PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE FÍSICA
INSTITUTO DE FÍSICA

Predicción semiautomatizada de respuesta a quimioterapia neoadyuvante en pacientes con cáncer de mama: protocolo de segmentación y modelo radiómico-clínico con imágenes de resonancia magnética

Por

M. Belén Ramírez Bunster

Informe de Tesis presentado a la Facultad de Física de la Pontificia Universidad Católica de Chile, como requisito para optar al grado académico de Magíster en Física Médica

Profesor Guía	:	Dr. Paola Caprile (PUC)
Comisión Revisora	:	Dr. Daniela Cornejo (PUC)
	:	Dr. Ignacio Espinoza (PUC)

Noviembre de 2023

Santiago, Chile

Agradecimiento

Quiero expresar mi más sincero agradecimiento a la profesora Paola Caprile por toda su motivación, paciencia, comprensión y ayuda en este proceso de tesis; difícil es describir cuán importante fue su apoyo.

Me gustaría ofrecer un agradecimiento especial a Catalina Vial por su tiempo y por su orientación respecto al reconocimiento de lesiones en las imágenes y otros, información que me fue de gran ayuda para la etapa de elaborar y establecer los protocolos de segmentación.

Quiero agradecer también al cuerpo académico del programa por darme la especial oportunidad de dar término a esta formación de magíster.

Resumen

El tratamiento de quimioterapia neoadyuvante (NACT, por sus singlas en inglés) es una de las estrategias utilizadas en pacientes con cáncer de mama para conseguir una reducción del tamaño del tumor, facilitando su posterior eliminación y permitiendo una cirugía menos agresiva y más conservadora, que no provoque mayor impacto psicológico en la paciente. Imágenes de resonancia magnética (MRI) posibilitan evaluar el estado de la lesión en distintas etapas del tratamiento, mientras que la eficacia de este último se mide con exámenes patológicos del tejido, logrando una respuesta completa patológica (pCR) como resultado positivo a la ausencia residual de la enfermedad.

En este trabajo se desarrolló un código computacional en lenguaje Python para construir modelos de predicción de pCR en base a información clínica y radiómica extraída de imágenes MRI de 59 pacientes con cáncer de mama sometidas a NACT. Específicamente, se utilizaron las secuencias de imágenes T1w y con contraste dinámico mejorado (DCE, por sus siglas en inglés), elaborando un protocolo de segmentación semiautomatizado del tumor y del parénquima de la mama lesionada para el posterior análisis y selección de atributos (*features*).

Se construyeron modelos uni- y multi-variados basados en *Machine Learning* utilizando distintos algoritmos supervisados de clasificación y, mediante la técnica de validación cruzada *k-fold* estratificada con repetición con $k = 3$ y $n = 500$ repeticiones, se evaluaron las métricas AUC y *Accuracy* para analizar el rendimiento de éstos como predictores de pCR del tumor a la terapia neoadyuvante en pacientes con cáncer de mama.

El modelo mejor evaluado se logró con el algoritmo de regresión logística como estimador y la configuración de *features*: *Kurtosis* (f.O), *ChusterShade* (glcm) y *Shpericity* (Sh), relativos al tumor; *10Percentil* (f.O), *GrayLevelNonUniformity* (glszm) y *DependenceVariance* (gldm), relativos al parénquima mamario; y el estado hormonal HR, como atributo clínico. El rendimiento AUC del modelo fue $0.87\ (0.08)\ [0.74; 1.00]$, con un *Accuracy* $0.76\ (0.08)$, desempeño equivalente a otros modelos desarrollados al realizar la comparación mediante análisis bayesiano, señalados en los capítulos de resultados y discusión.

Si bien los resultados son concordantes con otros estudios reseñados en la tesis, se identificaron algunas posibles limitaciones para la optimización del rendimiento del predictor, siendo los

principales factores el reducido tamaño de la cohorte y, por ende, la necesidad de un set de datos independiente para el testeo del modelo.

Índice

Índice de Figuras

Índice de tablas

CAPÍTULO 1

Introducción

Una de las principales causas de muerte actualmente en el mundo es por enfermedad del cáncer, presentando una tendencia significativa al alza en incidencia sobre la población mundial desde hace algunas décadas[1], debido principalmente al mejor registro de información y a los avances en su detección. En particular, el cáncer de mama es el cáncer diagnosticado con mayor incidencia, junto al de pulmón, y prevalencia sobre la población mundial, y el de mayor mortalidad en mujeres. [1]

Según los últimos registros de la Organización Mundial de la Salud (WHO, por sus siglas en inglés) y las estimaciones GLOBOCAN 2022 [2] desarrolladas por la Agencia Internacional para la Investigación del Cáncer (IARC), en el año 2022 el 15,4% de muertes por cáncer en mujeres (4.3 millones) fue por cáncer de mama. En cuanto a la prevalencia del cáncer en la población femenina a 5 años (27.8 millones), el de mama se tasó en un 29,3% con el mayor porcentaje. En el caso de Chile, estas mismas cifras en dicho período se estimaron en un 12,4% y un 33.7%, respectivamente, según el registro de DEIS (Departamento de Estadísticas e Información de Salud) del Ministerio de Salud[2]. [3]

Conforme al tipo de cáncer de mama diagnosticado (según estadío, fenotipo molecular y composición genética), el administrar un tratamiento de quimioterapia neoadyuvante (NACT) permite reducir el tamaño del tumor, simplificando su extirpación en cirugía con mejores resultados clínicos. Lograr una respuesta patológica completa (pCR) o positiva a NACT, evidenciando ausencia de enfermedad tumoral microscópica residual en la zona afectada, se asocia a una mayor probabilidad de conservar la mama, y a una supervivencia general y libre de enfermedad significativamente mejor. [4, 5]

[1] https://gco.iarc.fr/overtime/en

[2] https://deis.minsal.cl/

Es por ello, que el control de la respuesta del tumor primario al tratamiento tiene valor predictivo y que una evaluación temprana de la respuesta tumoral al tratamiento podría implicar cambios en la planificación del tratamiento para conducir a mejores resultados clínicos y cambios en la calidad de vida del paciente.

En la literatura se encuentran variados estudios respecto a la predicción de pCR a NACT en pacientes de cáncer de mama en base a parámetros cinéticos, farmacocinéticos y estadísticos (primer orden), tanto del tumor como del parénquima de la mama, obtenidos a partir de imágenes de resonancia magnética (MRI) en sus distintas secuencias de adquisición. [6–8] Sin embargo, la predicción de pCR basada en características cuantitativas (*features*) extraídas de imágenes MRI, utilizando *Machine Learning* (*ML*) para la construcción de los modelos predictivos, ha resultado ser cada vez más de gran interés dado los resultados significativos que se han evidenciado en estudios bibliográficos recientes. [9, 10] Para ello, las secuencias de imágenes de resonancia magnética mejorada con contraste han demostrado proporcionar mejor información sobre la fisiología del tumor, así como también información de su textura en la imagen.

Los principales desafíos que se han presentado en este tema para lograr un buen rendimiento de predicción de pCR al tratamiento neoadyuvante, con el objetivo de proporcionar un marcador más temprano y preciso de la respuesta tumoral al tratamiento, son: la selección adecuada de *features*, una segmentación precisa del tumor y/o área de interés y un algoritmo entrenado y ajustado óptimamente, además de una cohorte numerosa de pacientes; y son éstos, precisamente, los que motivan el desarrollo de esta tesis.

1.1 Objetivos

1.1.1 Objetivo general

Generar un modelo predictivo de clasificación, basado en información cuantitativa de imágenes de resonancia magnética e información clínica, para respuesta patológica tumoral (pCR / no-pCR) de pacientes con cáncer de mama sometidas a tratamiento de quimioterapia neoadyuvante.

1.1.2 Objetivos específicos

- Desarrollar y establecer un protocolo de segmentación semiautomático para los volúmenes de interés: tumor y parénquima mamario, a partir de las imágenes de resonancia magnética con secuencias ponderada en T1 y dinámica de contraste (DCE) previo al tratamiento neadyuvante.
- Extraer los atributos cuantitativos estandarizados de los volúmenes de interés: tumor y parénquima mamario, en la secuencia de imágenes generada como resultado de la sustracción entre la primera secuencia adquirida post-contraste y la secuencia pre-contraste.
- Generar un código en lenguaje Python que implemente: el análisis estadístico de los datos clínicos y *features* extraídos de las imágenes, y el desarrollo y configuración de distintos modelos de predicción de pCR (uni- y multi-variados) utilizando algoritmos de aprendizaje automático supervisado de clasificación, con su consiguiente evaluación mediante métricas e indicadores establecidos.
- Identificar los modelos más relevantes y compararlos entre sí mediante métricas e indicadores establecidos utilizando análisis bayesiano.
- Comparar y analizar los dos conjuntos de *features* seleccionados: manualmente mediante estadística inferencial y automáticamente mediante LASSO (*Least Absolute Shrinkage and Selection Operator*, por sus siglas en inglés).
- Identificar, entre los analizados, el mejor modelo de predicción de pCR en pacientes con cáncer de mama tratados con NACT.

La estructura del presente informe se organiza de la siguiente manera. El capítulo 2 presenta y detalla el marco teórico bajo el cual se desarrolla el estudio; mientras que en el capítulo 3 se describe la cohorte de pacientes, las secuencias y procesamiento de imágenes MRI a utilizar, la metodología para la extracción y selección de *features*, y el procedimiento para la aplicación de *ML*. En los capítulos 4 y 5 se presentan los resultados del estudio, su análisis y discusión, junto con las limitaciones, respectivamente. Finalmente, en el capítulo 6 se señalan las conclusiones y propuestas de trabajo futuro referente al tema tratado.

CAPÍTULO 2

Marco teórico

Este capítulo se organiza de la siguiente manera. En la primera parte se realiza una reseña general sobre el cáncer de mama, de su clasificación de acuerdo con el interés del presente estudio, y de la respuesta tumoral frente a un tratamiento neoadyuvante. A continuación, se introducen las imágenes médicas de resonancia magnética según las secuencias de adquisición que serán utilizadas como herramientas de estudio, junto a la descripción de *Radiomics*. En la tercera sección se detalla el análisis estadístico a utilizar para la selección de *features*; mientras que el contenido en la última sección se dedica a explicar el aprendizaje automático aplicado según los objetivos de esta tesis.

2.1 Cáncer de mama

El cáncer de mama puede presentarse tanto en mujeres como en hombres; sin embargo, éste es mucho más prevalente en mujeres, y en edad avanzada (sobre los 40 años). Por esta razón, es la población femenina el foco e interés de la mayor parte de los estudios referentes al tema.

La composición de la mama femenina, como se observa en la Figura 2.1, consiste básicamente en tejido adiposo y tejido fibroglandular, también llamado parénquima mamario. La proporción entre ellos es variable según la edad, principalmente, y determina la densidad de la mama. El parénquima mamario contiene lo que es tejido conectivo, los lóbulos y los lobulillos conectados mediante ductos.

