing(x_i) + b) \geqslant 1 - \epsilon_i, \epsilon_i \geq 0 \tag{2.31}$$

Con esto el modelo de SVM encuentra un hiperplano de separación lineal con el margen máximo. C>0 es el parámetro de penalización del término del error. También, la expresión $K(x_i, x_j) \equiv \phi(x_i)^T(x_j)$ es el kernel de la función. Existen los siguientes kernel más comunes (Awad & Khanna, 2015; Giovanetti, 2021)

Lineal $K(x_i, x^j) = x_i^T x_j$
Polinomial $K(x_i, x^j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
Función de base radial $K(x_i, x^j) = e^{(-\gamma \|x_i - x_j\|^2), \gamma > 0}$
Sigmoide$K(x_i, x^j) = \tanh(\gamma x_i^T x_j + r)$

Los valores γ, d, y r son los parámetros del kernel.

2.15 t-Distributed Stochastic Neighbor Embedding (t-SNE).

El método de t-Distributed Stochastic Neighbor Embedding (t-SNE), es un algoritmo de machine learning no supervisado, que se utiliza para la exploración y visualización de datos, proporcionando la intuición de cómo se organizan los datos en el espacio de alta dimensión (Maaten & Hinton, 2008).

Este algoritmo comienza aplicando Stochastic Neighbor Embedding (SNE) al conjunto de datos, que convierte las distancias euclidianas de alta dimensión en probabilidades condicionales que representan similitudes para cada par de datos. La similitud de los datos X_a con los datos X_b está representada por la probabilidad condicional $P_{a|b}$, definida en la siguiente ecuación (Anowar et al., 2021):

$$P_{a|b} = \frac{exp^{\frac{-\|x_b - x_a\|^2}{2\sigma^2}}}{\sum_{a \neq k} \frac{-\|x_b - x_a\|^2}{2\sigma^2}} \tag{2.32}$$

La ecuación mide qué tan cerca está un punto de datos X_a de otro punto de datos X_b, en la se considera una distribución gaussiana alrededor de X_b con una varianza dada σ^2 . Esta varianza difiere para cada dato y se elige de manera que los datos en áreas densas tengan una varianza menor que los datos en áreas dispersas. Posteriormente, en lugar de utilizar la distribución gaussiana, se utiliza una "distribución t de Student" con un grado de libertad, similar a la distribución de Cauchy, para obtener el segundo conjunto de probabilidades ($Q_{a|b}$) en el espacio de baja dimensión. Si los datos de alta dimensión X_a y X_b se asignan correctamente a los datos de baja dimensión y_a e y_a, entonces la similitud entre $P_{a|b}$ y $Q_{a|b}$ se vuelve igual. Por lo tanto, t-SNE minimiza la diferencia entre estas dos probabilidades desde los espacios de baja dimensión a los de alta dimensión. Esta diferencia se mide optimizando la función de costo (ϕ) de la suma de la divergencia de Kullback-Leibler como se muestra a continuación (Maaten & Hinton, 2008):

$$\phi = \sum_a \sum_b P_{a|b} log \frac{P_{a|b}}{Q_{a|b}} \tag{2.33}$$

2.16 Interpretabilidad y explicabilidad.

Las técnicas de ML interpretables y explicables surgen de la necesidad de diseñar sistemas de aprendizaje automático inteligibles, es decir, que puedan ser comprendidos, y poder explicar las predicciones hechas por los modelos, tales como modelos de redes neuronales o de grandient boosting (Marcinkevičs & Vogt, 2020). Las primeras investigaciones sobre el aprendizaje automático interpretable se remontan a la década de 1990 (Rudin & Radin, 2019), sin mencionar que muchos modelos estadísticos clásicos pueden considerarse interpretables. Por ejemplo existen algunos

modelos explicables en su estructura como los árboles de decisión (Tran, 2020), modelos de regresión lineal y logística (Giovanetti, 2021). Es importante referirse a que hay una diferencia entre los términos de interpretabilidad y explicabilidad, ya que se pueden tender a confundirse por su relación.

La interpretabilidad tiene relación con explicar el significado de algo o cual es el significado pretendido de algo, que tiene relación con hacer una interpretación. Además, como es una interpretación, podría estar influenciada por la experiencia previa, para lo cual, para tener una interpretación correcta es necesaria cierta cantidad de ensayos y errores con diferentes variaciones en los procesos (Tran, 2020).

La explicabilidad tiene relación con el grado en el que se puede explicar el comportamiento de un modelo, considerando las diferentes partes de un modelo, como son el resultado y el proceso que realiza (Giovanetti, 2021). Otra forma de entender esto es que se puede descubrir el significado entre las atribuciones de datos de entrada y las salidas del modelo, lo que le permite explicar la naturaleza y el comportamiento del modelo AI/ML (Amazon Web Services, 2021). En la figura 2.9, se puede observar la relación que existe entre interpretabilidad y desempeño de algunos modelos.

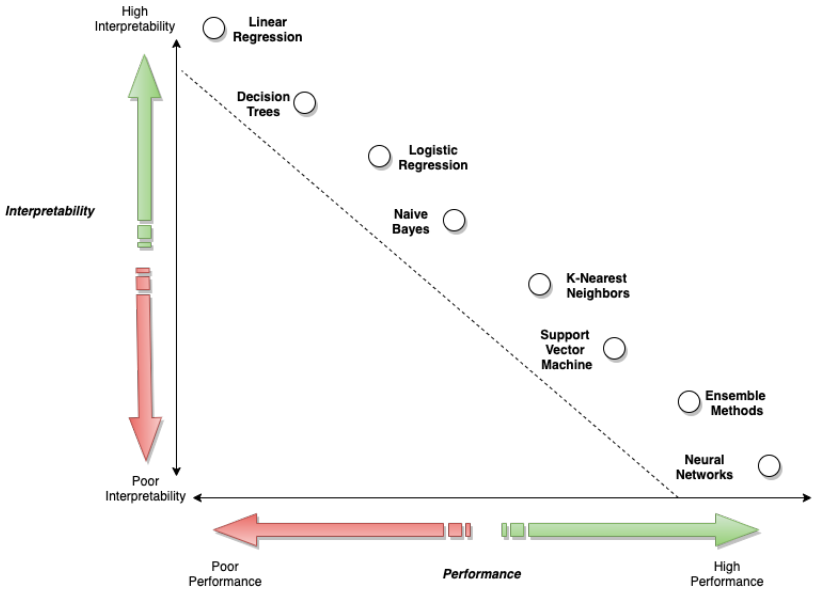

Figura 2.9: Interpretabilidad vs desempeño (Amazon Web Services)

Herzog menciona que la explicabilidad puede ser un principio ético, porque cuando se toman ciertas decisiones que pueden influir en una persona es necesaria que se entienda el modelo, o que responda a la pregunta ¿Cómo funciona?, por ejemplo, en la situación de la medicina. También sugiere que debe existir un equilibrio entre rendimiento y explicabilidad (Herzog, 2022).

La precisión de los modelos actuales de IA es notable, pero la precisión no es el único aspecto de suma importancia. Para dominios sensibles, también es importante una comprensión detallada del modelo y los resultados. Los algoritmos subyacentes de ML y DL construyen modelos complejos que son opacos para los humanos, el área de salud se encuentra entre los mayores desafíos para la IA (Holzinger et al., 2019).

Un ejemplo ilustrativo se presenta en la figura 2.10, donde se emplea el método LIME (Local Interpretable Model-Agnostic Explanations) para predecir si una persona está resfriada. Este enfoque desglosa las variables, como los síntomas de estornudos o dolor de cabeza, como relevantes para la gripe, mientras que la fatiga se considera menos relevante. A partir de esta explicación, un médico humano puede tomar una decisión informada sobre la precisión de la predicción. Es evidente que el médico, con su experiencia y conocimiento en el campo, está mejor capacitado para tomar esta decisión y puede confiar más plenamente en la predicción (Tran, 2020).

En el artículo de Burkart y Huber, llamado "A Survey on the Explainability ofSupervised Machine Learning" (Burkart & Huber, 2021), mencionan ocho razones para la explicabilidad:

- Confianza: La confianza juega un papel fundamental y es esencial en la implementación del modelo de predicción, dado que es necesario comprender tanto las fortalezas como las debilidades del mismo. Esta comprensión se convierte en un requisito previo para generar confianza en los seres humanos y, en última instancia, para llevar a cabo la implementación exitosa del modelo

- Causalidad: La causalidad puede resultar deseable cuando se busca obtener un modelo interpretable o un método de explicación, que genere hipótesis sobre relaciones causales entre variables. Por lo tanto, es frecuentemente preferible que la técnica detecte relaciones causales en lugar de simples asociaciones (Marcinkevičs, 2020).

- Transferibilidad: El modelo de predicción debe proporcionar una comprensión del comportamiento futuro para respaldar la toma de decisiones, permitiendo asi poder utilizar el modelo de predicción con datos no vistos. Solo cuando la persona que tome la decisión, sepa que el modelo generaliza bien, o mejor dicho, qué contexto generaliza bien, el modelo de predicción podrá guiar la toma de decisiones.

- Informatividad: Para que un sistema se implemente correctamente, es fundamental determinar si realmente cumple con los propósitos del mundo real para los cuales fue diseñado, en lugar de limitarse únicamente a los propósitos para los cuales fue entrenado

- Toma de decisiones justa y ética: Es fundamental en la sociedad tener acceso a los motivos detrás de una decisión tomada. Es importante destacar que los responsables de la toma de decisiones deben presentar los resultados de manera comprensible, de manera que se pueda evaluar su conformidad con los estándares éticos establecidos. Cualquier individuo afectado por una decisión automatizada debería recibir información clara y comprensible sobre la misma.

- Responsabilidad: Uno de los objetivos al incorporar la explicabilidad en el proceso de toma de decisiones es asegurar que un algoritmo sea responsable de sus acciones.

- Realización de ajustes: Comprender el modelo de predicción y los factores subyacentes permite a los expertos comparar el modelo de predicción con el conocimiento existente. La explicabilidad se convierte en un requisito previo para la capacidad de ajustar el modelo de predicción al incorporar conocimientos adicionales. Además, desde una perspectiva algorítmica, la explicabilidad brinda a los diseñadores de sistemas la capacidad de realizar cambios en el modelo de predicción, como ajustar parámetros. También resulta útil para los desarrolladores, ya que les permite identificar posibles modos de falla

- Funcionalidad de proxy: Cuando un sistema ofrece explicabilidad, se vuelve posible evaluarlo en función de criterios adicionales que no son fácilmente cuantificables, como la seguridad, la no discriminación, la privacidad, la solidez, la confiabilidad, la facilidad de uso, la equidad, la verificación y la causalidad. Por lo tanto, la explicabilidad puede servir como un proxy para analizar y evaluar estos aspectos.

En el libro Interpretable Machine Learning (Molnar, 2020), que tiene enfoque en ser una guía para la interpretabilidad de modelos que tienen "caja negra", indica que existen 3 tipos de métodos de interpretabilidad:

- Métodos globales: Métodos que describen el comportamiento global de los modelos como los Partial Dependence Plot (PDP)
- Métodos locales: Métodos que describen o explican las predicciones individuales. En este grupo se encuentran los shap values
- Métodos para Deep learning: Estos métodos están enfocados en visualizar las características y conceptos aprendidos por una red neuronal, explicando así las predicciones individuales y que simplifican la red neuronal.

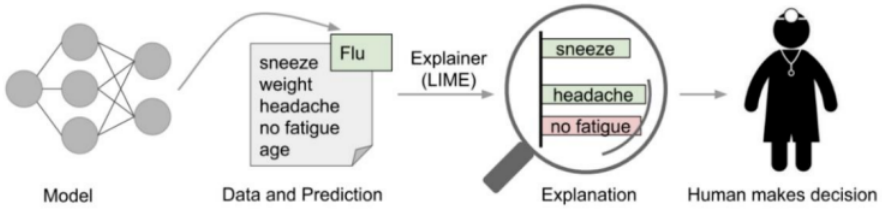

Figura 2.10: Explicación de predicción individual

2.17 Explicabilidad por medio de shap values

Los shap values están basados en los shapley values de la teoría de juegos, y específicamente en los juegos cooperativos, en como los jugadores toman sus decisiones conformando alianzas o coaliciones con el fin de obtener una recompensa, que luego será repartida según la proporcionalidad de las contribuciones de cada jugador (Shapley, 2016). En la siguiente ecuación (2.34), donde se muestra la fórmula del cálculo de los shap values de cada variable o jugador:

$$\Phi_i(v) = \sum_{S:i\epsilon S} \frac{(k-1)!(n-k!)}{n!}[v(S)] - v(S-i)] \tag{2.34}$$

Donde,
- K = tamaño de la coalición
- N = cantidad de participantes
- $V(s)$ = función característica que incluye al jugador
- $V(s-(i))$ = función característica que no incluye al jugar

Para poder utilizar esto, existe una librería llamada shap (SHapley Additive exPlanations)

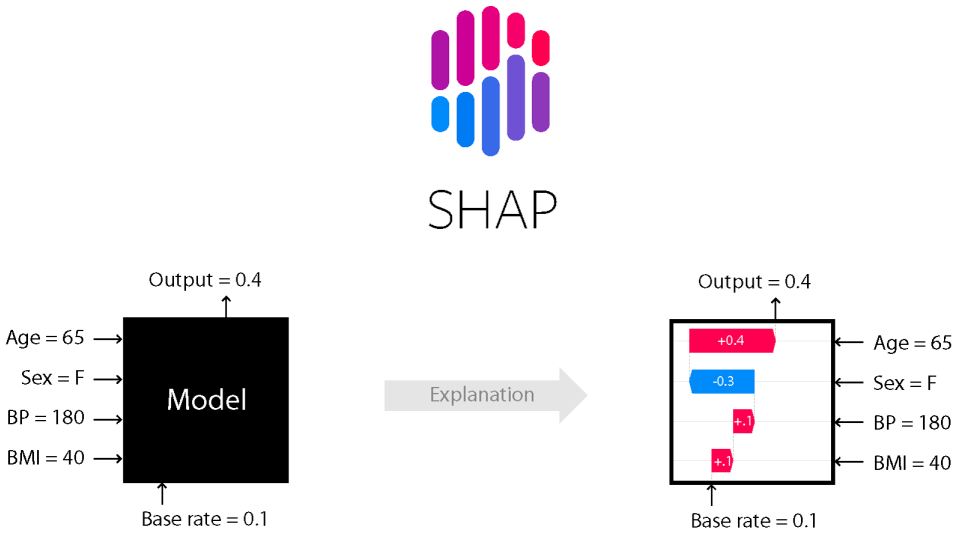

Figura 2.11: Shap values.