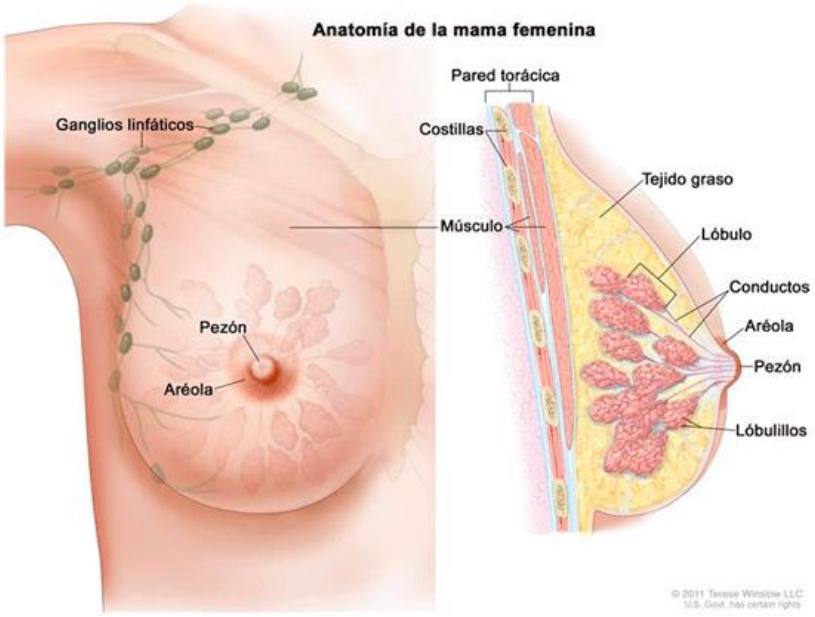


Figura 2.1: Anatomía de la mama femenina[3]

El cáncer de mama es un tipo de cáncer que se forma en las células mamarias, específicamente en las del revestimiento (epitelio) de los conductos o en las de los lóbulos del tejido glandular de la mama, cuando éstas se ven afectadas y comienzan a crecer de forma descontrolada originando un nódulo o tumor. Este cáncer *in situ* confinado al conducto o lóbulo, puede avanzar e invadir el tejido circundante, volviéndose de carácter invasivo. Con el tiempo puede propagarse a los ganglios linfáticos cercanos situados en la axila, e incluso a otros órganos del cuerpo más distantes. Esta evolución es la que da cuenta del estadío o etapa clínica del tumor, representados en la Figura 2.2.

Figura 2.2: Etapas clínicas del cáncer de mama, vinculadas a la ubicación y diseminación del tumor.

[3] https://www.teresewinslow.com/#/breast/

Dependiendo de estas características de origen del tumor en su diagnóstico, el cáncer es catalogado como carcinoma ductal (*in situ* o invasivo), carcinoma lobular (*in situ* o invasivo), o bien, sarcoma. Este último se origina en el tejido conectivo de la mama, compuesto de músculos, grasa y vasos sanguíneos.

Según las estadísticas 2023 de WHO [11], existe una considerable diferencia entre los países de ingresos elevados y los de ingresos bajos y medios en lo que respecta al cáncer de mama. En los primeros, el número de casos de supervivencia a 5 años es elevado gracias a los programas de detección temprana de este cáncer para su diagnóstico, en tanto que los países con menores ingresos económicos son los que presentan más elevados los casos de muerte. Por otra parte, mientras que el sexo (femenino) y la edad constituyen los mayores factores de riesgo para padecer esta enfermedad, existen otros tantos que aumentan dicho riesgo: la obesidad, el consumo perjudicial de alcohol, los antecedentes familiares de cáncer de mama, el historial de exposición a radiación, el historial reproductivo, el consumo de tabaco y la terapia hormonal posterior a la menopausia.

2.1.1 Clasificación molecular del cáncer de mama

A partir de un procedimiento de biopsia, luego de un diagnóstico confirmado de la enfermedad mediante imágenes, se analizan los genes de una muestra de tejido tumoral, proporcionando una clasificación más precisa de ésta. [12]

Se establecen cuatro tipos de cáncer de mama según análisis de inmunohistoquímica, dando cuenta del estado hormonal y de la composición genética respecto a la proteína del factor de crecimiento epidérmico humano 2 de las células cancerosas.

Grupo 1 – Luminal A

El tumor presenta receptor hormonal positivo (HR+), esto tanto para el receptor de estrógeno (ER+) como para el de progesterona (PR+), y no se expresan receptores de factor de crecimiento epidérmico humano 2 (HER2-).

Grupo 2 – Luminal B

Éste incluye tumores con receptor hormonal positivo (HR+), con la implicancia que los receptores de estrógeno son positivos (ER+), pero los de progesterona pueden ser tanto negativos como positivos. A su vez, los receptores del factor HER2 pueden ser tanto positivos (HER2+) como negativos (HER2-).

Grupo 3 – Enriquecido con HER2

El tumor es positivo para el receptor del factor de crecimiento epidérmico humano 2 (HER2+) y negativo para los receptores hormonales (HR-). Esto último implica que es negativo tanto para el receptor de estrógeno (ER-) como para el de progesterona (PR-), resultando no - luminal.

Grupo 4 – Tipo basal o triple negativo

El tumor no expresa receptor alguno, resultando negativo para los de estrógeno (ER-), progesterona (PR-) y factor de crecimiento epidérmico humano 2 (HER2-).

Conocer el estado de los receptores (presencia o ausencia) y la composición genética del cáncer en un paciente permite a los médicos elegir el tratamiento más eficaz para dicho tipo de cáncer específico, y el cual complementa a la cirugía oncológica (extirpación del tumor), con el objetivo de lograr finalmente la eliminación total del tumor.

2.1.2 NACT y respuesta del tumor al tratamiento

Las probabilidades de supervivencia al cáncer de mama pueden ser altas si la enfermedad es detectada de forma temprana y se suministra un tratamiento adecuado en combinación con la cirugía, que puede consistir en radioterapia, terapia endocrina (hormonal), quimioterapia, terapia dirigida o inmunoterapia, o combinaciones de éstas.

El tratamiento puede ser previo (‘neoadyuvante’) y/o posterior (‘adyuvante’) a la cirugía según los objetivos de su aplicación, los cuales son principalmente: frenar el avance del tumor hacia el pecho, ganglios linfáticos y áreas circundantes, tratar y/o reducir el riesgo de metástasis, reducir el tamaño del tumor, o destruir cualquier resto de células cancerosas, disminuyendo la probabilidad de recurrencia de la enfermedad.

En particular, el tratamiento de quimioterapia neoadyuvante (NACT, por sus siglas en inglés) es estándar para el cáncer de mama localmente avanzado e inflamatorio y es utilizado en pacientes con cáncer de mama catalogado en cualquiera de los 4 grupos previamente señalados, especialmente en los que no expresan receptores hormonales (estrógeno / progesterona); estos últimos se privan del beneficio de una terapia endocrina.

La evaluación del tumor por su respuesta y evolución frente al tratamiento neoadyuvante administrado se mide utilizando métodos como respuesta clínica, criterios de evaluación de respuesta en tumores sólidos (RECIST, por sus siglas en inglés) o mediante el análisis patológico que presenta el tejido luego de la cirugía planificada. Se considera 'respuesta patológica completa' (pCR) cuando no se obtiene enfermedad invasiva residual en la mama o en los ganglios linfáticos postcirugía, admitiendo la presencia de células tumorales residuales solo *in situ*. No obstante, sobre esta última condición existen diferencias entre algunos autores. [13] Por el contrario, de no cumplirse las condiciones mencionadas, la respuesta tumoral al tratamiento se considera entonces no completa (no-pCR).

2.2 Imágenes médicas

El cáncer de mama se presenta de diversas formas y síntomas, como por ej: un nódulo o engrosamiento de la mama, alteraciones en tamaño, forma o aspecto de ésta y/o del pezón, y enrojecimiento o grietas en la piel. Un examen médico completo y temprano, junto con la obtención de imágenes de la mama y biopsia del tejido para determinar si la masa es maligna o benigna, permite un pronto diagnóstico y tratamiento, aumentando las posibilidades de una cirugía conservadora y curación de la enfermedad.

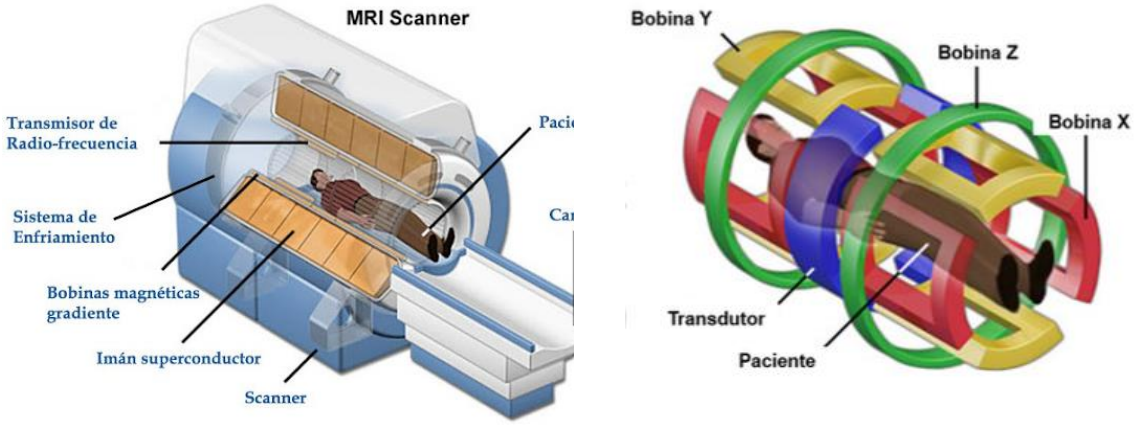
Después de un resultado positivo para cáncer mediante una biopsia o imágenes de mamografía, se le realizan a la paciente una serie de imágenes de resonancia magnética en distintos modos de adquisición para observar el alcance y características en detalle de la enfermedad.

2.2.1 Imágenes de Resonancia Magnética

Las imágenes de resonancia magnética (MRI) corresponden a una tecnología de imágenes no invasiva, sin el uso de radiación ionizante y basada en la estimulación y detección del cambio de dirección del eje de rotación de los protones contenidos en el agua de los tejidos vivos, que produce imágenes anatómicas tridimensionales de alta resolución como herramienta para detectar enfermedades, realizar un diagnóstico y/o monitorear tratamientos.[4]

[4] https://www.nibib.nih.gov/

En la Figura 2.3 se ilustra el interior de un scanner con los componentes principales para la generación de las señales y posterior reconstrucción de la imagen.



a. b.