Chapter 3

Estado del arte

Fabian Isensee y su equipo emplearon un modelo basado en U-net para la segmentación de tumores en el desafío del año 2017. En este desafío, lograron obtener un puntaje DICE de 0,896 y 0,797 para la segmentación de todo el tumor y del núcleo tumoral, respectivamente. Además, demostraron un buen desempeño en el conjunto de prueba, con un puntaje de 0,858 para la segmentación de todo el tumor y 0,775 para el núcleo tumoral (Isensee et al., 2018).

Una investigación del año 2017 exploró diferentes arquitecturas basadas en redes convolucionales para segmentar variables grados de glioblastomas en imágenes de resonancia magnética, proponiendo finalmente una variante propia de CNN (Convolutional Neural Network), con la intención de acelerar el proceso, que logró ser 30 veces más rápida que las arquitecturas publicadas hasta esa fecha para la base de datos BRATS 2013, y además, con una buena caracterización de propiedades locales y globales (Havaei et al., 2017).

En el otro artículo utilizaron 2 modelos de deep learning, el CNN4 y el VGG-16 con 2 tipos de entrenamiento cada uno. En este estudio concluyeron que las redes neuronales convolucionales son una herramienta útil para la identificación de hemorragias intracraneales en imágenes de tomografía computarizada y se puede utilizar como un apoyo en el diagnóstico de este tipo de patologías (Castro et al., 2020).

Según Numan Saeed en su estudio sobre predicción de marcador MGMT en el cual utilizaron modelos de CNN 2D, CNN 3D y vision Transformers, concluyeron que no existe conexión entre las imágenes de resonancia magnética y la metilación del tumor, por lo cual indican que la mejor opción para conocer esto sigue siendo la biopsia del tumor (Saeed et al., 2022).

Firas Baba fue el ganador del desafío del año 2021 para la predicción del marcador MGMT con un ROC de 0,62173, en el cual aplicó un modelo ResNet, junto a un algoritmo de segmentación llamado MONAI (Molnar, 2020).

Han y Kamdar usaron una base de datos del The Cancer Imaging Archive junto a la de The Cancer Genome Atlas, para la cual utilizaron una arquitectura de convolutional recurrent neural network (CRNN) en la predicción del marcador MGMT obteniendo un accuracy de 0,62, con precisión de 0,67 y un recall de 0,67 (Han & Kamdar, 2018)

En un estudio realizado por P. Franco, propusieron como objetivo analizar y extraer múltiples

patrones de correlación de patrones hemodinámicos a partir de imágenes de resonancia magnética en pacientes sanos y en pacientes con alteración hemodinámica en válvula aortica bicúspide con dilatación y sin dilatación aortica, con el fin de encontrar posibles biomarcadores hemodinámicos usando técnicas de machine learning (Franco et al., 2022).

Pálsson para el desafío de predicción de MGMT en el cual entrenaron un modelo de deep learning basado en la red U-net para segmentación, a continuación, obtuvieron los features por medio de Radiomics para finalmente entrenar un modelo random forest logrando un AUC de 0.598 (Pálsson et al., 2022). Además, aplicarón una técnica de selección de variable, obteniendo los que se indican en la figura 3.1:

Category	Feature name	Modality	Region
Shape	Maximum 3D Diameter	-	Enh-core
First order	Interquartile Range	T1-ce	Core
First order	Mean Absolute Deviation	T1-ce	Core
First order	Mean	T1-ce	Core
First order	Median	T1-ce	Core
First order	Median	T1-ce	Whole
First order	Variance	T1-ce	Core
First order	10Percentile	FLAIR	Core
GLRLM	Graylevel non-uniformity normalized	FLAIR	Whole
GLRLM	Graylevel variance	FLAIR	Whole
GLSZM	Small area emphasis	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	FLAIR	Whole
GLSZM	Small area low graylevel emphasis	FLAIR	Whole
NGTDM	Busyness	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	T2	Whole

Figura 3.1: Features seleccionados en el estudio de Palsson

En un estudio en el que se empleó un modelo XGBoost para el entrenamiento, se fusionaron la extracción de variables mediante radiomics y la selección de características basada en el F-score. Los resultados de este estudio indicaron una mejora en el rendimiento del modelo. El conjunto de datos utilizado en esta investigación constó de 53 pacientes diagnosticados con glioblastomas y se evaluó la presencia del gen MGMT. Los valores de rendimiento obtenidos fueron notables, con una precisión (accuracy) de 0.887 y un área bajo la curva (ROC) de 0.896. Además, se identificaron y seleccionaron las siguientes características como las más relevantes para el modelo (Le et al., 2020):

No.	Feature Name	Modality	Matrix	Type
1	HISTO_ET_T2_Bin6	T2	First Order	Histogram
2	TEXTURE_GLRLM_ED_T2_GLV	T2	GLRLM	Texture
3	TEXTURE_GLSZM_NET_FLAIR_ZP	FLAIR	GLSZM	Wavelet Texture
4	TEXTURE_GLSZM_NET_FLAIR_SZE	FLAIR	GLSZM	Wavelet Texture
5	TEXTURE_GLSZM_NET_FLAIR_ZSN	FLAIR	GLSZM	Wavelet Texture
6	TEXTURE_GLSZM_NET_T1_ZSN	T1	GLSZM	Wavelet Texture
7	TEXTURE_GLSZM_NET_T1_SZE	T1	GLSZM	Wavelet Texture
8	HISTO_ED_T2_Bin5	T2	First Order	Histogram
9	TEXTURE_GLSZM_NET_T1_ZP	T1	GLSZM	Wavelet Texture

Figura 3.2: Features seleccionados en el estudio de Le

Un desafío que se debe abordar en radiómica es la variabilidad asociada con el proceso de adquisición y reconstrucción de imágenes. Algunas de las fuentes de variación a considerar incluyen (Forghani et al., 2019):

- Distintos parámetros de adquisición utilizados.
- Variaciones entre fabricantes o tipos de modelos de equipos médicos.
- Diferencias en los métodos de reconstrucción de las imágenes.

Chapter 4

Metodología

La Figura 4.1 muestra un diagrama del flujo sobre el proceso de radiómica integrado con técnicas de Aprendizaje Automático.

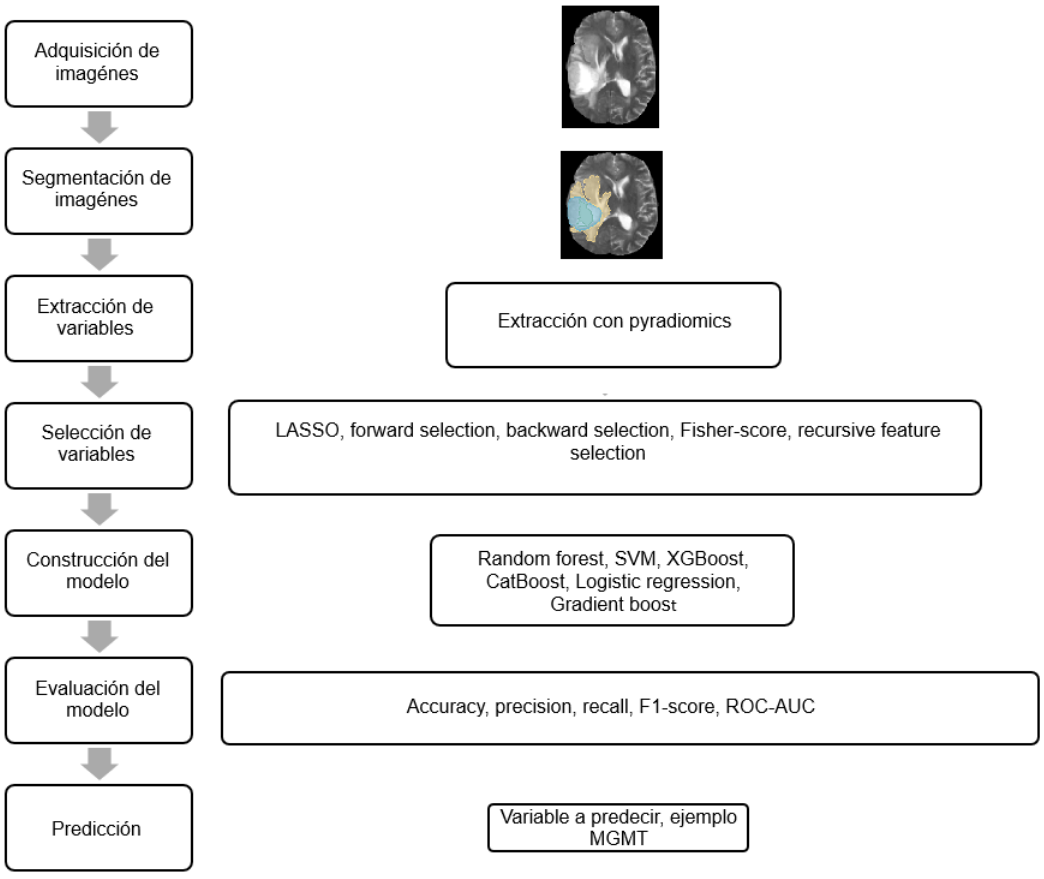

Figura 4.1: Diagrama de flujo para radiómica

4.1 Software

Para el procesamiento y análisis de la base de datos se utilizó el software Visual Studio Code, con el lenguaje de programación Python junto a la extensión de Jupyter Notebooks. También se importaron las siguientes librerías o bibliotecas:

- Pandas: Es una biblioteca de Python , utilizada para la manipulación y análisis de datos (McKinney, 2010). Proporciona estructuras de datos y herramientas para trabajar con datos tabulares o heterogéneos de manera eficiente, útil para análisis cuantitativo. Algunas de las funciones que permite realizar son: limpiar, filtrar, agrupar, y transformar datos, facilitando la preparación y exploración de los mismos para su posterior análisis o visualización.

- Numpy: Es una biblioteca de Python que proporciona soporte para arrays y matrices de gran tamaño y multidimensionales, junto con una colección de funciones matemáticas para operar con estas estructuras de datos de manera eficiente (Harris et al., 2020). NumPy permite realizar operaciones numéricas complejas y rápidas, aprovechando operaciones optimizadas a nivel de hardware.

- Sklearn: Scikit-learn, comúnmente referido como sklearn, es una biblioteca en Python que ofrece una variedad de herramientas simples y eficientes para el análisis predictivo de datos y la construcción de modelos de machine learning (Pedregosa et al., 2011). Ofrece algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad, así como flujos de trabajo de modelado con validación cruzada y búsqueda de hiperparámetros, facilitando la creación y validación de modelos machine learning con pocas líneas de código.

- Matplotlib: Es una biblioteca de Python enfocada en la visualización de datos, que permite la creación de gráficos estáticos, interactivos y animados en 2D (Hunter, 2007). Es muy personalizable y capaz de crear prácticamente cualquier gráfico o diagrama. Desde gráficos de barras y líneas hasta diagramas de dispersión y histogramas, Matplotlib es una herramienta esencial para la visualización de datos y el análisis exploratorio de datos.

- Shap: La biblioteca SHAP (SHapley Additive exPlanations) es una herramienta en Python que ofrece una manera de explicar los resultados de modelos de machine learning (Lundberg & Lee, 2017). Basada en la teoría de juegos y, en particular, en los valores de Shapley, SHAP asigna a cada característica un valor que indica su importancia en la predicción de un modelo. Esto es útil para interpretar modelos complejos y no lineales, como redes neuronales profundas o modelos ensamblados, ya que permite a los usuarios y científicos de datos entender cómo y por qué un modelo toma una decisión particular, contribuyendo a la transparencia y confiabilidad de los modelos predictivos.