Figura 2.3: Diagrama del interior de un scanner de resonancia magnética; a.) componentes y su distribución en el interior; b.) distribución detallada de las bobinas magnéticas de gradiente en cada eje para la codificación espacial de las señales según posición. (Coyne, 2012)

El paciente se sitúa al interior del scanner de resonancia magnética conformado por un imán principal de gran campo magnético que actúa sobre la muestra. En detalle, el elemento sobre el cual opera el campo magnético configurado es el átomo de hidrógeno (^{1}H), por abundar de forma natural en las personas, específicamente en la grasa y en el agua contenida en los tejidos blandos (órganos, entre otros) y por su particularidad de poseer un solo protón (impar) en su núcleo, con un momento angular intrínseco fijo de *spin* igual a ½. Sin la presencia de un campo magnético externo $\vec{B}$, los *spins* apuntan en direcciones aleatorias, por lo que la suma total de sus momentos magnéticos es igual a cero. Considerando entonces estas propiedades, la formación de imágenes por resonancia magnética se resume en 4 etapas generales: alineación de los *spins*, excitación de estos mismos, lectura de la señal, y reconstrucción de la imagen.[5]

Al aplicar un campo magnético externo constante $\overrightarrow{B_0}$, de magnitud según diseño del scanner y sobre un eje z, los *spins* mencionados se alinean con éste en forma paralela (*spin up*) y antiparalela

[5] www.mri.cl; Uribe, S.: Imágenes de Resonancia Magnética Visión Rápida. Documento clase académica, Pontificia Universidad Católica de Chile.

(*spin down*) y 'precesan' en torno al eje vectorial de dicho campo en un ángulo fijo θ y a una frecuencia angular determinada ω_0, llamada frecuencia de Larmor, según:

$$\omega_0 = -\gamma B_0 \tag{2.1}$$

donde γ es la constante giromagnética del núcleo y B_0 es la intensidad del campo externo. Para los núcleos de hidrógeno (en los tejidos del paciente) en un scanner con campo magnético de 1.5 T se tiene $\gamma = 42.58\ MHz/T$ y, por ende, $\omega_0 = 63.87\ MHz$.

En la Figura 2.4 a. y b. se representa visualmente la información señalada; en el plano *xy* perpendicular al campo magnético, la orientación de los *spins* sigue siendo aleatoriamente dispersa cuya suma total de los momentos magnéticos es cero, mientras que en el eje *z* predominan las componentes de *spin*s orientados en paralelo por sobre antiparalelo, equivalentes a estados de menor y de mayor energía respectivamente. Como resultado se obtiene una magnetización total constante en el tiempo $\overrightarrow{M_0}$, proporcional a la cantidad de átomos de hidrógeno y a la magnitud del campo magnético externo e inversamente proporcional a la temperatura, en la misma dirección que el campo magnético externo $\overrightarrow{B_0}$, como se observa en la Figura 2.4c.

Esta disposición corresponde al estado de equilibrio para los protones ante estas condiciones, y es la posición a la que regresarán naturalmente los protones después de cualquier perturbación.

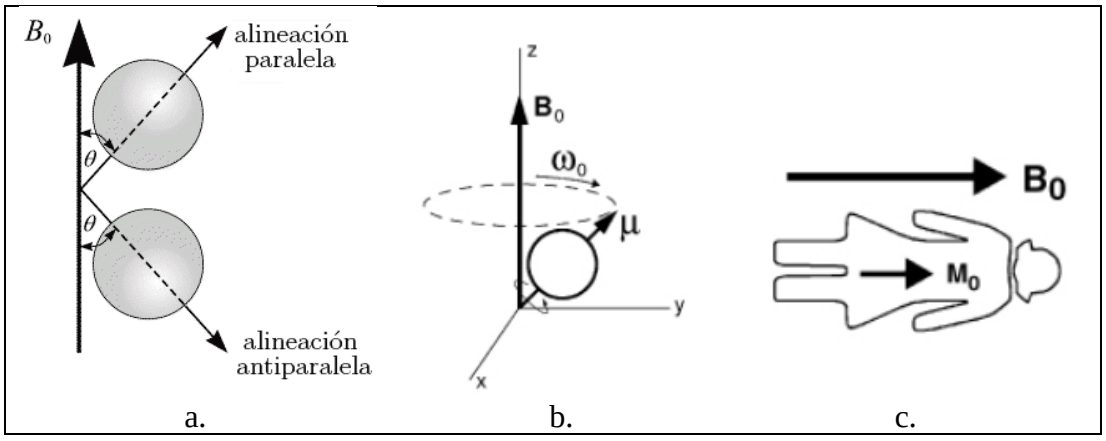


Figura 2.4: Polarización de los *spins* en un scanner de resonancia magnética. a.) Alineación paralela y antiparalela sobre la dirección del campo magnético externo; b.) precesión del *spin* en torno al eje de dicho campo, y donde μ corresponde al vector del momento magnético asociado; c) magnetización en los tejidos del cuerpo conforme a los átomos de hidrógeno. (Uribe, S.)

El núcleo de hidrógeno, en la configuración establecida previamente, es capaz de absorber y emitir energía de radiofrecuencia (RF), generando una señal detectable que es recibida por antenas situadas muy cerca de la anatomía que se examina y que contribuirá a la reconstrucción de la imagen. Por ello, se aplica a continuación un pulso de radiofrecuencia perpendicular a $\overrightarrow{B_0}$ y a la misma frecuencia de Larmor de los *spins* de interés, excitándolos y llevándolos hacia el plano transversal *xy* mediante un movimiento en espiral al seguir el campo rotatorio generado por el pulso de RF. La inclinación que alcanzan los *spins* respecto a la dirección del campo externo $\overrightarrow{B_0}$ (eje z) definen la nueva dirección del vector de magnetización $\vec{M}$ mediante el ángulo α (*flip angle*), el cual varía dependiendo de la amplitud y la duración del pulso de RF programado. Además, se aplican campos magnéticos adicionales no simultáneos, mediante las bobinas magnéticas especificadas en la Figura 2.3b., que varían linealmente en el espacio y que actúan como gradientes que modifican la magnitud del campo magnético principal según la posición, como se aprecia en la Figura 2.5 según coordenada. En específico, al gradiente aplicado sobre el eje perpendicular a los planos de corte deseados (x: sagital, y: coronal, z: axial) se le denomina 'gradiente de frecuencia' o 'de selección de corte', ya que provoca una modificación en la frecuencia de resonancia de los *spins* (frecuencia de Larmor), según Ec. (2.1), a lo largo de dicho eje. Este gradiente actúa en conjunto con el pulso de RF seteado en la frecuencia de Larmor asociado a cada plano transversal, produciendo una 'excitación selectiva' que permite seleccionar cada corte para la obtención de imágenes del volumen de interés.

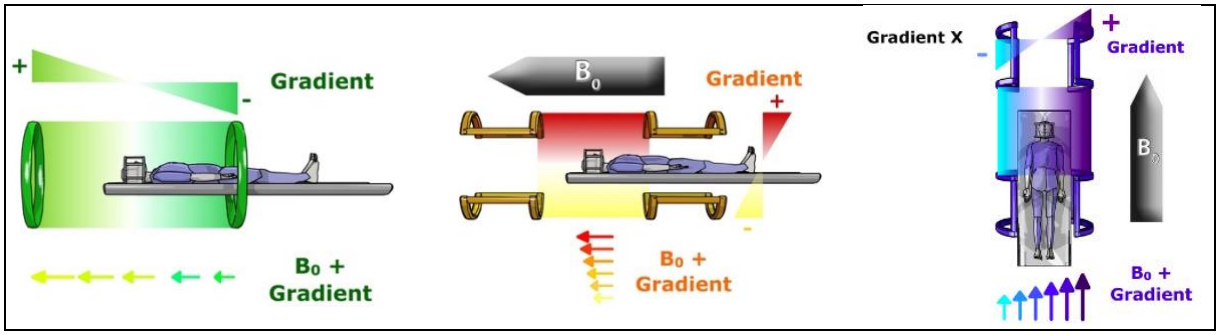


Figura 2.5: Gradientes que actúan sobre el campo magnético principal de forma lineal en cada una de las coordenadas x, y, z.[6]

[6] https://www.imaios.com/en

Tras la excitación, los *spins* retornan 'precesando' en torno a $\overrightarrow{B_0}$ (movimiento en espiral) a su estado de equilibrio hasta alinearse con el campo principal, de modo que la magnetización longitudinal (M_z) vuelve al equilibrio $\overrightarrow{M_0}$ y la magnetización transversal (M_{xy}) decae a cero. El tiempo que demora la magnetización longitudinal en recuperar un 63% de su valor inicial previo a la excitación se denomina T1 y el tiempo que demora la magnetización transversal en decaer a un 37% de su valor inicial tras la excitación, T2; información que se utiliza para lograr contrastes de tejidos en la imagen. Esta relajación del vector de magnetización induce un voltaje en la bobina receptora (antena) cuya señal contiene la información de todos los *spins* del plano o corte. Para localizar en el plano la señal (sinusoidal) emitida por cada *spin*, se aplican de forma contigua los otros dos gradientes previamente señalados, llamados gradientes de 'codificación de fase' y 'codificación de frecuencia', provocando una variación y dispersión de sus fases en un eje y de su frecuencia en el otro. Con ello, se da lugar a la lectura de la señal y su codificación espacial.

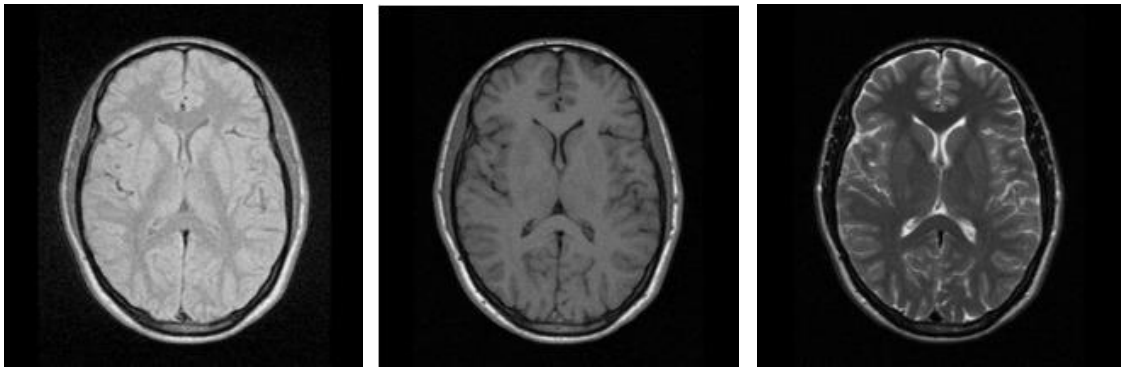
Finalmente, la reconstrucción de la imagen en cada corte se logra al aplicar la transformada de Fourier sobre el espacio-*k* generado por dichas señales.

El tiempo que transcurre entre los pulsos RF sucesivos para generar la imagen de un corte, por una parte, y el tiempo que transcurre entre la entrega de un pulso de RF y la recolección de la señal o eco, por otra, se denominan respectivamente: 'tiempo de repetición' (TR) y 'tiempo echo' (TE). Éstos, junto con el ángulo α (*flip angle*), corresponden a los parámetros básicos de secuencia que pueden modificarse para lograr los distintos tipos y grados de contraste; de este modo, para la misma configuración en un mismo scanner se logra uniformidad en la adquisición de imágenes, posibilitando la comparación directa de las intensidades de contraste entre las imágenes de diferentes pacientes.