- Seaborn: Seaborn es una biblioteca de visualización de datos en Python que está basada en Matplotlib, que proporciona una interfaz de alto nivel para la creación de gráficos estadísticos atractivos y detallados (Waskom, 2021). Algunos gráficos que se pueden crear son los diagramas de dispersión, de líneas, histogramas, y mapas de calor, con una mayor facilidad de uso y estilos por defecto más estéticos que Matplotlib. Seaborn está diseñado para trabajar bien en un contexto de análisis de datos, integrándose perfectamente con las estructuras de datos de Pandas.

- Scikit-image: Scikit-image, o skimage, es una biblioteca de procesamiento de imágenes para Python que está construida sobre NumPy y SciPy (Van der Walt et al., 2014). Proporciona una colección de algoritmos y herramientas para diversas tareas de procesamiento de imágenes, como

la manipulación, mejora, segmentación, detección de características y más. Esta librería tiene aplicaciones que involucran análisis de imágenes, permitiendo transformar imágenes para extraer información, realizar mejoras visuales o preparar los datos para otros tipos de análisis o procesamiento de machine learning.

4.2 Conjunto de datos

Para este estudio, se emplearon dos conjuntos de datos de imágenes preoperatorias de resonancia magnética, provistos por el desafío BraTS (Brain Tumor Segmentation Challenge) 2021 (Radiological Society of North America, 2021). Este evento es auspiciado conjuntamente por entidades del campo de la radiología y el procesamiento de imágenes médicas, como son la RSNA (Radiological Society of North America), la ASNR (American Society of Neuroradiology) y el congreso MICCAI (Medical Image Computing and Computer Assisted Intervention).

El primer conjunto de datos corresponde a la tarea 1 del desafío y está orientado hacia la evaluación de técnicas avanzadas en la segmentación de subregiones heterogéneas del glioblastoma cerebral en imágenes de MRI. De este conjunto, se han utilizado las máscaras de segmentación para identificar distintas áreas tumorales, incluyendo la zona tumoral no contrastada, la zona tumoral realzada y el edema periférico.

El segundo conjunto, asignado a la tarea 2 del desafío BraTS 2021, se enfoca en la evaluación de métodos de clasificación para predecir el estado de metilación del promotor MGMT, clave en la respuesta al tratamiento por quimioterapia en glioblastomas.

Las secuencias de imágenes de resonancia magnética, son las siguientes:

- Fluid Attenuated Inversion Recovery (FLAIR)
- T1-weighted pre-contrast (T1)
- T1-weighted post-contrast (T1CE)
- T2-weighted (T2)

4.3 Análisis exploratorio

Se realizó histogramas con las medias de los valores de intensidad de los píxeles de cada secuencia, como parte del análisis de variabilidad de los contrastes de las imágenes. También, una vez que fueron extraídas las variables, se realizó un segundo análisis exploratorio por medio de análisis de componentes principales (PCA) y un modelo de t-SNE.

4.4 Armonización de imágenes

Se aplicó una armonización de las imágenes por medio de una coincidencia por histogramas, con el fin de disminuir la variabilidad de contraste de las imágenes entre una misma secuencia. La imagen

de referencia se escogió mediante el análisis de los histogramas, seleccionando una de las imágenes que estuviera en el intervalo que presentara una mayor frecuencia.

La coincidencia de histograma o la especificación de histograma, en el campo de procesamiento de imágenes, es la transformación de una imagen para que su histograma coincida con un histograma específico (Gonzalez & Woods, 2006).

Según el sitio de scikit-image (Walt et al., 2014), la coincidencia de histograma se puede utilizar como una normalización para el procesamiento de imágenes, especialmente en circunstancias en las que las imágenes se han tomado de diferentes fuentes. En la figura 4.2 se presenta un ejemplo, obtenido del sitio de scikit-image:

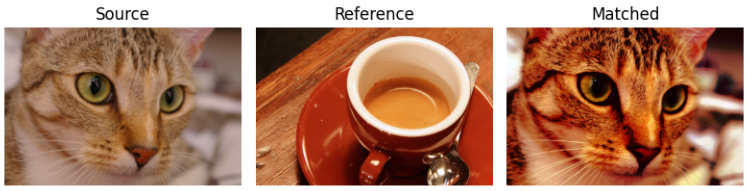

Figura 4.2: Coincidencia de histograma

4.5 Extracción de variables.

Se utilizó el software 3Dslicer (Fedorov et al., 2012), al cual se importaron las imágenes de cerebro y sus respectivas máscaras de segmentación tumoral, para así extraer las variables correspondientes a la zona del tumor que no capta contraste (net), la zona que si capta contraste (ed), la zona de edema tumoral (ed), el tumor, y finalmente el tumor junto a la zona peritumoral, estas últimas 2 se formaron en base a las primeras 3. Luego, se aplicó la extensión pyradiomics, y se extrajeron todos los conjuntos de características cuantitativas disponibles.

4.6 Selección de variables.

A continuación, se aplicaron las siguientes técnicas de selección de variables:

LASSO

El método LASSO (*Least Absolute Shrinkage and Selection Operator*) es un método de selección de variables y contracción de algunos parámetros hacia 0. Siendo un B el vector de coeficientes, el método LASSO lo estima mediante el problema de mínimos cuadrados, como se muestra a continuación (González, 2015):

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \tag{4.1}$$

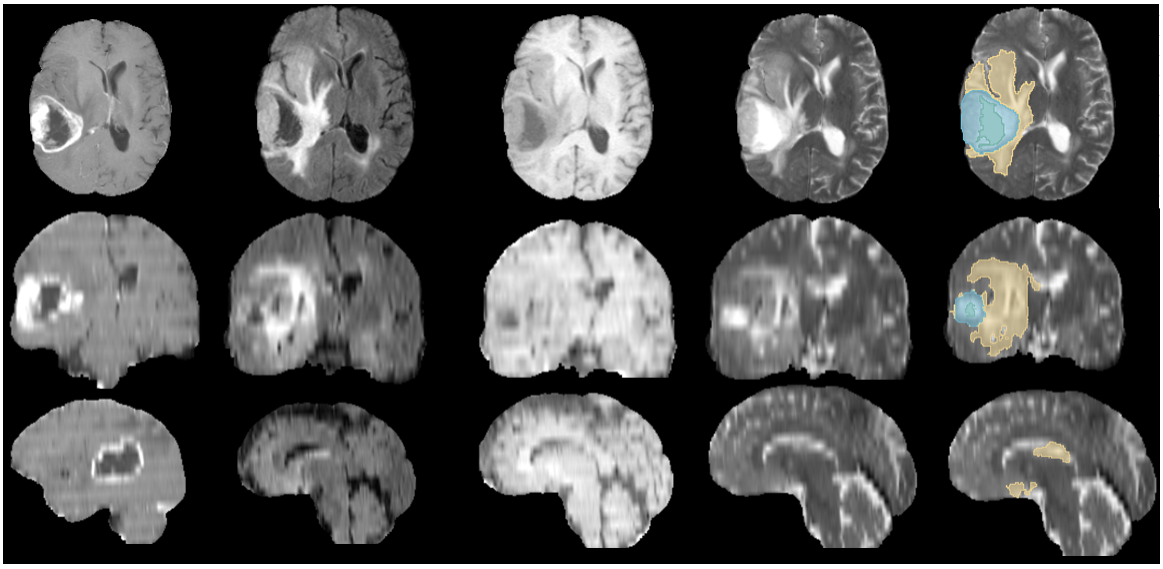

Figura 4.3: Secuencia T1-ce, secuencia FLAIR, secuencia T1, secuencia T2 y secuencia T1CE con las mascaras de zona del tumor que no capta contraste(verde), la zona que si capta contraste (celeste), y la zona de edema tumoral (amarillo)

Sujeto a:

$$\sum_{j=1}^{p} |\beta_j| \leq t \tag{4.2}$$

Otra forma minimizada es:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.3}$$

Siendo t y λ los parámetros de regularización.

Fisher-score

Los métodos de filtro utilizan las propiedades estadísticas de las variables para poder seleccionar las que entregan mayor información al modelo. Dentro de estos métodos se puede utilizar el Fisher-score (F-score), según lo siguiente:

$$F(j) = \left| \frac{u_j^+ - u_j^-}{(\sigma_j^+)^2 - (\sigma_j^-)^2} \right| \tag{4.4}$$

Donde u_j^+, u_j^- es la media del j-ésimo atributo en la clase positiva y negativa. Y σ_j^+, $-\sigma_j^-$ las correspondientes desviaciones estándar(Maldonado, 2015). Luego se eligen las variables con mejor valor de Fisher-score.

Recursive Feature elimination

Este método se realiza en conjunto con la construcción del clasificador y es especifico de cada técnica de selección (Maldonado et al., 2015). Consta de las siguientes partes, construcción del modelo sobre el conjunto de variables, calculo de los valores de importancia de cada variables, y eliminación de los predictores menos importantes (Guyon et al., 2002).

Forward Selection

Este método consiste en que se comienza de un modelo inicial vacío al cual se agrega una sola variable, y luego se van incluyendo las otras, una a una, en cada iteración, según el siguiente algoritmo (Guerra, 2016):

- Se escoge como variable inicial (x1) la que tenga mejor correlación con la variable respuesta (y).
- A continuación, se calcula un modelo de regresión lineal simple entre x1 e y, y las correlaciones parciales entre (x2, ..., xn) e y, habiendo eliminado de la variable respuesta el efecto de la variable x1
- Luego se introduce al modelo la variable con mayor correlación con y, siendo un x2.
- Se vuelve a calcular un modelo con ambas variables y se observa el estadístico t para coeficiente de x2, Bgorro2, si es significativo. Después se tienen 2 opciones, una opción es que, si es significativo, se introduce una variable x que tenga mayor correlación, en cambio si no es significativo, se elimina la variable y el proceso termina. Se continua con esto hasta que no queden más variables, o se obtenga un valor t no significativo.

Backward Selection

En el caso de backward selection, comienza con todas las variables explicativas, como se indica a continuación (Guerra, 2016):

- Comienza el modelo con todas las variables y se fija un valor del estadístico t para elegir las variables
- Se calcula el modelo de regresión lineal y los estadístico t para cada coeficiente
- Se eliminan las variables que tenga menor valor que el fijado
- Se calcula el modelo con las n-1 variables y los estadísticos t para los coeficientes
- Se itera hasta no haya más variables o el t no sea significativo

4.7 Modelos y evaluación de desempeño

Luego con las variables seleccionadas se utilizaron los siguientes modelos, con sus grillas de hiperparámetros correspondientes, cada uno con 5 validaciones cruzadas:

Table 4.1: Modelos con grillas de hiperparámetros

Modelo	Grilla de hiperparámetros
Logistic Regression	penalty = [l2, L1, ElasticNet, None], C = [0.1, 1, 10, 100] solver = [lbfgs, liblinear, newton-cg, newton-cholesky, sag, saga]
Gradient Boost	learning rate = [0.01, 0.1, 0.5], n estimators = [50, 100, 500, 1000] subsample = [0.1,0.5,0.9]
Random Forest	n estimators = [100, 500, 1000,2000], max features = [None, log2, sqrt]
Support Vector Classifier	C = [0.001, 0.1, 1, 10, 100, 1000], gamma = [0.0001, 0.01, 0.1, 1, 10 kernel = [rbf, linear, poly, sigmoid]
Xgboost	learning rate = [0.001,0.1,0.3,0.9], n estimators = [50,250,500]
Catboost	depth = [4,6,8,10],learning rate = [0.01,0.02,0.03,0.04] iterations = [10,20,50,70,100]

Para evaluar el desempeño de los modelos se usan las métricas de desempeño o de rendimiento, las cuales buscan facilitar y visualizar de mejor manera la elección de un modelo. Estas métricas están basadas en la matriz de confusión (Awad & Khanna, 2015), la cual es una tabla que resume los resultados obtenidos con un modelo de clasificación prediciendo. Esta tabla resumen esta compuesta por un eje que representa la etiqueta que el modelo predijo, y otro eje con la etiqueta real, como se muestra en la figura 4.4 (Burkov, 2019):

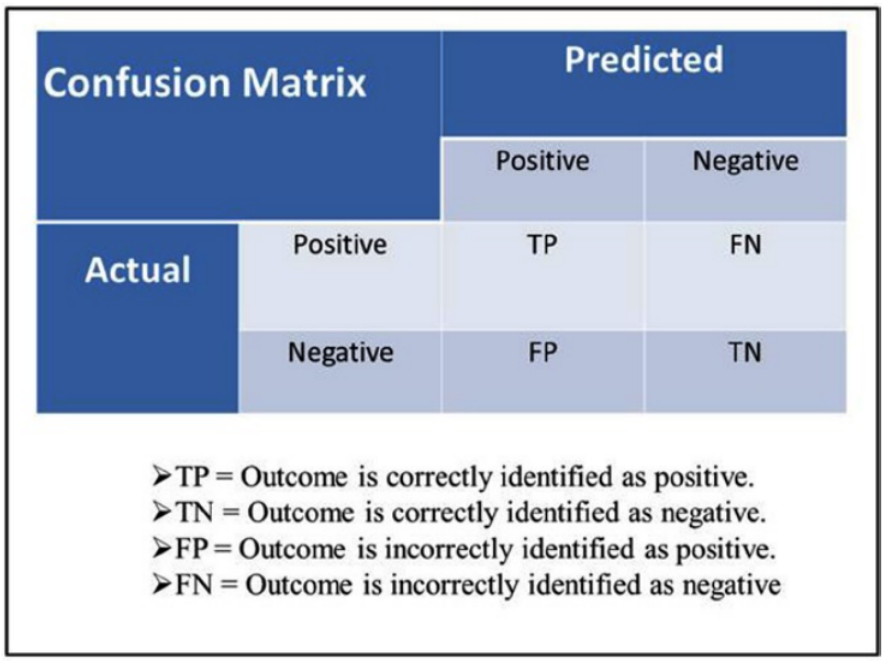

Figura 4.4: Matriz confusion obtenida del libro efficient learning machines

Las métricas que se considerarán, se definen a continuación:

- Accuracy o exactitud: es el número de ejemplos clasificados correctamente dividido por el número total de ejemplos .

$$Accuracy = \frac{TP + TN}{TOTAL} \tag{4.5}$$

- Recall o sensibilidad: es la relación entre las predicciones positivas coreectas y el número total de ejemplos positivos.