Existen distintos tipos de secuencias de imágenes que se configuran al momento de su adquisición y que dependen de factores extrínsecos a los tejidos como los mencionados previamente, otorgando los distintos tipos de contraste según interés: ponderadas por densidad de protones (PD), ponderadas en T1 (T1w), ponderadas en T2 (T2w), ponderadas por difusión, sensibles al flujo, entre otras. En la Figura 2.6 se observa un ejemplo de tres secuencias distintas de adquisición para un mismo corte axial en el cerebro. Existen también otras opciones que complementan las propiedades en la adquisición, como atenuación en grasa o líquido, o la mejora de realce de intensidad mediante un agente de contraste.

T1w - MRI

En la secuencia de imágenes ponderadas en T1 brillan más los tejidos caracterizados por un tiempo de relajación longitudinal T1 corto al emitir una señal más intensa, por ejemplo: la grasa, sangre (sin flujo o uno muy lento), sustancias proteicas, melanina y agentes de contraste paramagnéticos. Por otro lado, no brillan o se ven más oscuros (hipointensos) elementos y estructuras como el aire, ligamentos, tendones, tejido fibroso, hueso cortical y flujo sanguíneo.



a. b. c.

Figura 2.6: Corte axial a la altura del cerebro, donde se aprecian los ventrículos y líquido cefalorraquídeo, adquirido en distintas secuencias: a.) secuencia PD; b.) secuencia T1w; c.) secuencia T2w. (Uribe, S.)

DCE - MRI

Las imágenes de resonancia magnética con contraste mejorado son favorables por sobre otras secuencias de adquisición para aplicaciones como: caracterizar lesiones con apariencias anormales, evaluar neoplasias especialmente neurológicas, ciertos tumores y diferentes tipos de malformaciones, entre otros. El contraste consiste en un agente paramagnético, basado en gadolinio, que se inyecta por vía intravenosa o intraarticular y que modifica la respuesta de las moléculas de agua (compuestas por hidrógeno) frente al campo magnético y a las ondas de radio en el scanner de MRI. Como consecuencia, se altera a conveniencia la señal de los tejidos donde se concentra.

Las imágenes de este tipo corresponden a secuencias T1w con supresión de grasa, de modo que la captación del contraste no quede camuflada, y se adquieren repetidamente en distintas instancias de tiempo: la primera secuencia consiste en la basal y se realiza previo a la inyección del contraste en el paciente, y al menos dos secuencias más se adquieren después de un determinado tiempo posterior a la administración del contraste, comúnmente a los 2 minutos y 7 minutos.

La alta resolución temporal que caracteriza la secuencia de imágenes ultrarrápidas de la resonancia magnética dinámica de contraste permite conocer el fenómeno fisiológico de la distribución del contraste por medio de la curva de 'intensidad de señal – tiempo' que se obtiene durante las adquisiciones. Del análisis de dicha curva se deduce información relevante sobre la vascularización y perfusión tisular, la permeabilidad capilar, y el espacio intersticial del tumor.

2.2.2 *Radiomics*: Extracción de información cuantitativa de imágenes

Radiomics se refiere al campo de estudio médico que contempla la extracción de alto rendimiento de características cuantitativas denominadas *features* de imágenes médicas estándar de alguna región de interés (ROI), como es el caso de alguna lesión o tumor, mediante algoritmos de caracterización de datos. [14] El análisis matemático avanzado de esta información permite apoyar decisiones clínicas, mejorando la precisión diagnóstica, pronóstica y predictiva.

Los *features*, tal como lo describe Mayerhoefer en la bibliografía referenciada, se agrupan dependiendo de la técnica de extracción o cálculo: basados en la forma del ROI (2D y/o 3D), basados en estadística de primer y segundo orden, y basados en estadística de orden superior que implican transformaciones y filtrados matemáticos de las imágenes. A continuación, se realiza una descripción general de ellos.

Features de forma (shape, Sh)

Éstos describen las propiedades geométricas de la región de interés en 2D y 3D, de modo que la información es independiente de la distribución de intensidad del nivel de gris de los pixeles que componen el ROI, y se extraen directamente del cálculo sobre los vóxeles que conforman la segmentación de dicha región. La compacidad (*Compactness*) y la esfericidad (*Sphericity*), que describen cómo la forma del ROI difiere de un círculo (2D) o esfera (3D), son dos ejemplos de este tipo de *features*.

Features de primer orden (f.O.)

Éstos corresponden a descriptores estadísticos basados en el histograma global del nivel de gris o intensidad de los valores de los pixeles o vóxeles, por si solos, que conforman el ROI. Algunos ejemplos son: la media (*Mean*), máximo valor (*Maximum*), mínimo valor (*Minimum*), percentiles de nivel de gris (*Percentile*), entropía (*Entropy*) y curtosis (*Kurtosis*); esta última hace referencia a la

forma de la distribución de los datos graficados, describiendo la falta de cola en ésta respecto a una distribución gaussiana debido a valores atípicos.

Features de segundo orden o textura

Estas características texturales proporcionan información sobre las intensidades y la geometría de los vóxeles en su conjunto a partir de relaciones entre vóxeles vecinos, y se subdividen en distintos tipos conforme a dicha relación.

Features derivados de la matriz de co-ocurrencia de nivel de gris (glcm): la matriz refleja las relaciones espaciales de pares de píxeles o vóxeles con intensidades de nivel de gris predefinidas, en diferentes direcciones y con una distancia determinada entre dichos píxeles o vóxeles. Algunos de estos *features* son: el promedio conjunto (*Joint Average*), el contraste (*Constrast*), la correlación (*Correlation*), la tendencia de cúmulos (*Cluster Tendency*) y la sombra de cúmulos (*Cluster Shade*); este último mide la asimetría y uniformidad de la matriz de co-ocurrencia, de modo que un tono de cúmulo más alto implica una mayor asimetría respecto a la media.

Features derivados de la matriz de longitud de secuencia de nivel de gris (glrlm): proporcionan información sobre la distribución espacial, cuantificando las secuencias de pixeles consecutivos con el mismo valor de gris según longitud, en una o más direcciones. La falta de uniformidad en los niveles de gris (*Gray Level Non-Uniformity*) y la varianza en la intensidad del nivel de gris en las secuencias (*Gray Level Variance*) son dos ejemplos que pertenecen a esta clasificación.

Features derivados de la matriz de zonas de tamaño de nivel de gris (glszm): se relacionan con la cuantificación de las zonas de nivel de grises en la región de interés en la imagen, donde las zonas corresponden al número de pixeles o vóxeles conectados en cualquier dirección con una misma intensidad de gris. Dos ejemplos son la entropía de zona (*Zone Entropy*) y la falta de uniformidad en los niveles de gris (*Gray Level Non-Uniformity*) de zona, que refleja la variabilidad de los valores de intensidad del nivel de gris en el ROI de tal manera que un valor más bajo indica una mayor homogeneidad en dichos valores.

Features derivados de la matriz de diferencia de tonos de grises vecinos (ngtdm): se basan en la cuantificación de la diferencia del nivel de gris entre un pixel o vóxel central y el valor promedio de su vecindario conectado. Los *features* incumben análisis de dependencia que reflejan heterogeneidad u homogeneidad, así como también de uniformidad que evalúa la similitud en los niveles de gris.

Features derivados de la matriz de dependencia del nivel de gris (gldm): la matriz en que se basan para su cálculo cuantifica las dependencias del nivel de gris, definidas como el número de vóxeles conectados que dependen del vóxel central, en el ROI de la imagen. Algunos de estos *features* son: énfasis en la pequeña dependencia (*Small Dependence Emphasis*) y alta dependencia (*Small Dependence Emphasis*), y la varianza de dependencia (*Dependence Variance*), que mide la variación en el tamaño de dependencia en la región de interés en la imagen.

Son 107 los *features* referentes a las categorías ya mencionadas que se encuentran incorporadas en la librería de código abierto *pyradiomics* [7] para *Python*, y los que fueron considerados en el cálculo de extracción de *features* en el presente estudio. Como parte del resultado, se obtiene también información general de la imagen completa y de la 'máscara' que contiene la segmentación, como el valor medio de intensidad de nivel de gris, el mínimo y el máximo valor, el número de vóxeles, entre otros.

2.3 Estadística analítica para la selección de *features*

La estadística analítica, cuyo rol es interpretar y hacer proyecciones, inferencias y comparaciones al analizar poblaciones, permite realizar deducciones y seleccionar variables que se asocien significativamente con cierto resultado en el ámbito clínico. Las herramientas que se utilizan para este objetivo son diversas, destacando las pruebas o test de hipótesis que se describen a continuación. [15]

2.3.1 Test de hipótesis de normalidad

Existen distintos tipos de pruebas de normalidad que cuantifican la probabilidad de que una muestra provenga de una distribución de tipo gaussiana, algunos basados en la distribución empírica y otros basados en regresión y correlación; y su elección queda condicionada a su vez por la cantidad de datos disponibles como muestra.

[7] https://pyradiomics.readthedocs.io/en/latest/features.html

Test de Shapiro-Wilk

Esta prueba plantea como hipótesis nula H_0 que la muestra en estudio procede de una población que distribuye normal con media μ y desviación estándar σ, y otra antagónica como hipótesis alternativa H_A; lo que se expresa por:

$$H_0 : X \sim N(\mu, \sigma^2)$$
$$H_A : X \nsim N(\mu, \sigma^2) \tag{2.2}$$

Para el testeo de la hipótesis se calcula el valor estadístico W, definido como:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.3}$$

donde x_i corresponde al valor de los datos según posición *i* luego de ser ordenados de menor a mayor, *n* indica el tamaño de muestra, $\bar{x}$ es el promedio de dicha muestra, y a_i son constantes tabuladas (Tablas Shapiro & Wilk) para cada posición *i*. El estadístico calculado *W* oscila entre 0 y 1.

Según los valores de *n* y *W*, se obtiene igualmente por tabulación el valor de probabilidad *p*, el cual se compara con un nivel de significancia estadística α para aceptar o rechazar la hipótesis nula según la evidencia que presentan los datos. Así, si:

$$p \leq \alpha : H_0 \text{ se rechaza; distribución no normal}$$
$$p > \alpha : H_0 \text{ no se rechaza; distribución normal} \tag{2.4}$$

Normalmente se asigna como valor un 1%, 5% o 10% a α, y por ende, un 99%, 95% o 90%, respectivamente, al nivel de confianza.

Gráficos cuantil-cuantil (QQ)

Los cuantiles corresponden a puntos tomados a intervalos regulares, que comprenden la misma proporción de datos, de la función distribución de una variable aleatoria.

Estos gráficos QQ son una herramienta para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación; en específico, permiten identificar si el conjunto de datos proviene de una distribución normal al comparar la dispersión del conjunto de cuantiles de la muestra versus la de una

distribución teórica normal. Si la muestra posee un comportamiento normal, entonces se observará una alineación de los datos a 45º respecto a los ejes.

2.3.2 Test de hipótesis de comparación de poblaciones

Este tipo de prueba contrasta para una variable aleatoria continua si dos muestras proceden de poblaciones equidistribuidas o no (independencia); o bien, en caso de variables discretas evalúa mediante test exacto si las proporciones de pacientes según clase son diferentes significativamente dependiendo del valor que éstas adquieran.

Se dispone de una variedad de pruebas para los dos casos mencionados y su elección se condiciona por: los tamaños de muestra a comparar, si estos datos provienen de un mismo conjunto de individuos o no (muestras dependientes o independientes), el tipo de distribución, y la información estadística conocida de éstas.

En términos generales, para dos muestras aleatorias e independientes con valor medio $\overline{x}_1$ y $\overline{x}_2$, y desviación estándar s_1 y s_2, que provienen de dos poblaciones con medias desconocidas μ_1 y μ_2 respectivamente, se establecen una hipótesis nula H_0 y una alternativa H_A que permiten comparar ambos parámetros desconocidos y determinar así si corresponden a poblaciones dependientes o no.

$$\begin{aligned} &H_0 : \mu_2 - \mu_1 = 0 \text{ , no hay diferencia entre ambas poblaciones} \\ &H_A : \mu_2 - \mu_1 \neq 0 \text{ , hay diferencia entre ambas poblaciones} \end{aligned} \tag{2.5}$$

Test t-Student

En este test paramétrico se calcula el valor estadístico t y los grados de libertad υ de la distribución *t-Student* descrita, dados por:

$$t = \frac{(\overline{x}_2 - \overline{x}_1) - (\mu_2 - \mu_1)}{\sqrt{({s_1}^2/n_1) + ({s_2}^2/n_2)}} \tag{2.6}$$

$$\upsilon = \frac{(\overline{x}_2 - \overline{x}_1) - (\mu_2 - \mu_1)}{\sqrt{({s_1}^2/n_1) + ({s_2}^2/n_2)}} \tag{2.7}$$

con $s_1 \neq s_2$, y donde n_1 y n_2 corresponden al tamaño de cada muestra, mientras que los demás parámetros ya fueron definidos anteriormente. Mediante la tabla de distribución de *t- Student* [8] y en función de los grados de libertad υ calculados y del nivel de significancia α evaluado, se precisa un

[8] https://www.tdistributiontable.com/

valor *t crítico* respecto al cual se compara el estadístico t. Si el primero es menor que el segundo, se infiere que el valor de probabilidad p es menor a α, rechazando en consecuencia H_0. De ello se concluye que para la variable analizada existe una diferencia estadísticamente significativa entre ambas muestras y, por lo tanto, son independientes.

Test Wilcoxon-Mann-Whitney

Esta prueba, llamada también test de suma de rangos de Wilcoxon, se presenta como una alternativa no paramétrica al test *t-Student* en los casos en que las dos muestras analizadas no provengan de una población con distribución normal.

Ambos grupos se combinan, enlistando los datos de la variable de menor a mayor, a los que se les asigna un índice o rango en orden ascendente. Como resultado, si los datos se mezclan aleatoriamente en dicho orden establecido, se deduce que éstos no presentan diferencia significativa y, por ende, provendrían de una misma población. Por el contrario, si existen conglomerados de datos de pacientes de una misma clase al combinar las muestras, entonces ambos grupos presentarían independencia para la variable evaluada.

En concreto, para analizar el rechazo o no de la hipótesis nula H_0: $(\mu_2 = \mu_1)$ se calcula el estadístico U de cada muestra:

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i \tag{2.8}$$

donde *=1;2* representa las dos muestras comparadas, n_i es el tamaño de la muestra *i*, y R_i corresponde a la suma de sus rangos según *i*. El estadístico menor entre los dos anteriores se compara con el correspondiente valor crítico $U(n_1, n_2, \alpha)$ registrado en las tablas estadísticas[9], rechazando la hipótesis nula en caso de que el primero sea mayor que el estadístico tabulado. Este paso es equivalente a calcular la probabilidad *p* a partir de las tablas y contrastarla con el nivel de significancia α escogido.

Test exacto de Fischer

En el caso de variables discretas, tal como se explica en el libro de Jim Frost previamente reseñado, se configuran primero las tablas de contingencia para clasificar los datos recopilados. Luego, el test

[9] https://real-statistics.com/statistics-tables/mann-whitney-table/

de Fisher se aplica para evaluar cuantitativamente la asociación entre dichas variables y determinar si ésta es significativa o no. En este caso, la desviación de la hipótesis nula o valor *p* que se calcula para aceptar o rechazar H_0 corresponde a un valor exacto.

2.3.3 Intervalos de confianza

Las observaciones de una variable aleatoria continua en una muestra pueden resumirse y representarse en un único valor mediante el promedio; sin embargo, resulta más representativo caracterizar los datos mediante un intervalo de valores entre los cuales es factible encontrar dicho parámetro de interés.

Bajo el fundamento del 'Teorema central del límite', un intervalo de confianza (I.C.) se determina a partir de definir una variable estandarizada *Z*, dada por la expresión:

$$Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}} \tag{2.9}$$

μ es la media poblacional, $\bar{X}$ corresponde a la media muestral con desviación estándar σ, y n es el tamaño de la muestra. Ésta sigue una distribución normal estándar con media igual a 0 y varianza igual a 1, dada por:

$$z = \frac{1}{\sqrt{2\,\pi}}\, e^{-x^2/2} \tag{2.10}$$

Se busca determinar dos valores ($\pm\, z_{\alpha/2}$) asociado al nivel de confianza evaluado, de tal manera que la probabilidad:

$$P = \left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \tag{2.11}$$

Mediante álgebra de probabilidades, se expresa finalmente el I.C. para la variable estudiada como:

$$\left(\bar{X} - z_{\alpha/2}\, \frac{\sigma}{\sqrt{n}}\, , \bar{X} + z_{\alpha/2}\, \frac{\sigma}{\sqrt{n}}\right) \tag{2.12}$$

y cuya representación gráfica se observa en la Figura 2.7.

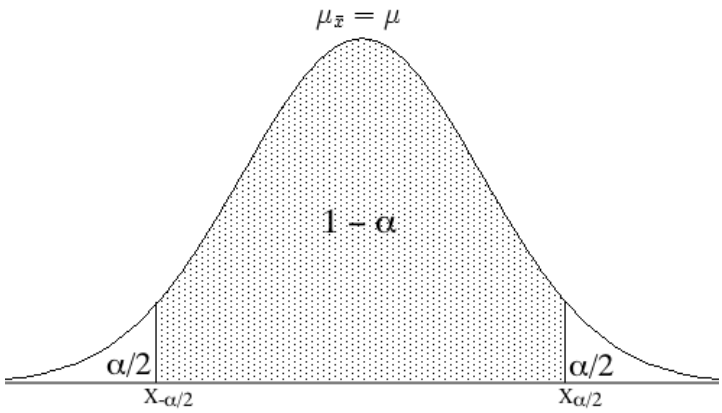


Figura 2.7: Distribución de una variable independiente continua con media muestral . El intervalo de confianza, determinado por el nivel de significancia escogido, limita el rango de valores de la distribución en que se estima el parámetro poblacional correspondiente. [10]

Cuando la muestra no es muy numerosa, se utilizan técnicas de remuestreo que permiten mejorar el cálculo del intervalo y la estimación poblacional.

Bootstrap

Este método de remuestreo consiste en obtener múltiples submuestras (*n*) de cierto tamaño *m* a partir de la muestra original de datos, luego calcular la media de la variable en cada una de ellas (submuestras), y conformar con estos resultados una nueva 'muestra alternativa' de tamaño *n* para la variable evaluada. Finalmente, se determina la media de la distribución que esta 'muestra alternativa' describe y su concerniente I.C., estimando así el parámetro de la población desde donde procederían los datos. Para este proceso se utiliza un muestreo con reposición, seleccionando en forma aleatoria un dato cada vez con la muestra global a disposición hasta completar cada submuestra, de tal forma que algunos datos no serán seleccionados y otros lo podrán ser más de una vez en cada muestreo. [16]

2.3.4 Test de correlación de variables

Un test de correlación de variables analiza la relación entre dos variables, independiente de su escala de medida, cuantificando la diferencia entre sus distribuciones o intensidad en su relación.

Se calcula un coeficiente de correlación que varía entre -1 y 1 como indicador, siendo positivo o negativo según si la dependencia es directa o inversamente proporcional entre las variables

[10] Imagen *ConfIntervNormalP.png*

analizadas. Un valor resultante de r igual a 0 implica que no hay dependencia, mientras que uno cercano a los extremos de su intervalo evidencia una fuerte correlación.

Si las variables se comportan según una distribución normal, se utiliza un test paramétrico para el análisis. En el test de correlación de Pearson, el coeficiente está dado por:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2.13}$$

donde x_i e y_i son los valores de ambos *features* evaluados, $\bar{x}$ e $\bar{y}$ sus medias respectivas, y n el tamaño de la cohorte.

En el caso contrario, es decir, para distribuciones no normales de la variable, se presenta el test no-paramétrico de Spearman como herramienta. El coeficiente se define entonces:

$$r = 1 - \frac{6 \cdot \sum_{i=1}^{n} D^2}{N\,(N^2-1)} \tag{2.14}$$

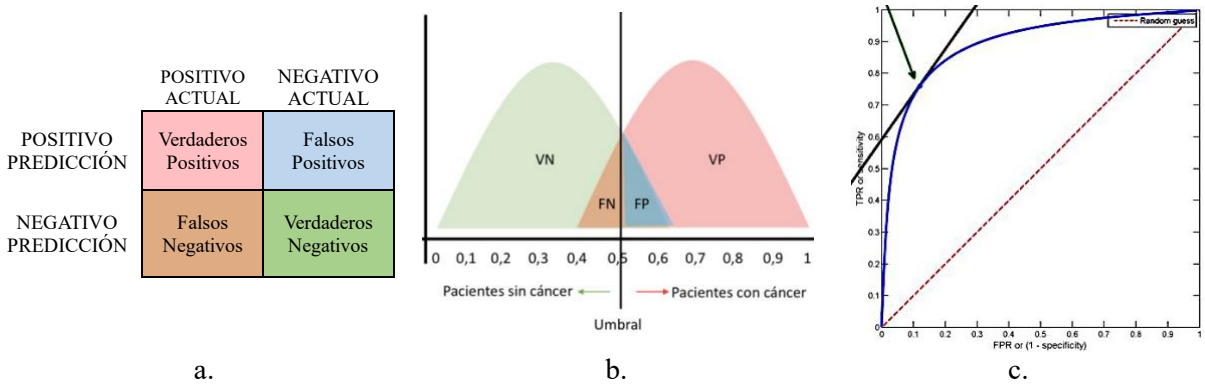
donde *N* es el número de parejas de datos, equivalente al tamaño de la cohorte, y *D* es la diferencia entre los estadísticos *orden(x)* y *orden(y)* para cada pareja de datos (*features*) comparados. Estos estadísticos se obtienen al ordenar de forma ascendente e independiente los datos para los dos *features* analizados y asignar a cada uno el valor correspondiente a su posición, comenzando desde 1.