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

- Precisión: la precisión es la relación entre las predicciones positivas correctas, y el número total de predicciones positivas.

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

- F1-score: es la media armónica entre precisión y recall.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.8}$$

- Área bajo la curva ROC (AUC): La sigla ROC viene de la frase "características operativa del recepto" . Las curvas ROC utilizan una combinación de la tasa de verdaderos positivos (TVP), y una tasa de falsos positivos (TFP) para construir un gráfico que resume el rendimiento de la clasificación. Un modelo con un AUC mayor a 0.5, es mejor que un modelo aleatorio, y mientras más se acerque al valor 1, es un mejor modelo.

$$TVP = \frac{TP}{TP + FN} \tag{4.9}$$

$$TFP = \frac{FP}{FP + TN} \tag{4.10}$$

4.8 Shap values para explicabilidad

Finalmente, se realizó un análisis de shap values para la explicabilidad del modelo, mediante la librería shap para obtener un gráfico con las diez variables más influyentes, para así poder observar que características presentan una mayor influencia en la predicción de la metilación del promotor de MGMT.

Chapter 5

Resultados

5.1 Histogramas

Al observar los histogramas de las secuencias FLAIR, T1 y T1CE se aprecia una mayor variedad en las intensidades de los pixeles, e incluso se pueden identificar 2 o 3 grupos, los cuales pueden pertenecer a diferentes grupos de adquisición, donde se utilizarón diferentes parámetros o equipos. La secuencia T2, es la que presenta valores más homogéneos.

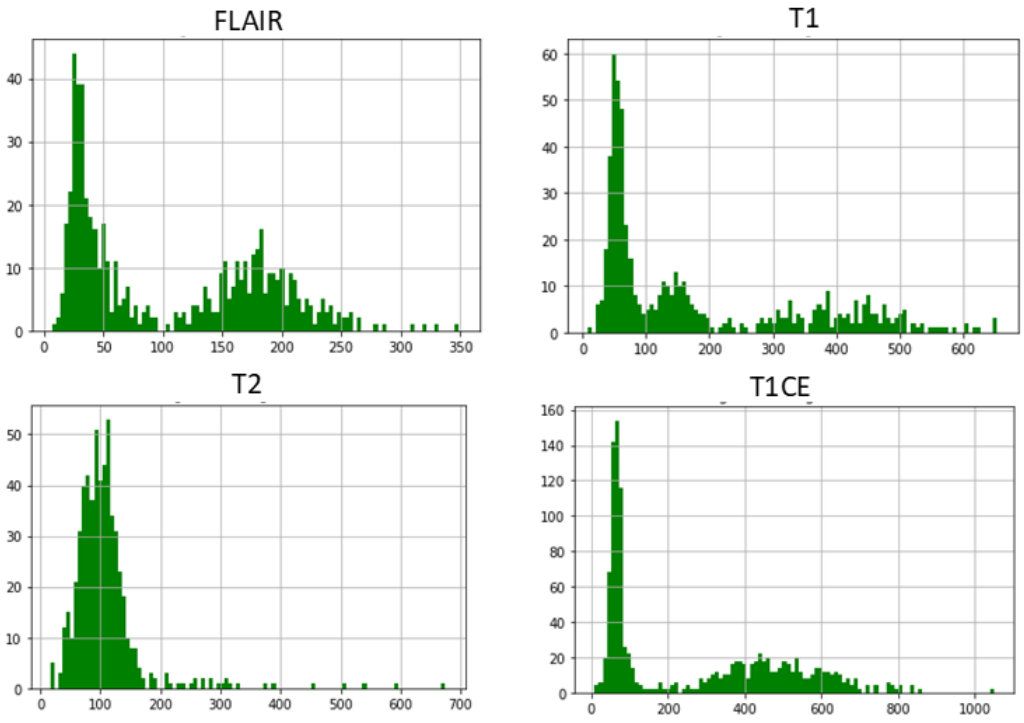

Figura 5.1: Histograma de intensidades de pixeles de cada secuencia

Lo anteriormente visto en el histograma, se ve ejemplificado en la figura 5.2, en la cual se puede ver que dentro de una misma secuencia, tienen una escala de contraste diferente.

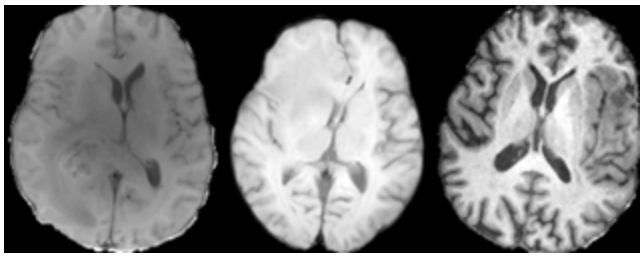

Figura 5.2: Secuencia T1 con diferente contraste

5.2 Armonización por coincidencia de histogramas

Después de aplicar la técnica de armonización mediante la coincidencia de histogramas utilizando una imagen de referencia, los histogramas de medias cambiaron en su distribución, como se ve en la figura 5.3:

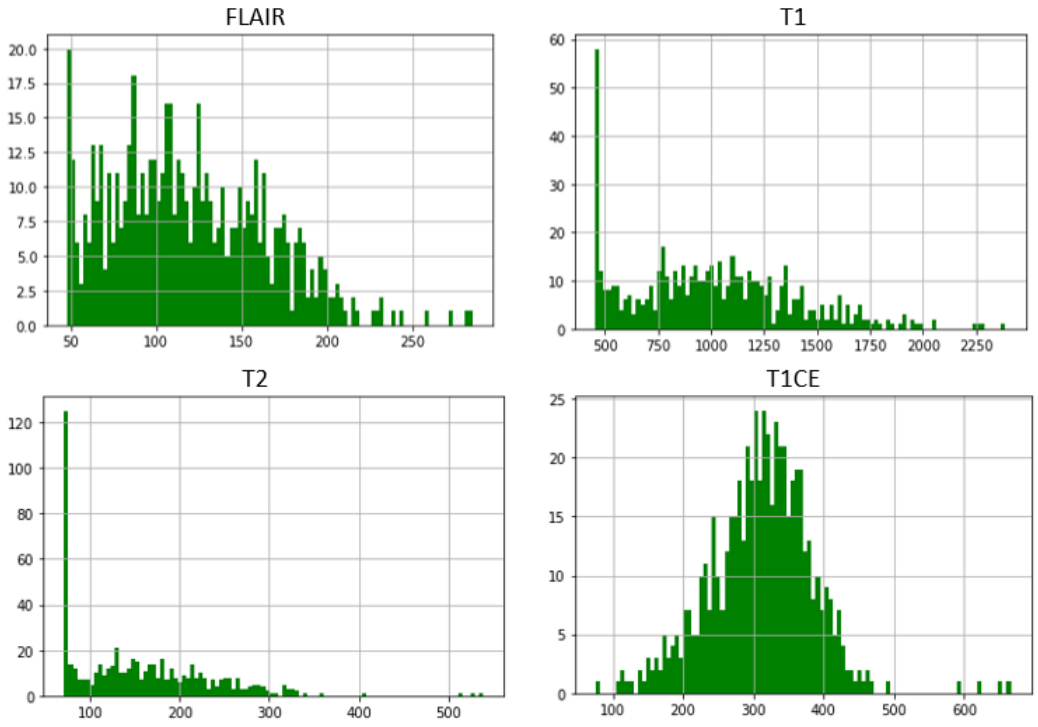

Figura 5.3: Histograma de intensidades de pixeles de cada secuencia

A continuación en las figuras 5.3, 5.4, 5.5, 5.6 y 5.7, se presentan los conjuntos de imágenes por secuencia, con los resultados de la armonización, pero desde el punto de vista de la imagen. En el lado izquierdo se encuentran las imágenes originales, en el centro se muestra la imagen de referencia y en el lado derecho se encuentran las imágenes resultantes:

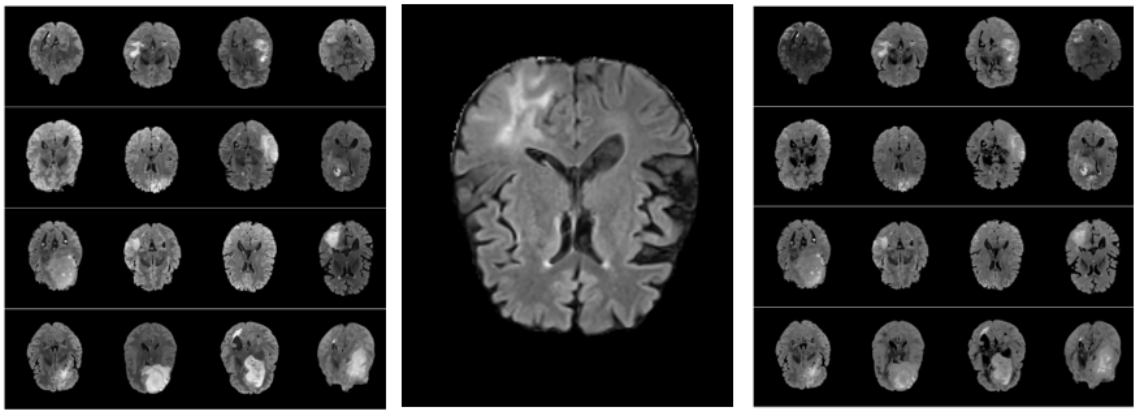

Figura 5.4: Secuencia FLAIR

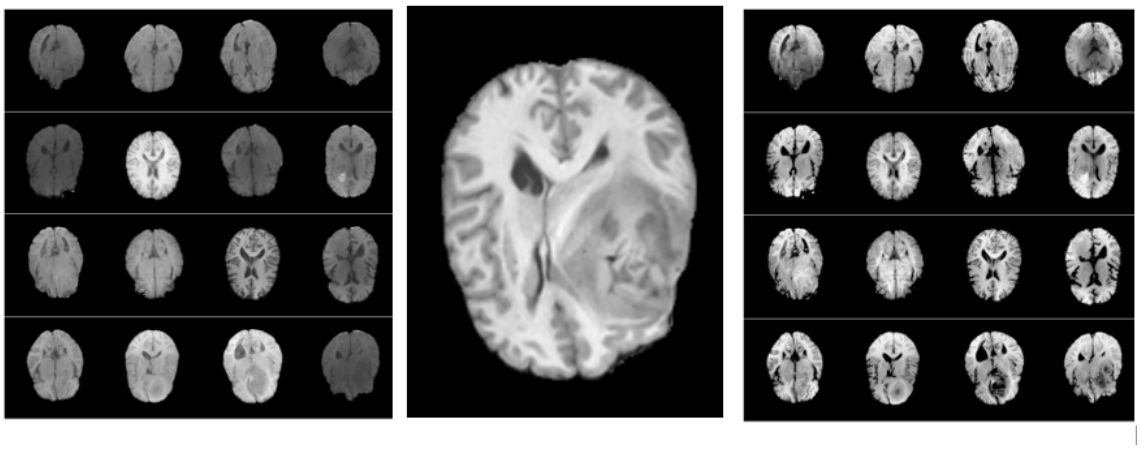

Figura 5.5: Secuencia T1

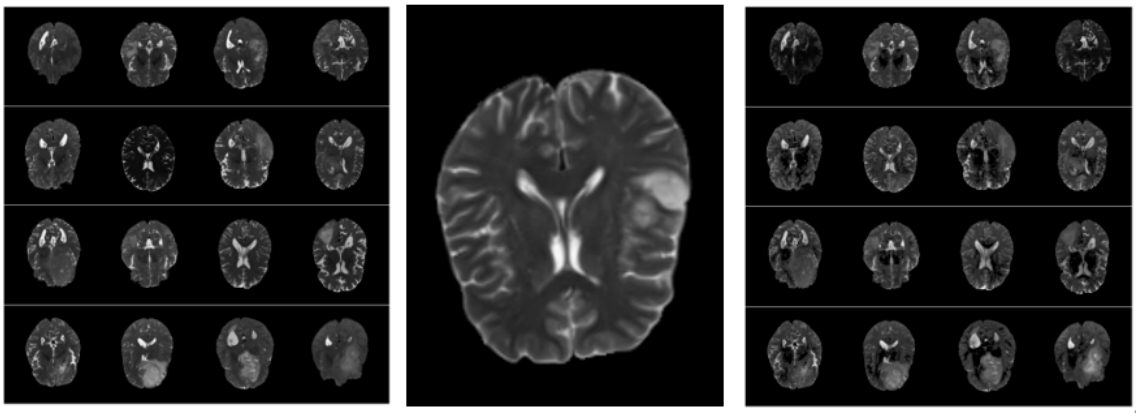

Figura 5.6: Secuencia T2

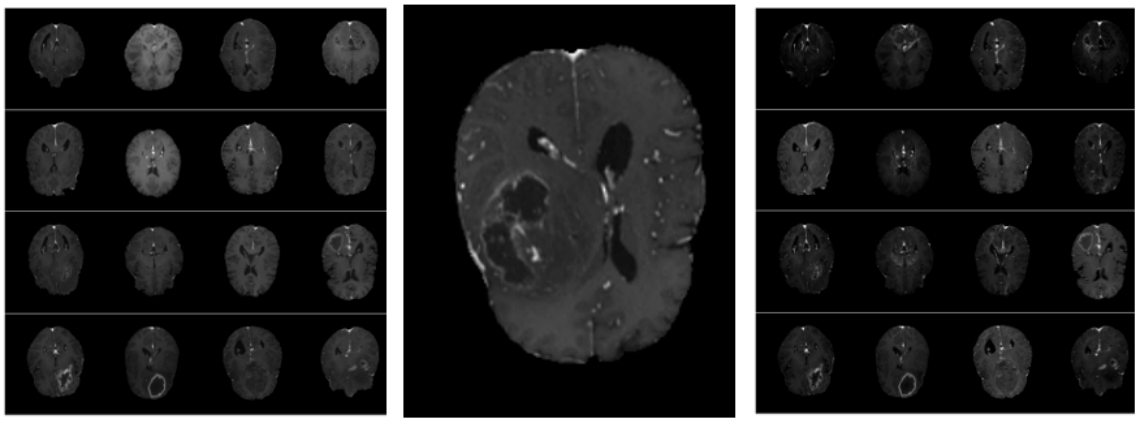

Figura 5.7: Secuencia T1CE

5.3 Base de datos generada con variables extraídas

Una vez extraídas las variables, se generó una base de datos con valores cuantitativos, la cual quedó compuesta por 558 observaciones y 2141 variables, correspondiendo a 535 variables para cada secuencia de imagen, y una variable vector (MGMT).

Table 5.1: Resumen de variables extraídas

Base de datos		
Tipo	Comentario	Cantidad
Textura	Todas las características de niveles de grises	75
Primer orden	Se consideran todos, a excepción de la desviación estándar	18
Shape	Solo las de forma 3D	14
Features por segmentación		107
Features por secuencia	Cada secuencia tiene 5 segmentaciones	535
Total	Considerando las 4 secuencias y 5 segmentaciones	2140

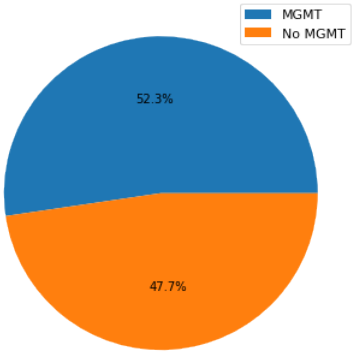

Figura 5.8: Gráfico de porcentajes correspondientes a cada clase

Se excluyeron de la base de datos las siguientes cantidades de conjuntos:

- Diez conjuntos de imágenes que carecían de sus correspondientes segmentaciones.
- Dieciocho conjuntos de imágenes que mostraban valores nulos en, al menos, una de sus variables extraídas.

5.4 Análisis exploratorio con PCA y t-SNE

Al analizar los gráficos generados con t-SNE, y al explorar la posibilidad de separar los datos en tres componentes principales utilizando PCA, no se ha podido identificar una clara distinción entre los dos grupos en ninguna de las secuencias.