2.3.5 Curva ROC y área bajo la curva AUC

La curva 'característica del operador de receptor' (ROC) consiste en una representación gráfica de la razón de verdaderos positivos (VPR) frente a la razón de falsos positivos (FPR) obtenidos al evaluar una variable aleatoria, o un conjunto de ellas, en un modelo clasificador o predictor binario, y en donde cada par (x, y) se obtiene al variar de forma continua el punto de corte o umbral de discriminación para dicha variable analizada. [17]

VPR o 'sensibilidad' es la proporción de los pacientes con clasificación 'positivo' correctamente identificados (VP) por el predictor sobre el total de la muestra de clase 'positivo', mientras que FPR o '(1 – especificidad)' es la proporción de los pacientes con clasificación 'negativo' incorrectamente identificados (FP) por el predictor sobre el total de la muestra de clase 'negativo'. En la Figura 2.8a. se representa mediante una matriz de confusión los resultados de una predicción binaria bajo un cierto valor de umbral de discriminación; mientras que en b. se observan las distribuciones de ambas

muestras según clase para la variable analizada y sus secciones equivalentes a VP y FP determinadas a partir de dicho umbral. Finalmente, en c. se grafica en color azul la curva ROC como resultado de variar el valor del umbral; se señala en el gráfico, mediante la flecha, un punto sobre la curva que correspondería a un umbral óptimo para el cual el predictor tiene un buen comportamiento, con una sensibilidad y especificidad alta próxima o superior al 80%. La línea roja que se aprecia sobre este mismo gráfico corresponde a la curva ROC de un clasificador aleatorio, incapaz de categorizar la clase para una nueva observación. Mientras más próxima a esta recta se encuentre la curva ROC de un predictor, peor es el rendimiento que éste posee como tal; por el contrario, mientras más próximo a la coordenada $(0,1)$ se encuentre el punto de corte óptimo de la curva ROC, el rendimiento será superior.



a. b. c.

Figura 2.8: Clasificador binario. a.) Matriz de confusión construida en base a los resultados entregados por un modelo predictor (o clasificador) para los valores actuales según clase (Positivo / Negativo); b.) distribución de la variable aleatoria para ambas muestras según clase, identificándose las proporciones equivalentes con los resultados en la matriz de confusión: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN); c.) Curva ROC (azul) obtenida para el modelo evaluado. [11]

El área bajo esta curva ROC, llamada comúnmente AUC, mide el rendimiento o la capacidad del predictor para distinguir entre ambas clases en un rango de valores entre 0.5 y 1, cuyos límites representan un predictor aleatorio y un predictor perfecto respectivamente. Un valor calculado menor a 0.5 en el AUC supone un predictor que clasifica al revés la clase para la variable analizada, es decir, predice la clase 'positivo' como 'negativo' y viceversa. El rendimiento del predictor se calcula en este caso como $1 - AUC$, o bien, se ajusta el mecanismo de clasificación.

[11] https://aprendeia.com/curvas-roc-y-area-bajo-la-curva-auc-machine-learning/

2.4 Aprendizaje Automático

El aprendizaje automático, o *Machine Learning*, es una rama de la inteligencia artificial que emplea herramientas de estadística, probabilidades y optimización mediante *software* para reconocer patrones en un conjunto de datos, relacionarlos entre sí y construir un modelo representativo de dicho conjunto con el objetivo de utilizar este aprendizaje para clasificar o predecir un resultado a nuevos datos adquiridos.

2.4.1 Aprendizaje automático supervisado

Este tipo de aprendizaje automático trabaja con una serie de datos de entrenamiento que consisten en pares de objetos o vectores cuya primera componente del par corresponde a los atributos o variables de entrada y la segunda, a las 'etiquetas', ya sea en forma categórica o en forma numérica obtenidos de alguna evaluación. Cuando la predicción tiene como resultado un valor numérico, el modelo corresponde al de una regresión, y cuando la predicción es de clases, el modelo consiste en uno de clasificación.

Los algoritmos computacionales utilizados como herramienta deben aprender a mapear estos datos ingresados y encontrar una función que, dados los valores de las variables de entrada, sea capaz de distinguir las muestras según clase; el aprendizaje se detiene cuando el algoritmo alcanza un nivel aceptable de rendimiento medido mediante alguna métrica de evaluación. De este modo, se espera que la función o modelo generado asigne la etiqueta adecuada de salida a algún nuevo dato de entrada, comportándose como un predictor de dicho resultado. En la Figura 2.9 se muestra un esquema del funcionamiento de un sistema de aprendizaje automático; el conjunto de datos de entrenamiento incluye entradas y salidas correctas, permitiendo que el modelo aprenda con el tiempo y genere, posteriormente, la salida deseada.[12]

Esta subcategoría de aprendizaje automático ha sido la más utilizada en la investigación y aplicación en el ámbito médico, específicamente en el estudio del cáncer con enfoque en el diagnóstico de la enfermedad, y en la predicción o pronóstico de algún resultado referente a esta misma, como la respuesta del tumor al tratamiento, la susceptibilidad de pacientes a desarrollar un tumor, la supervivencia y la recurrencia a esta enfermedad. [18]

[12] https://www.ibm.com/es-es/topics/machine-learning-algorithms

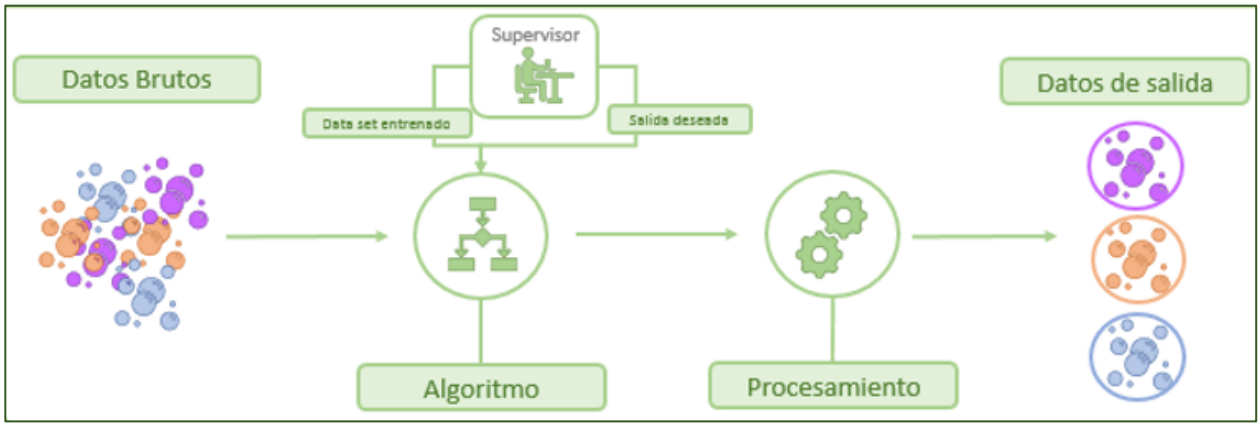


Figura 2.9: Esquema que representa un proceso de aprendizaje automático supervisado.

2.4.2 Tipos de algoritmos de clasificación

Como se mencionó previamente, en los procesos de aprendizaje supervisado se utiliza una variedad de algoritmos y técnicas de cálculo para predecir la probabilidad de una variable dependiente categórica, algunos de los cuales se describen brevemente a continuación. [19]

Regresión Logística (LR)

Éste es un modelo basado en probabilidad que estima la relación entre la variable dependiente binaria y las variables independientes, utilizando como método de cálculo la función logística o sigmoide. Esta función puede tomar cualquier número de valor real, asignando a un valor entre 0 y 1 relativo a la probabilidad de clasificar en cierta clase. Así, si la salida es mayor o igual que 0.5, el resultado se clasifica como 'positivo' o 1, y si es menor que 0.5, se clasifica como 'negativo' o 0.

Vecino más cercano - k (KNN)

Este algoritmo clasifica los puntos de datos en función de su proximidad y asociación con otros datos disponibles, asumiendo que los que se encuentran cerca entre sí corresponden a puntos de datos similares. Para ello, el modelo calcula la distancia entre puntos, generalmente mediante distancia euclidiana, y asigna una clase basada en la categoría más frecuente. Estos modelos son de clasificación rápida, útiles en problemas no lineales, y robustos en cuanto a nuevas variables de entrada; sin embargo, la complejidad computacional aumenta significativamente al aumentar el conjunto de datos de entrada.

Árbol de decisión (DT)

Este algoritmo trabaja en un esquema de nodos y ramas, donde cada nodo representa una única variable de entrada con un punto de división de dos ramas, mutuamente excluyentes, que representan la variable de salida que se utiliza para hacer la predicción de clases. El proceso de poda que se realiza permite resolver problemas de sobreajuste; sin embargo, el modelo final de árbol de decisión es sensible y depende del orden establecido en la selección de atributos o *features*.

Bosque aleatorio (RF)

La denominación "bosque" hace referencia a una colección de árboles de decisión no correlacionados entre sí, y que se fusionan en su mecanismo de cálculo para reducir la varianza y obtener predicciones de datos más precisas.

Navie Bayes (NB)

Este modelo predictor se basa en el principio de independencia condicional del Teorema de Bayes entre cada par de atributos dado el valor de clase, de modo que la presencia de un atributo no afecta a la presencia de otro en la probabilidad de un resultado dado. El cálculo en el modelo determina la posibilidad de si un punto de datos pertenece a una categoría definida. Una ventaja de este modelo es que un pequeño set de datos bastaría para su entrenamiento siempre cuando la cantidad de atributos o variables de entrada no fuese considerable, en una proporción aproximada de 5:1 según la reseña bibliográfica señalada de J. Cruz.

Máquinas de soporte vectorial (SVM)

En estos modelos se establece que la distancia entre dos clases de puntos de datos es la máxima y se construye un hiperplano, conocido como 'límite de decisión', en un espacio de dimensión N, equivalente a la cantidad de atributos de entrada, que separa ambas clases de puntos de datos a ambos lados del plano.

Estos algoritmos descritos involucran hiperparámetros, que corresponden a variables de configuración externa utilizados para administrar el entrenamiento de los modelos y que son definidos numéricamente previos al proceso de aprendizaje. Algunos de éstos, por ejemplo, son: la cantidad de

árboles de decisión en modelos RF, el número de vecinos k considerados en modelos KNN, y el parámetro C en el caso de modelos SVM cuyo rol es controlar la penalización por clasificación errónea durante el entrenamiento.

LASSO (Least Absolute Shrinkage and Selection Operator)

Este tipo de regresión combina un algoritmo de búsqueda y una función criterio con el objetivo de encontrar el mejor subconjunto de *features* para predecir la clasificación en cuestión. De este modo, su mecanismo permite reducir el número de variables de entrada en un modelo, identificando los atributos más importantes para resolver eficientemente el problema. El algoritmo integra un parámetro constante α que penaliza la norma de los pesos de los atributos considerados, controlando la fuerza de regularización. Como resultado, este modelo entrega un vector de coeficientes, correlacionados con los atributos que constituyen el vector de entrada y de los cuales la mayoría toman el valor cero.

2.4.3 Conjunto de datos: entrenamiento, validación y testeo

Cuando la base de datos disponibles es de gran volumen, el mejor enfoque para el desarrollo del modelo es dividir este conjunto en tres subconjuntos de forma aleatoria: un set de entrenamiento, un set de validación y un set de prueba. El primer set, tal como lo indica su nombre, corresponde a la muestra utilizada para entrenar y ajustar el modelo, aprendiendo los datos para cada una de las dos clases consideradas; el set de validación actúa como intermediario entre la formación del modelo y su evaluación final, estimando un error de predicción o eficacia para la selección del mejor modelo; y finalmente, el set de prueba se utiliza para una evaluación del ajuste final del modelo elegido, determinando el 'error de generalización' que corresponde a la medida de precisión con la que el modelo predice las clases para nuevos datos nunca antes vistos. [20]

Para situaciones donde la cantidad de datos no es numerosa se utilizan otras metodologías para la división del conjunto que permiten un entrenamiento y una evaluación satisfactoria; la validación cruzada (CV: *Cross Validation*) es una de ellas. Ésta elimina el requisito explícito de un conjunto de validación, utilizando el conjunto de datos completo mediante un remuestreo en su división para entrenar y ajustar el modelo, por una parte, y estimar su rendimiento, por otra parte. Convencionalmente se considera adecuada una división 80/20 del conjunto para entrenamiento/testeo, evitando sobreajustes y variaciones demasiado altas en la estimación de parámetros y en la estadística de rendimiento tal como se detalla en la literatura señalada.

En particular, para la validación cruzada *k-fold* se divide aleatoriamente el conjunto en k subconjuntos y se entrena el modelo con $k-1$ subconjuntos de datos, mientras que el subconjunto k-ésimo se utiliza para validar y testear el modelo mediante una métrica de evaluación. A continuación, se repite el mismo procedimiento destinando para cada iteración un subconjunto distinto como conjunto de prueba. Finalmente, el rendimiento del modelo para predecir el *outcome* deseado queda determinado por la media de las métricas de evaluación calculadas en las k iteraciones.

Adicionalmente, cuando los datos se encuentran desequilibrados entre ambas clases es preferible utilizar la técnica de validación cruzada *k-fold* estratificada, que mantiene la proporción de clase del conjunto total de datos (original) en cada subconjunto, permitiendo un entrenamiento del modelo tanto en la minoría como en la mayoría, o bien, de forma homogénea. En la imagen de la Figura 2.10 se aprecia el esquema de una validación cruzada *k-fold* estratificada, con $k = 5$, para un conjunto de datos que distribuyen en dos clases.

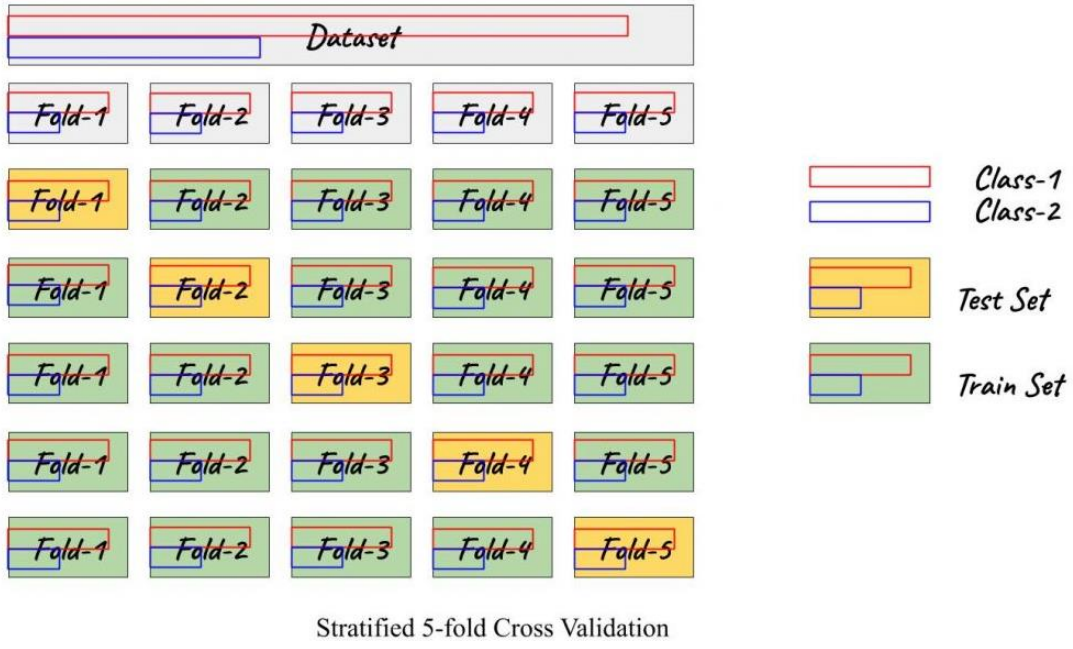


Figura 2.10: Esquema de remuestreo por validación cruzada *k-fold* estratificada, con , para un conjunto de datos clasificados en forma desbalanceada en clase 1 (*Class-1*) y clase 2 (*Class-2*). Cada uno de los subconjuntos de datos (*Fold*) mantienen la proporción entre clases del conjunto original (*Dataset*). Para cada iteración *k*, los subconjuntos que entrenan el modelo (*Train Set*) y el subconjunto utilizado para su testeo (*Test Set*) varían. [13]

[13] https://www.mantralabsglobal.com/blog/model-selection-cross-validation-quest-for-elite-model/

Una validación cruzada *k-fold* estratificada repetida realiza reiteradas veces el mismo procedimiento del esquema señalado anteriormente; la división estratificada de los datos en los distintos subconjuntos se realiza aleatoriamente para cada *n*-ésima repetición, de modo que sus configuraciones son diferentes cada vez.

2.4.4 Métricas de evaluación de modelos

Existen diferentes métricas que miden el rendimiento del modelo de predicción automatizado, estimando y evaluando su capacidad para predecir correctamente las clases ante nuevos datos de entrada[14]. Se señalan a continuación las utilizadas en el presente estudio.

AUC y curva ROC

El puntaje AUC y curva ROC fueron definidos y detallados en la Sección 2.3.5; sin embargo, el cálculo de la sensibilidad y especificidad implícito en ellos difiere para este caso en su método, el cual considera un vector con la clase verdadera de los datos y un vector con las estimaciones de probabilidad de la clase 1 (o con la etiqueta mayor) para estos mismos datos como resultado de ajustar y entrenar el modelo de predicción.

Accuracy

Esta métrica corresponde a una de las medidas más utilizadas y directa para evaluar un modelo automatizado, y representa el porcentaje de valores verdaderos o correctamente clasificados, tanto positivos (VP) como negativos (VN), sobre el total de las predicciones.

2.4.5 Comparación del rendimiento entre dos modelos

En una primera aproximación, el comparar los resultados de las métricas de evaluación entre los modelos permitiría destacar uno por sobre otros como mejor modelo predictor según interés. Sin embargo, existe la posibilidad de que los modelos desarrollados presenten similitud o discrepancia en sus resultados debido a simple casualidad estadística, de modo que se torna necesario aplicar pruebas de significancia estadística para analizar si los valores de la métrica obtenida provienen de la misma

[14] https://scikit-learn.org/stable/modules/model_evaluation.html

distribución o no, e inferir respecto a la superioridad de alguno de los modelos conforme al modelo predictor deseado. Benavoli *et al.* describen en su tutorial [21] distintos test como herramientas para la comparación, dos de los cuales se describen a continuación.

Frequentist correlated t-test

Con los múltiples resultados de la métrica de evaluación, AUC en este caso, debido a las N repeticiones de una validación cruzada al entrenar y evaluar un modelo, se construye un vector $\boldsymbol{x} = \{x_1, \dots, x_N\}$ de diferencia entre pares de valores AUC de los dos modelos a comparar.

Se define y determina entonces el estadístico como:

$$t = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 \left(\frac{1}{N} + \frac{n_{te}}{n_{te} + n_{tr}}\right)}} \tag{2.15}$$

donde $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ y $\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$ corresponden, respectivamente, a la media y a la desviación estándar del vector $\boldsymbol{x}$; n_{te} y n_{tr} representan el tamaño del conjunto de datos de teste y entrenamiento; y μ es el valor medio de la población a la que pertenece el vector $\boldsymbol{x}$.

Bajo la hipótesis nula H_0: $\mu = 0$, y la alternativa H_A: $\mu \neq 0$, el *p-value* del estadístico se calcula como:

$$p = 2 \cdot \left(1 - \mathcal{T}_{n-1}(|t(\boldsymbol{x}, 0)|)\right) \tag{2.16}$$

donde $\mathcal{T}_{n-1}(|t(\boldsymbol{x}, 0)|)$ corresponde a la distribución acumulada de la distribución de Student estandarizada con $n - 1$ grados de libertad en $t(\boldsymbol{x}, \mu)$ con $\mu = 0$.

Si p es menor a α, se rechaza H_0 y se concluye que ambos modelos difieren en su rendimiento, mientras que en caso contrario ambos modelos tendrían un comportamiento predictivo significativamente igual. Estas dos conclusiones son las únicas posibles que se obtienen de este análisis, sin cuantificar dicha diferencia o similitud.

Bayesian correlated t-test

Esta prueba, al igual que el caso anterior, considera la correlación debida a la superposición de conjuntos de datos para el entrenamiento y validación, y trabaja con un vector de diferencia entre pares de valores de la métrica de los modelos a comparar conocido como ‘distribución posterior’. La

diferencia en el análisis radica en que mediante la prueba se estima la probabilidad de que el rendimiento de dos modelos predictores sea igual o diferente, y cuán diferente.

Para los parámetros con los que se trabaja en este estudio, la probabilidad p de la prueba –t correlacionada bayesiana y el valor p de la prueba –t correlacionada frecuentista son numéricamente equivalentes; sin embargo, las inferencias extraídas son diferentes debido a la mayor información que entrega como resultado, cuantificando la diferencia de rendimiento entre los modelos si la hubiese, e identificando cuál de ellos es mejor o peor. Este análisis permite entonces tomar decisiones de selección del mejor modelo que cumpla el objetivo propuesto en esta tesis.

La distribución posterior (vector) se calcula como la diferencia de la métrica evaluada entre (MODELO 2 – MODELO 1). Se define una región de equivalencia práctica (*rope*), equivalente a un rango de valores para la distribución posterior, mediante el cual se estiman 3 resultados de probabilidad:

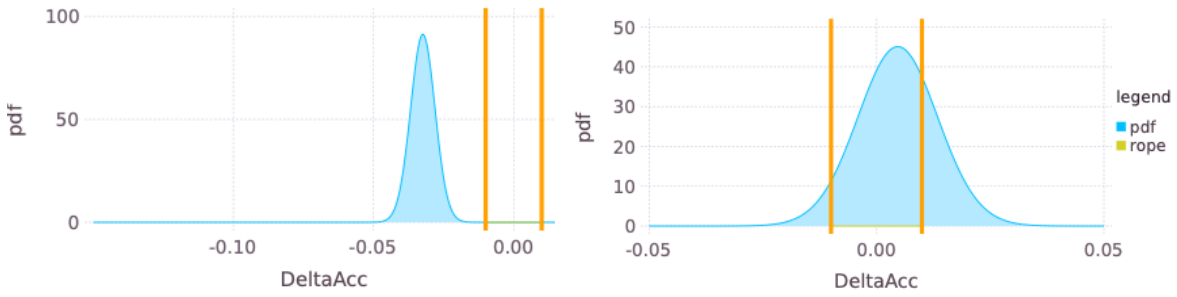
- La probabilidad $P\ (MODELO\ 2\ =\ MODELO\ 1)$ de que ambos modelos sean equivalentes estadísticamente en rendimiento, dada por la integral de la distribución en la región *rope*.