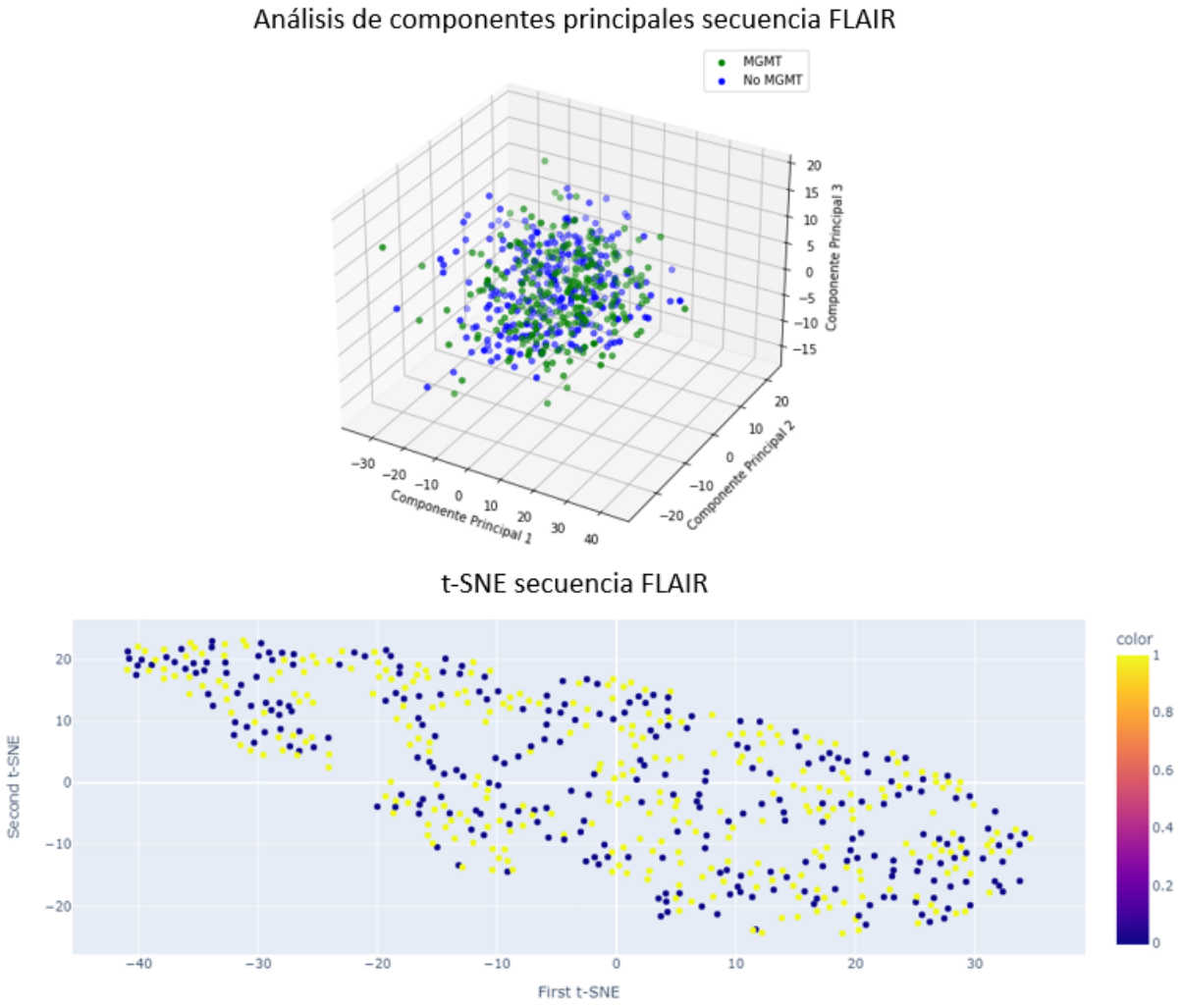

Figura 5.9: Secuencia FLAIR

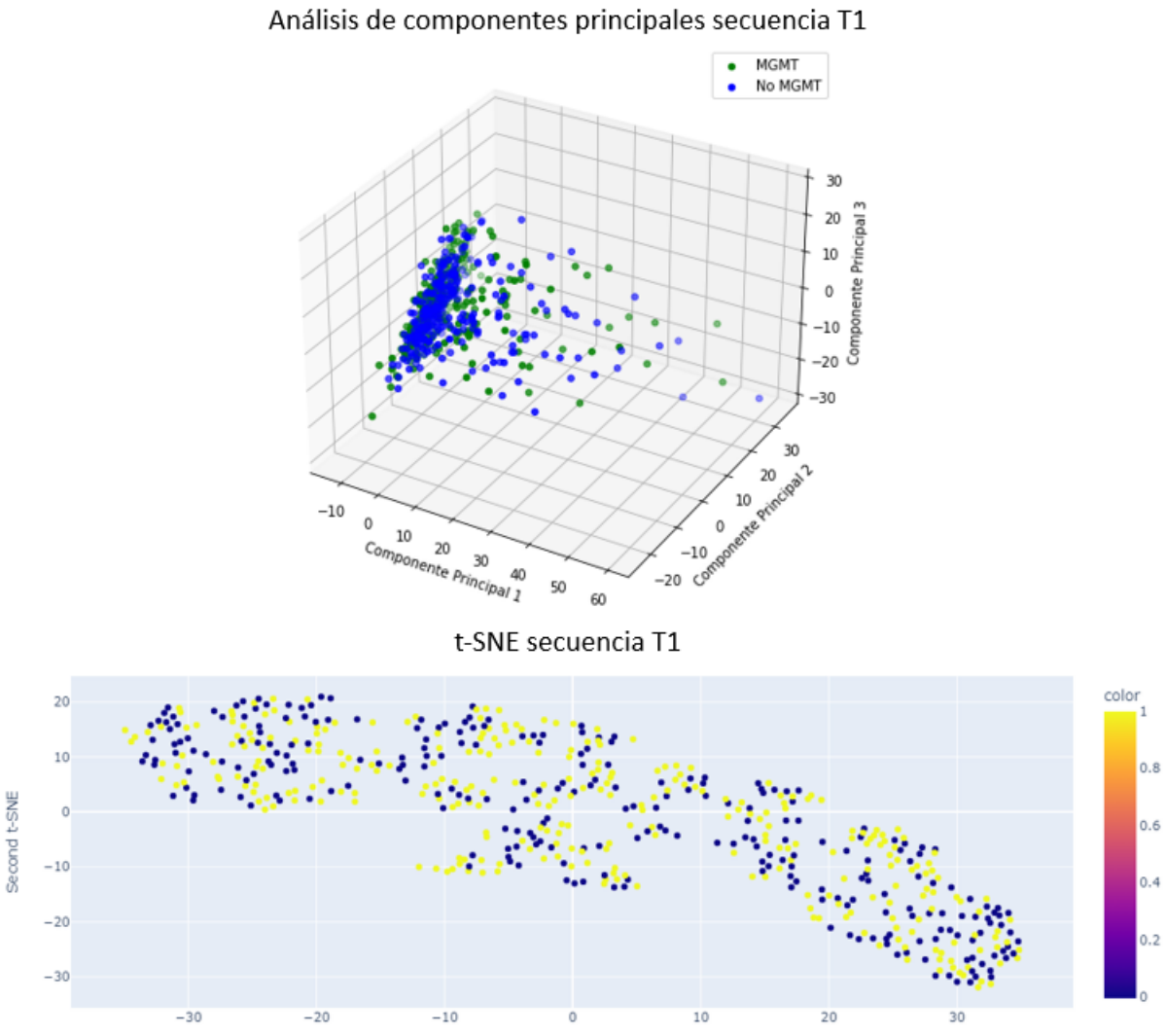

Figura 5.10: Secuencia T1

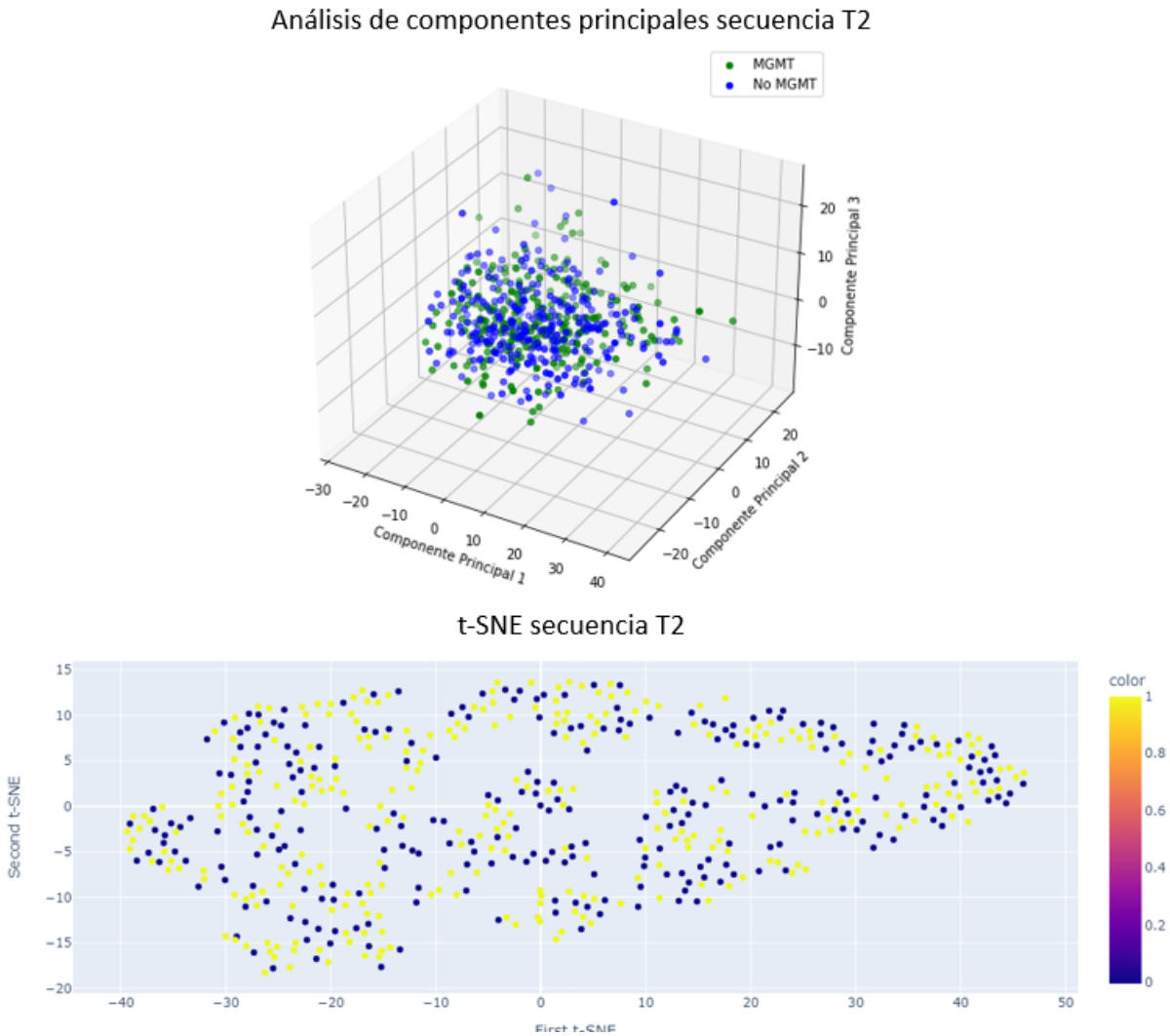

Figura 5.11: Secuencia T2

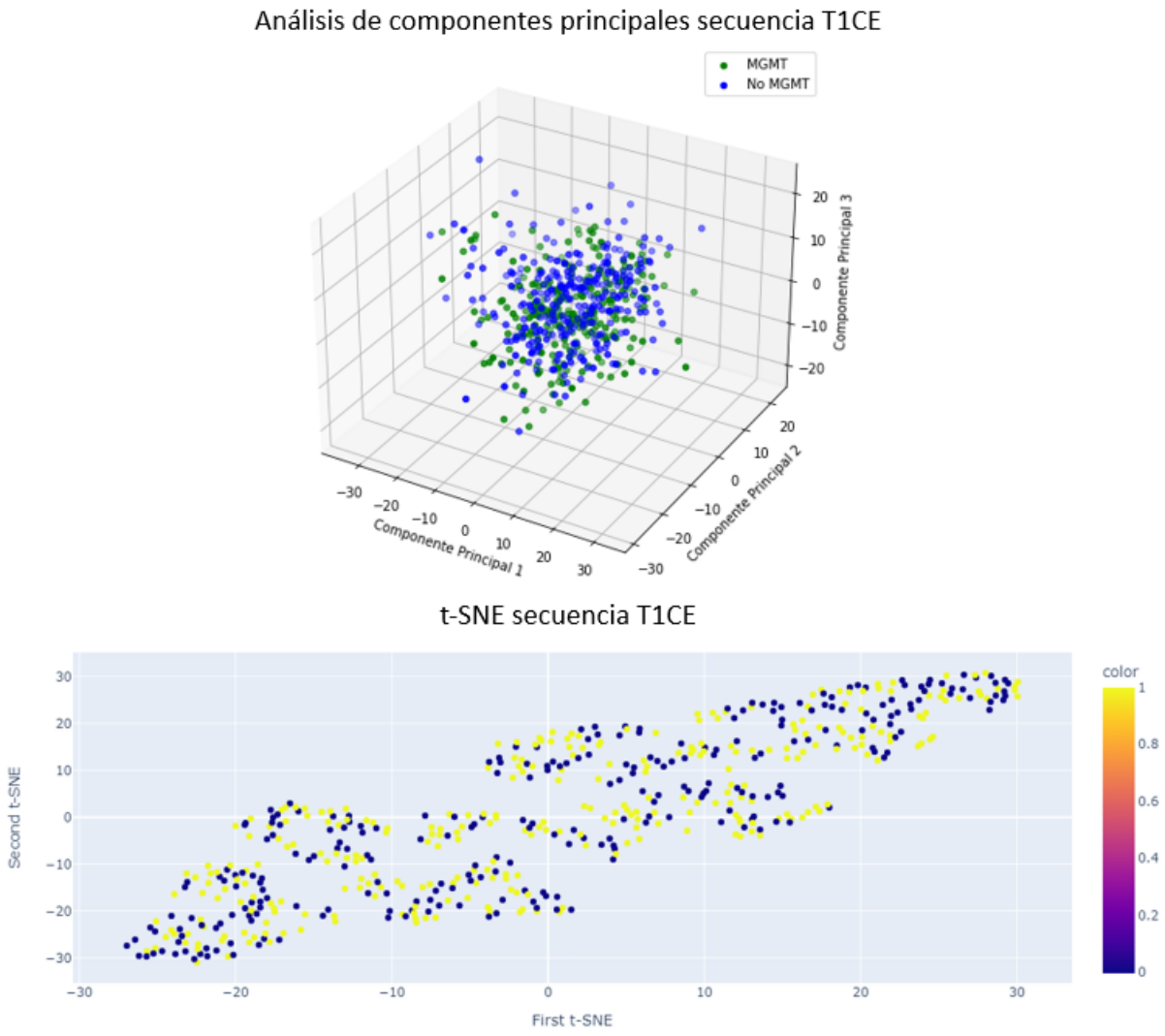

Figura 5.12: Secuencia T1-ce

5.5 Métricas de los modelos

Las siguientes métricas corresponden a los modelos considerendo todas las variables extraídas.

Métricas de desempeño						
Modelo	Clase	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0	0.55	0.56	0.55	0.55	0.59
	1		0.54	0.55	0.54	
Gradient boosting	0	0.50	0.51	0.45	0.48	0.54
	1		0.48	0.54	0.51	
Random forest	0	0.54	0.55	0.54	0.54	0.52
	1		0.52	0.54	0.53	
Support vector machine	0	0.56	0.60	0.34	0.43	0.55
	1		0.53	0.77	0.62	
Catboost	0	0.51	0.53	0.53	0.53	0.51
	1		0.49	0.49	0.49	
XGBoost	0	0.51	0.52	0.55	0.53	0.51
	1		0.50	0.48	0.49	

Table 5.2: Métricas considerando todas las variables

El modelo con mejor rendimiento corresponde a Logistic Regression con técnica selección LASSO, en el cual se obtuvieron las siguientes métricas:

Metricas de desempeño							
Modelo	Tecnica de seleccion	Clase	accuracy	precision	recall	f1-score	roc-auc
Logistic Regression	LASSO (153 variables)	0	0.61	0.62	0.60	0.61	0.67
		1		0.60	0.61	0.60	

Table 5.3: Métricas del modelo con variables reducidas con mejor rendimiento

En los gráficos de las curvas ROC, se muestran los valores de desempeño de todos los modelos, ya sea aquellos en los que se consideraron todas las variables o aquellos en los que se redujo la cantidad de variables. Se puede apreciar que en la mayoría de los modelos, la reducción de variables conduce a una mejora en el rendimiento.

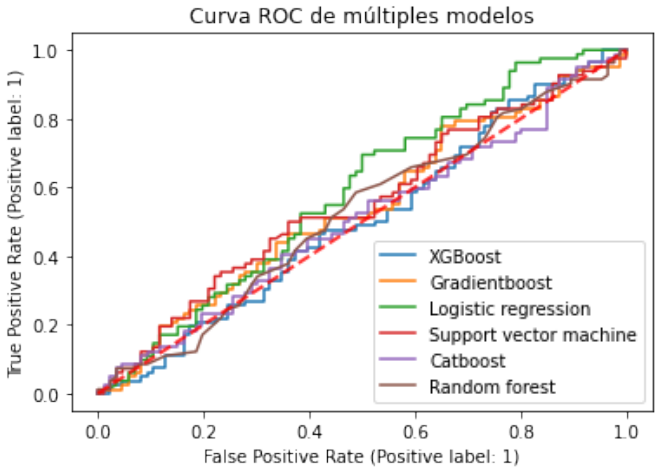

Figura 5.13: Curva ROC de los modelos con todas las variables

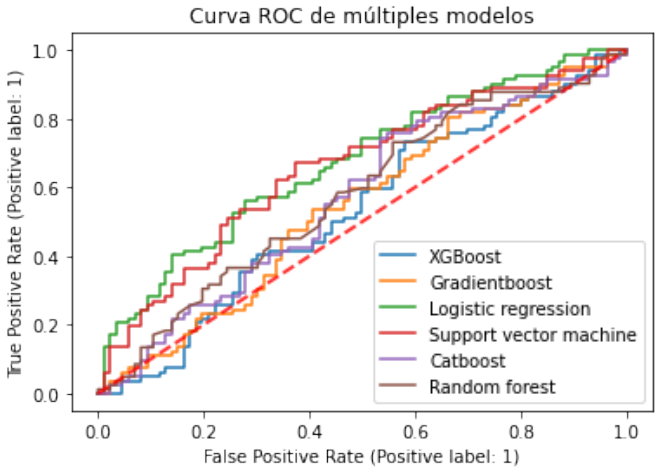

Figura 5.14: Curva ROC de los modelos con reducción de variables

5.6 Shap values

Las tres variables de mayor importancia fueron las siguientes: una variable de secuencia T1CE relacionada a la textura, una variable de FLAIR relacionada al tipo de estadísticas de primer orden, y una variable T1 del tipo de estadísticas de primer orden. De las diez variables más destacadas, predominan principalmente las características relacionadas con la textura y estadísticas de primer orden, lo que implica que las variables relacionadas con la forma no se encuentran dentro de las diez más relevantes.

También se puede observar que, de las diez variables, seis de ellas se derivan de segmentos más extensos, específicamente aquellos que abarcan la totalidad del tumor (la suma de los segmentos que captan contraste y los segmentos que no captan contraste), así como los segmentos que incorporan tanto el tumor como el edema. Por último, es relevante señalar que de las diez variables, ocho de ellas tienen un efecto negativo en la predicción cuando sus valores son altos, lo que indica que ejercen una mayor influencia en la probabilidad de que no esté presente el biomarcador genético.

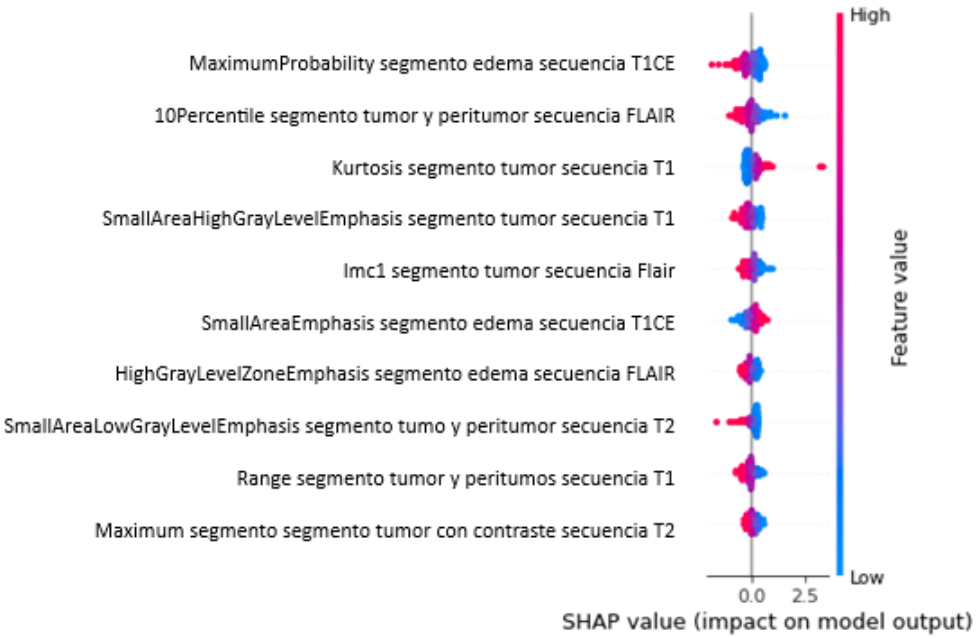

Figura 5.15: Gráfico de shap values

Chapter 6

Discusión

Al utilizar todas las variables extraídas en diferentes modelos, se observa un rendimiento insuficiente. Esto hace que la aplicación de técnicas de selección de características, o "feature selection", sea esencial para mejorar la eficacia de estos modelos. Los resultados coinciden con el estudio de Le (Le et al., 2020), que también indican una mejora en el desempeño al aplicar estas técnicas..

El modelo que mostró el mejor rendimiento fue logistic regression con la técnica de selección LASSO, logrando un AUC de 0,67. Esto es ligeramente superior al estudio "Prediction of MGMT Methylation Status of Glioblastoma using Radiomics and Latent Space Shape Features" (Pálsson et al., 2022), que obtuvo un AUC de 0,598 usando un modelo de random forest y la técnica de selección LASSO (15 variables seleccionadas). El incremento en el desempeño podría atribuirse a la segmentación manual de imágenes de las zonas tumorales por expertos, a diferencia del estudio mencionado, que usó una segmentación automática, y también a que se utilizó una armonización de imágenes. Además, se obtuvo un mejor resultado que Firas Baba (Baba, n.d.), que fue el ganador de la competencia en Kaggle con un roc-auc de 0,62173 usando un modelo ResNet, sin una aplicar metodología basada en radiómica.

En el estudio "XGBoost Improves Classification of MGMT Promoter Methylation Status in IDH1 Wildtype Glioblastoma" (Le et al., 2020), donde se utilizó una base de datos de 53 pacientes, se obtuvo un desempeño de ROC-AUC de 0.896 con un modelo XGBoost y selección de variables mediante F-score (9 variables seleccionadas). No se menciona el uso de preprocesamiento en este caso.

Comparado con el estudio "Radiomics signature: A potential biomarker for the prediction of MGMT promoter methylation in glioblastoma", nuestro modelo obtuvo un rendimiento menor, donde alcanzaron un accuracy 85,59% con un modelo de support vector machine y selección LASSO, sobre una base de datos propia de ellos, que estaba compuesta por 98 conjuntos de imágenes de pacientes (Xie et al., 2015).

Es importante considerar que los movimientos, artefactos y protocolos de adquisición de imágenes, que pueden variar según los parámetros de adquisición o la marca y modelo del equipo, generan diferencias en el contraste y la resolución. Esto puede llevar a problemas en la extracción de características y su análisis posterior, ya que la variabilidad puede afectar el rendimiento del modelo. Un ejemplo de variabilidad en las imágenes, es el que se presenta en la figura 6.1

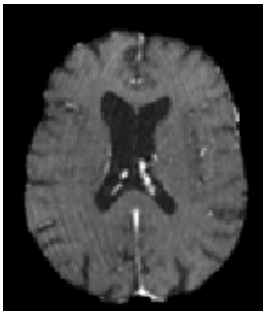

Figura 6.1: Imagen con movimiento

Para abordar los problemas mencionados anteriormente, existen estudios que proponen herramientas específicas para el preprocesamiento de datos. En el caso de esta investigación, se empleó una algoritimo simple de procesamiento de imágenes como fue la armonización por histograma, lo cual mejora levemente el desempeño, en comparación a cuando no se utiliza esta técnica. Existe también un método llamado ComBat diseñado para mitigar la variabilidad de ciertos parámetros, como el grosor de corte (Orlhac et al., 2022). Sin embargo, una de las limitaciones de este método es que requiere el conocimiento de los parámetros de adquisición de la imagen para que funcione con eficacia.

Otra estrategia a considerar para futuras investigaciones, es explorar valores adicionales de hiperparámetros que pudiesen mejorar el rendimiento de los modelos. También considerar armonizaciones de imágenes por algún medio diferente al de armonizacion por histograma con imagen de referencia.

Por lo observado en las técnicas de PCA y t-SNE, se demuestra que el objetivo del estudio, es un problema complejo, ya que no se logra identificar de manera simple alguna separación notoria entre los grupos, lo cual explicaría los valores bajos en las métricas de desempeño. Para lograr una mejor separación de grupos y mejorar el desempeño, podría ser beneficioso explorar otra variable objetivo, como un gen diferente que esté igualmente relacionado con la sensibilidad a la quimioterapia u otros tipos de tratamiento.

En cuanto a la base de datos de imágenes utilizada, una de las desventajas principales es el desconocimiento de los parámetros de adquisición. En futuros estudios, sería beneficioso adquirir las imágenes en un centro hospitalario local para tener acceso directo a estos datos. Además, se podría enriquecer el análisis incorporando factores clínicos de los pacientes, como la edad, género o signos radiológicos. Un estudio de Pszczolkowski (Pszczolkowski et al., 2021), enfocado en la expansión de hematomas cerebrales mediante imágenes de tomografía computada, concluyó que la combinación de características basadas en radiómica con factores clínicos mejora significativamente el desempeño.

Finalmente, sería útil explorar otras técnicas de selección de características, métodos de preprocesamiento, algoritmos de armonización y estrategias para explicar los resultados, lo que podría ofrecer una mayor comprensión y optimización del proceso.

Chapter 7

Conclusión

En este estudio, se logró desarrollar una metodología de radiómica para poder identificar y evaluar biomarcadores de tumores cerebrales en resonancia magnética. Entre las técnicas utilizadas, la selección de variables LASSO, combinada con un modelo de Logistic Regression, resultó ser la más efectiva.

A través de la aplicación de técnicas de minería de datos se logró extraer y reducir la cantidad de variables de las imágenes, obteniendo un rendimiento en los modelos en la predicción de la clasificación comparable al estado del arte, en los casos que se utilizó la misma base de datos. Sin embargo, es importante reconocer que estos resultados podrían mejorarse mediante un ajuste diferente de hiperparámetros, y la exploración de diferentes técnicas de selección de variables y preprocesamiento de datos

El análisis de shap values reveló que las tres variables de mayor importancia fueron una variable de secuencia T1CE relacionada a la textura, una variable de FLAIR del tipo de estadísticas de primer orden, y una variable T1 del tipo de estadísticas de primer orden de igual manera. Esta información no sólo es relevante desde una perspectiva técnica, sino que también subraya la importancia de llevar a cabo un análisis de explicabilidad para obtener una comprensión más profunda e interpretación más precisa de los modelos de aprendizaje automático.

Para futuras investigaciones, sería recomendable considerar la creación de una base de datos propia o local, pues esto permitiría acceder a información más detallada sobre la adquisición y condiciones de los pacientes. Como se señaló anteriormente en la discusión, los estudios que emplearon bases de datos propias tendían a obtener resultados con mejor desempeño. Esta estrategia podría, por tanto, representar una vía para estudiar de mejor manera el problema, para identificar si existe una mejora en el desempeño en la detección y evaluación de biomarcadores de tumores cerebrales, ya que se dispondría de información más detallada y completa.

Referencias

Alba Vega, D. A., & Calle Jara, J. F. (2020, April). *Aplicación de técnicas de machine learning basado en información sísmica para profundizar la probabilidad de terremotos mediante el uso de regresión logística y redes neuronales.* [bachelorThesis]. Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales. [Accepted: 2020-08-17T19:02:29Z]. Retrieved November 1, 2023, from http://repositorio.ug.edu.ec/handle/redug/48862

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and unsupervised learning for data science* (pp. 3–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1

Amazon Web Services. (2021). *Interpretability versus explainability - model explainability with AWS artificial intelligence and machine learning solutions.* Retrieved November 1, 2023, from https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Atlas, S. (2016). *Magnetic resonance imaging of the brain and spine* (4ta edicion). Lippincott Williams & Wilkins.

Avanzo, M., Wei, L., Stancanello, J., Vallières, M., Rao, A., Morin, O., Mattonen, S. A., & El Naqa, I. (2020). Machine and deep learning methods for radiomics. *Medical Physics*, *47*(5), e185–e202. https://doi.org/10.1002/mp.13678

Avery, E., Sanelli, P. C., Aboian, M., & Payabvash, S. (2022). Radiomics: A primer on processing workflow and analysis. *Seminars in Ultrasound, CT and MRI*, *43*(2), 142–146. https://doi.org/10.1053/j.sult.2022.02.003

Awad, M., & Khanna, R. (2015). Support vector regression. In M. Awad & R. Khanna (Eds.), *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Baba, F. (n.d.). *Monai simple prediction from flair.* Retrieved November 1, 2023, from https://kaggle.com/code/rinnqd/monai-simple-prediction-from-flair

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T. J., Brennan, M. C., Li, M., Church, E. W., Brandmeir, N. J., Rakszawski, K. L., Patel, A. S., Rizk, E. B., Suki, D., Sawaya, R., & Glantz, M. (2016). Association of the extent of resection with survival in glioblastoma: A systematic review and meta-analysis. *JAMA oncology*, *2*(11), 1460–1469. https://doi.org/10.1001/jamaoncol.2016.1373

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov. https://books.google.cl/books?id=0jbxwQEACAAJ

Castro, J., Chabert, S., Saavedra, C., & Salas, R. (2020). Convolutional neural networks for detection intracranial hemorrhage in CT images.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Contreras, L. E. (2017). Epidemiología de tumores cerebrales. *Revista Médica Clínica Las Condes*, *28*(3), 332–338. https://doi.org/10.1016/j.rmclc.2017.05.001

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Costa, J., & Jerez, J. (2021). *Resonancia magnética dirigida a técnicos superiores en imagen para el diagnóstico*. Elsevier Health Sciences. https://books.google.cl/books?id=k7gZEAAAQBAJ

del Cancer, I. N. (2022). *Marcadores tumorales* [Archive Location: nciglobal,ncienterprise]. Retrieved November 3, 2023, from https://www.cancer.gov/espanol/cancer/diagnostico-estadificacion/diagnostico/hoja-informativa-marcadores-de-tumores

Dolecek, T. A., Propp, J. M., Stroup, N. E., & Kruchko, C. (2012). CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the united states in 2005-2009. *Neuro-Oncology*, *14 Suppl 5*, v1–49. https://doi.org/10.1093/neuonc/nos218

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito [Publisher: Facultad de Ingeniería, UNAM]. *Ingeniería, investigación y tecnología*, *21*(3). https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S. N., Hidalgo, O. F., Vanaclocha, V., Baylin, S. B., & Herman, J. G. (2000). Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *The New England Journal of Medicine*, *343*(19), 1350–1354. https://doi.org/10.1056/NEJM200011093431901

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., & Kikinis, R. (2012). 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

Filos, J. (2008, October). *Parallel transmission mri* [Doctoral dissertation].

Forghani, R., Savadjiev, P., Chatterjee, A., Muthukrishnan, N., Reinhold, C., & Forghani, B. (2019). Radiomics and artificial intelligence for biomarker and prediction model development in oncology. *Computational and Structural Biotechnology Journal*, *17*, 995–1008. https://doi.org/10.1016/j.csbj.2019.07.001

for Health & Care Excellence, N. I. (2018, July 11). *Overview — brain tumours (primary) and brain metastases in over 16s — guidance — NICE* [Publisher: NICE]. Retrieved November 4, 2023, from https://www.nice.org.uk/guidance/ng99

Franco, P., Sotelo, J., Guala, A., Dux-Santoy, L., Evangelista, A., Rodríguez-Palomares, J., Mery, D., Salas, R., & Uribe, S. (2022). Identification of hemodynamic biomarkers for bicuspid aortic valve induced aortic dilation using machine learning. *Computers in Biology and Medicine*, *141*, 105147. https://doi.org/10.1016/j.compbiomed.2021.105147

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Giovanetti, M. (2021). *Modelos explicables de aprendizaje automático para la predicción de complicaciones perinatales en mujeres embarazadas con diabetes* [Doctoral dissertation, Universidad de Valparaíso].

Gonzalez, R. C., & Woods, R. E. (2006, January). *Digital image processing (3rd edition)*. Prentice-Hall, Inc.

González, A. (2015). *Selección de variables: Una revisión de métodos existentes* [Doctoral dissertation, Universidad de Coruña].

Guerra, A. (2016). *Técnicas de selección de variables en minería estadística de datos* [Doctoral dissertation] [Accepted: 2016-07-19T12:03:11Z]. https://idus.us.es/handle/11441/43804

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Han, L., & Kamdar, M. R. (2018). MRI to MGMT: Predicting methylation status in glioblastoma patients using convolutional recurrent neural networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *23*, 331–342.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Hegi, M. E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., de Tribolet, N., Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., Bromberg, J. E. C., Hau, P., Mirimanoff, R. O., Cairncross, J. G., Janzer, R. C., & Stupp, R. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *The New England Journal of Medicine*, *352*(10), 997–1003. https://doi.org/10.1056/NEJMoa043331

Hernández Muñiz, S., & Mitjavila Casanovas, M. (2006). Introducción a la tomografía computarizada. *Revista Española de Medicina Nuclear*, *25*(3), 206–214. https://doi.org/10.1157/13088421

Herzog, C. (2022). On the risk of confusing interpretability with explicability. *AI and Ethics*, *2*(1), 219–225. https://doi.org/10.1007/s43681-021-00121-9

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312. https://doi.org/10.1002/widm.1312

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost classifier with other machine learning methods [Number: 11 Publisher: The Science and Information (SAI) Organization Limited]. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(11). https://doi.org/10.14569/IJACSA.2020.0111190

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2018). Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 chal-

lenge. In A. Crimi, S. Bakas, H. Kuijf, B. Menze, & M. Reyes (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 287–297). Springer International Publishing. https://doi.org/10.1007/978-3-319-75238-9_25

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, *18*(14), 7346. https://doi.org/10.3390/ijerph18147346

Katti, G., Ara, S., & Shireen, D. (2011). Magnetic resonance imaging (mri) - a review. *Intl J Dental Clin*, *3*.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J. W. L., Dekker, A., Fenstermacher, D., Goldgof, D. B., Hall, L. O., Lambin, P., Balagurunathan, Y., Gatenby, R. A., & Gillies, R. J. (2012). Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, *30*(9), 1234–1248. https://doi.org/10.1016/j.mri.2012.06.010

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G. P. M., Granton, P., Zegers, C. M. L., Gillies, R., Boellard, R., Dekker, A., & Aerts, H. J. W. L. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer (Oxford, England: 1990)*, *48*(4), 441–446. https://doi.org/10.1016/j.ejca.2011.11.036

Larue, R. T. H. M., Defraene, G., De Ruysscher, D., Lambin, P., & van Elmpt, W. (2017). Quantitative radiomics studies for tissue characterization: A review of technology and methodological procedures. *The British Journal of Radiology*, *90*(1070), 20160665. https://doi.org/10.1259/bjr.20160665

Le, N. Q. K., Do, D. T., Chiu, F.-Y., Yapp, E. K. Y., Yeh, H.-Y., & Chen, C.-Y. (2020). XGBoost improves classification of MGMT promoter methylation status in IDH1 wildtype glioblastoma. *Journal of Personalized Medicine*, *10*(3), 128. https://doi.org/10.3390/jpm10030128

Lee, S.-H., Park, H., & Ko, E. S. (2020). Radiomics in breast imaging from techniques to clinical applications: A review. *Korean Journal of Radiology*, *21*(7), 779–792. https://doi.org/10.3348/kjr.2019.0855

Lévano Loayza, S. A., Sovero Gaspar, A. T., Lévano Loayza, S. A., & Sovero Gaspar, A. T. (2020). Evaluación anatómica de la articulación temporomandibular mediante resonancia magnética. artículo de revisión [Publisher: Universidad Peruana Cayetano Heredia]. *Revista Estomatológica Herediana*, *30*(4), 285–293. https://doi.org/10.20453/reh.v30i4.3882

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., & Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, *114*(2), 97–109. https://doi.org/10.1007/s00401-007-0243-4

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. Retrieved November 5, 2023, from http://jmlr.org/papers/v9/vandermaaten08a.html

Maldonado, S., Flores, A., & Weber, R. (2015). Selección de atributos y support vector machines adaptado al problema de fuga de clientes. *Revista Ingeniería de Sistemas*, *24*, 85–107.

Mannil, M., von Spiczak, J., Manka, R., & Alkadhi, H. (2018). Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: Unveiling the invisible. *Investigative Radiology*, *53*(6), 338–343. https://doi.org/10.1097/RLI.0000000000000448

Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour [Artwork Size: 24 p. Medium: application/pdf Publisher: ETH Zurich], 24 p. https://doi.org/10.3929/ETHZ-B-000454597

Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., & Cook, G. (2020). Introduction to radiomics. *Journal of Nuclear Medicine*, *61*(4), 488–495. https://doi.org/10.2967/jnumed.118.222893

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Medelo Ballesteros, H., Espinosa-Castro, J.-F., Rodríguez, J. E., Pazmiño Calero, A., Palacios Serrano, S., Pérez Granja, A., Ruiz Chávez, P., Añez, R., & Bermúdez, V. (2019). Valoración mediante resonancia magnética del glioblastoma [Accepted: 2020-01-28T15:44:07Z Publisher: Sociedad Venezolana de Farmacología y Terapéutica]. *Archivos Venezolanos de Farmacología y Terapéutica AVFT*. Retrieved November 4, 2023, from https://bonga.unisimon.edu.co/handle/20.500.12442/4562

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Molnar, C. (2020). *Interpretable machine learning*. Leanpub. https://books.google.cl/books?id=jBm3DwAAQBAJ

Neisius, U., El-Rewaidy, H., Nakamori, S., Rodriguez, J., Manning, W. J., & Nezafat, R. (2019). Radiomic analysis of myocardial native t1 imaging discriminates between hypertensive heart disease and hypertrophic cardiomyopathy. *JACC. Cardiovascular imaging*, *12*(10), 1946–1954. https://doi.org/10.1016/j.jcmg.2018.11.024

Newton, H. (2016). *Handbook of neuro-oncology neuroimaging*. Elsevier Science. https://books.google.cl/books?id=yY2rCQAAQBAJ

Nguyen, N., Redfield, J., Ballo, M., Michael, M., Sorenson, J., Dibaba, D., Wan, J., Ramos, G. D., & Pandey, M. (2021). Identifying the optimal cutoff point for MGMT promoter methylation status in glioblastoma. *CNS oncology*, *10*(3), CNS74. https://doi.org/10.2217/cns-2021-0002

Nieto, P. (2016). *Aplicación de algoritmos de aprendizaje maquina para el analisis predictivo del mercado forex* [Doctoral dissertation, Universidad de Sevilla].

Orlhac, F., Eertink, J. J., Cottereau, A.-S., Zijlstra, J. M., Thieblemont, C., Meignan, M., Boellaard, R., & Buvat, I. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, *63*(2), 172–179. https://doi.org/10.2967/jnumed.121.262464

Ostrom, Q. T., Gittleman, H., Stetson, L., Virk, S. M., & Barnholtz-Sloan, J. S. (2015). Epidemiology of gliomas. *Cancer Treatment and Research*, *163*, 1–14. https://doi.org/10.1007/978-3-319-12048-5_1

Pálsson, S., Cerri, S., & Van Leemput, K. (2022). Prediction of MGMT methylation status of glioblastoma using radiomics and latent space shape features. In A. Crimi & S. Bakas (Eds.), *Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 222–231). Springer International Publishing. https://doi.org/10.1007/978-3-031-09002-8_20

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pegg, A. E. (2000). Repair of o(6)-alkylguanine by alkyltransferases. *Mutation Research*, *462*(2), 83–100. https://doi.org/10.1016/s1383-5742(00)00017-x

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6639–6649.

Pszczolkowski, S., Manzano-Patrón, J. P., Law, Z. K., Krishnan, K., Ali, A., Bath, P. M., Sprigg, N., & Dineen, R. A. (2021). Quantitative CT radiomics-based models for prediction of haematoma expansion and poor functional outcome in primary intracerebral haemorrhage. *European Radiology*, *31*(10), 7945–7959. https://doi.org/10.1007/s00330-021-07826-9

Quillien, V., Lavenu, A., Karayan-Tapon, L., Carpentier, C., Labussière, M., Lesimple, T., Chinot, O., Wager, M., Honnorat, J., Saikali, S., Fina, F., Sanson, M., & Figarella-Branger, D. (2012). Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, MethyLight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze o6-methylguanine-DNA-methyltranferase in a series of 100 glioblastoma patients. *Cancer*, *118*(17), 4201–4211. https://doi.org/10.1002/cncr.27392

Radiological Society of North America. (2021). Brain tumor ai challenge 2021. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/brain-tumor-ai-challenge-2021

Ramírez Giraldo, J. C., Arboleda Clavijo, C., & McCollough, C. H. (2008). TOMOGRAFÍA COMPUTARIZADA POR RAYOS X: FUNDAMENTOS Y ACTUALIDAD [Publisher: Escuela de Ingeniería de Antioquia CES-, Universidad CES]. *Revista Ingeniería Biomédica*, *2*(4), 54–66. Retrieved November 5, 2023, from http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S1909-97622008000200008&lng=en&nrm=iso&tlng=es

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? a lesson from an explainable AI competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Saeed, N., Hardan, S., Abutalip, K., & Yaqub, M. (2022, February 26). Is it possible to predict MGMT promoter methylation from brain tumor MRI scans using deep learning models? https://doi.org/10.48550/arXiv.2201.06086

Saidman, J. M., Falconi, M., Arenaza, D. P. d., Gutiérrez-Vallecillo, M. J., Gentile, E., & Ulla, M. (2022). Utilidad de la tomografía computarizada multicorte para la estimación de la composición del derrame pericárdico. *Revista argentina de radiología*, *86*(3), 199–210. Retrieved November 5, 2023, from https://www.redalyc.org/journal/3825/382573236006/html/

Satake, H., Ishigaki, S., Ito, R., & Naganawa, S. (2022). Radiomics in breast MRI: Current progress toward clinical application in the era of artificial intelligence. *La Radiologia Medica*, *127*(1), 39–56. https://doi.org/10.1007/s11547-021-01423-y

Shapley, L. S. (2016, March 2). A value for n-person games. In *Contributions to the theory of games II* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Shukla, G., Alexander, G. S., Bakas, S., Nikam, R., Talekar, K., Palmer, J. D., & Shi, W. (2017). Advanced magnetic resonance imaging in glioblastoma: A review. *Chinese Clinical Oncology*, *6*(4), 40. https://doi.org/10.21037/cco.2017.06.28

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., Howe, F. A., & Ye, X. (2018). Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels. *Computer Methods and Programs in Biomedicine*, *157*, 69–84. https://doi.org/10.1016/j.cmpb.2018.01.003

Stupp, R., Hegi, M. E., Mason, W. P., van den Bent, M. J., Taphoorn, M. J., Janzer, R. C., Ludwin, S. K., Allgeier, A., Fisher, B., Belanger, K., Hau, P., Brandes, A. A., Gijtenbeek, J., Marosi, C., Vecht, C. J., Mokhtari, K., Wesseling, P., Villa, S., Eisenhauer, E., ... Mirimanoff, R.-O. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year

analysis of the EORTC-NCIC trial. *The Lancet Oncology*, *10*(5), 459–466. https://doi.org/10.1016/S1470-2045(09)70025-7

Suh, C. H., Kim, H. S., Jung, S. C., Choi, C. G., & Kim, S. J. (2018). Clinically relevant imaging features for MGMT promoter methylation in multiple glioblastoma studies: A systematic review and meta-analysis. *AJNR. American journal of neuroradiology*, *39*(8), 1439–1445. https://doi.org/10.3174/ajnr.A5711

Torres Courchoud, I., & Pérez Calvo, J. I. (2016). Biomarcadores y práctica clínica [Publisher: Gobierno de Navarra. Departamento de Salud]. *Anales del Sistema Sanitario de Navarra*, *39*(1), 5–8. Retrieved November 5, 2023, from https://scielo.isciii.es/scielo.php?script=sci_abstract&pid=S1137-66272016000100001&lng=es&nrm=iso&tlng=es

Tran, T. (2020). *Explainability vs. interpretability and methods for models improvement* [Doctoral dissertation, Universidad de Hamburgo].

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ*, *2*, e453.

van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J.-C., Pieper, S., & Aerts, H. J. W. L. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, *77*(21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, *11*(1), 91. https://doi.org/10.1186/s13244-020-00887-2

Wagner, M. W., Namdar, K., Biswas, A., Monah, S., Khalvati, F., & Ertl-Wagner, B. B. (2021). Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology*, *63*(12), 1957–1967. https://doi.org/10.1007/s00234-021-02813-9

Walt, S. v. d., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python [Publisher: PeerJ Inc.]. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xie, H., Tubbs, R., & Yang, B. (2015). Detection of MGMT promoter methylation in glioblastoma using pyrosequencing. *International Journal of Clinical and Experimental Pathology*, *8*(1), 636–642. Retrieved November 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4348884/

Zhou, M., Scott, J., Chaudhury, B., Hall, L., Goldgof, D., Yeom, K. W., Iv, M., Ou, Y., Kalpathy-Cramer, J., Napel, S., Gillies, R., Gevaert, O., & Gatenby, R. (2018). Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches. *AJNR. American journal of neuroradiology*, *39*(2), 208–216. https://doi.org/10.3174/ajnr.A5391

Agradecimientos

Esta tesis fue apoyada por el proyecto FONDECYT 1231268, "A Physics-informed neural network framework for robust and precise estimation of the brain hemodynamic response function in functional magnetic resonance imaging at low magnetic field", y por el proyecto FONDECYT Regular 1221938, "An Explainable Deep Neuro-Fuzzy Inference System for the segmentation of brain tumors on multi-contrast magnetic resonance imaging".