- La probabilidad $P\ (MODELO\ 2\ \ll\ MODELO\ 1)$, dada por la porción de la distribución a la izquierda de *rope*, de que el MODELO 1 sea superior en rendimiento.

- La probabilidad $P\ (MODELO\ 2\ \gg\ MODELO\ 1)$, dada por la porción de la distribución a la derecha de *rope*, de que el MODELO 2 sea superior en rendimiento.

En estudios realizados [22] se ha sugerido distintos rangos en los cuales podría establecerse *rope*, definido principalmente como un rango de un parámetro estandarizado. Para modelos lineales y pruebas t se ha generalizado como $[-0.01 \cdot \sigma_y\ ,0.01 \cdot \sigma_y]$, en base a la desviación estándar de la distribución posterior. Para modelos logísticos y de correlación, la definición varía; sin embargo, para otros modelos con resultado binario se recomienda especificar manualmente el argumento empleado para *rope*, a pesar de que los intervalos $[-0.1\ ,0.1]$ y $[-0.1 \cdot \sigma_y\ ,0.1 \cdot \sigma_y]$ se han establecido como los más utilizados.

En general, se define un umbral de significancia como regla de decisión, igual a 95% comúnmente, con el que se comparan las probabilidades calculadas para interpretar directamente si los dos modelos comparados son estrictamente equivalentes o cuál de ellos es mejor. Otro criterio es evaluar la media de la distribución posterior y analizarla en el rango de *rope*.



a. b.

Figura 2.11: Comparación de dos modelos de clasificación para el diagnóstico de 2 enfermedades (a. y b.), evaluados mediante métrica *accuracy*, utilizando análisis *Bayesian correlated t-test*. La distribución posterior (*pdf*) corresponde al vector de diferencia entre sus métricas de evaluación (*accuracy*). La región *rope* se encuentra delimitada entre -0.01 y 0.01.

En la Figura 2.11 se aprecian dos gráficos, obtenidos de la referencia bibliográfica señalada, como ejemplo del análisis descrito. Mientras que en a. se evidencia significativamente una diferencia en el rendimiento entre los dos modelos, destacándose al MODELO 1 como mejor, en b. el resultado es discutible entre autores, ya que no es estrictamente equivalente dado por la probabilidad calculada en *rope*, pero la media de la distribución posterior sí presenta su media en el rango y una probabilidad mayor al 60%

CAPÍTULO 3

Materiales y Métodos

En el presente capítulo se describe, en primera instancia, la información clínica disponible y las secuencias de imágenes MRI de la cohorte de pacientes con cáncer de mama sometidas a NACT a utilizar. En la segunda sección se explica la metodología y protocolos de segmentación empleados en el procesamiento de imágenes y extracción de las características radiómicas, indicados en la Sección 2.2.2. Posteriormente, se señalan las técnicas para el análisis estadístico y selección de datos, detallados en la Sección 2.3, a partir de los cuales se desarrollarán los modelos predictivos de pCR. Finalmente, se describe la metodología para la construcción y análisis de rendimiento de dichos modelos, basados en los algoritmos de *ML* mencionados en la Sección 2.4.2.

3.1 Cohorte de pacientes

Se escogió la colección *"ISPY-1"* disponible en el repositorio de acceso público de imágenes médicas para la investigación del cáncer '*The Cancer Imaging Archive'* (TCIA)[15]. Esta colección, conformada por estudios seriales multicéntricos con información de respuesta patológica completa (pCR) y supervivencia libre de recurrencia (RFS), reúne imágenes de resonancia magnética de pacientes mujeres con cáncer de mama en estadío 2 o 3 que recibieron tratamiento de quimioterapia neoadyuvante, ya sea con un régimen único de antraciclina-ciclofosfamida en un esquema de 4 ciclos o también seguido de 4 ciclos de taxano.

Cada una de estas pacientes registra 4 sets de imágenes, referentes a distintas etapas del tratamiento, con diversas técnicas de adquisición según los protocolos documentados [23, 24]. Para

[15] https://www.cancerimagingarchive.net/

el desarrollo de esta tesis se utilizaron las series adquiridas en la etapa previa al tratamiento neoadyuvante, específicamente las secuencias: ponderada en T1 (T1w), y dinámica de realce de contraste (DCE-MRI) pre- y la primera post- inyección intravenosa del agente de contraste, el que consistió en gadopentetato de dimeglumina en una dosis de 0.1 mmol/kg de peso corporal. Esta última secuencia se adquiere a los 2 minutos 30 s (2min) de ser administrado dicho contraste.

De las 222 pacientes que contempla la colección, se seleccionaron aquellas que registran la misma información técnica asociada a las imágenes de interés, expuestas en la Tabla 3.1, de manera que se asume el uso de un único scanner junto a un mismo protocolo para la adquisición de secuencias. Posteriormente, se descartan también las pacientes que no cuenten con la información clínica necesaria, como el resultado positivo o negativo a pCR y receptores hormonales. Las imágenes por evaluar se obtuvieron de un scanner de 1.5 Tesla de intensidad de campo, utilizando una bobina de radiofrecuencia de mama dedicada; esto implica una camilla especialmente diseñada para una posición prono del paciente con aberturas para las mamas rodeadas por bobina receptora de señal. La secuencia T1w corresponde a una exploración de localización realizada en plano axial de ambas mamas, mientras que las series DCE se realizaron unilateralmente sobre la mama sintomática y en orientación sagital. Estas últimas consistieron en una secuencia de eco de gradiente de alta resolución (resolución espacial en el plano ≤1 mm) tridimensional, con supresión de grasa, ponderada en T1, campo de visión de 16-18 cm, matriz mínima 256x192, y grosor de corte ≤ 2,5 mm.

Indicadores selección pacientes ISPY-1		
ID sitio / Protocolo clínico	AAI	
Fabricante / Modelo	GE MEDICAL SYSTEM / GENESIS SIGNA	
Intensidad campo magnético	1.5 T	
Datos de adquisición de imágenes MRI		
	T1w	**DCE**
TR	500 ms	8.4 / 8.9 ms
TE	14 ms	4.2 ms
TI	0 ms	15 ms
flip angle	90º	20º

Tabla 3.1: Datos y parámetros de las secuencias de imágenes bajo los cuales se seleccionaron las pacientes de la cohorte, para la etapa de desarrollo del modelo de predicción.

En la Tabla 3.2 se presenta la descripción de la cohorte finalmente seleccionada en cuanto a la información clínica registrada, en forma separada para las pacientes con resultado positivo (pCR) y negativo (no-pCR) a la evaluación de respuesta patológica completa de la lesión, y también en forma conjunta (Total).

Atributos	**Total** (n=59)	**pCR** (n=20)	**no-pCR** (n=39)
Edad ($\bar{x} \pm \sigma$) (años)	46.76 ± 9.57	44.42 ± 8.76	47.95 ± 9.85
LD Tumor baseline (mm)	62.03 ± 21.81	61.15 ± 17.26	62.49 ± 24.00
Raza			
Caucásico	47	15	32
Americano Africano	3	0	3
Asiático	7	3	4
Nativo hawaiano/ isleño Pacífico	1	0	1
Múltiple	1	1	0
Estado HR			
HR- (ER/PR)	25	13	12
HR+ (ER/PR)	34	7	27
Estado HER2			
HER2-	41	13	28
HER2+	18	7	11
Categoría (inmunohistoquímico)			
Luminal A	26	4	22
Luminal B	8	3	5
Triple Negativo	15	9	6
HER2+ (HR-)	10	4	6
Lateralidad			
Izquierda	26	8	18
Derecha	33	12	21

Tabla 3.2: Descripción clínica de la cohorte de pacientes, en forma conjunta y según resultado de respuesta patológica completa (pCR) o no (no-pCR) a la enfermedad.

La edad y la longitud mayor del tamaño del tumor pre-NACT (*LD Tumor baseline*) corresponden a variables continuas representadas en la tabla por la media y desviación estándar acorde a la muestra.

Se incluye, además, la raza según origen familiar, el estado de receptor hormonal (HR), el estado del Receptor-2 del factor de crecimiento epidérmico humano (HER2), la categoría según análisis inmunohistoquímico en que se clasifica el tumor de seno, y la lateralidad de la lesión. En el caso del estado HR, se representa como positivo (HR+) a las pacientes que poseen positividad para el receptor de estrógeno (ER) y/o para el de progesterona (PR), mientras que el negativo (HR-) representa la ausencia de ambos. La importancia de analizar el estado del HR, HER2 y categorizar el tumor para realizar un tratamiento efectivo fue señalada previamente en la Sección 2.1.

3.2 Procesamiento de imágenes

Como plataforma para la visualización, el procesamiento y el análisis de las imágenes se emplea el software gratuito y de código abierto *3DSlicer versión 4.10.2*, incorporando las extensiones de los módulos *"SegmentEditorExtraEffects"* y *"SlicerRadiomics"* para la segmentación de volúmenes de interés y la extracción de sus *features*, respectivamente. Este último módulo integra la librería *pyradiomics* para *Python*, que implementa el cálculo de una variedad de *features* mediante una interfaz gráfica que permite configurar tanto una 'personalización manual' para seleccionar los *features* a calcular.

Seleccionada ya la cohorte de pacientes, se efectúa como paso previo una exploración en ambas secuencias de imágenes (T1w y DCE- MRI), a modo de registrar las dimensiones y espaciado de las imágenes y de su volumen proyectado.

Para llevar a cabo el procesamiento de las imágenes se establece un protocolo, el cual permite disponer y trabajar las secuencias de cada paciente de igual manera entre sí, segmentando bajo los mismos criterios o normas las áreas y volúmenes de la mama de interés, su parénquima y el tumor, para posteriormente realizar el cálculo y extracción de sus *features* de manera idéntica. En su elaboración se consideraron indicaciones referenciadas por un radiólogo del área de imagenología mamaria, informes de la resonancia magnética disponible en el repositorio de imágenes TCIA [25] y el estudio realizado por *Aghaei F. et al*. [26]

En términos generales, se realiza primero una segmentación volumétrica de la mama de interés a partir de la secuencia T1w, debido al alto contraste entre el fondo de la imagen y el tejido mamario graso, y la que posteriormente es utilizada como zona restrictiva o editable. En esta misma secuencia se realiza la segmentación del parénquima mamario, que luego es aplicada sobre las imágenes mapeadas generadas por la sustracción entre la primera secuencia DCE-MRI posterior a la inyección de contraste (2min) y la secuencia base (antes de la inyección del agente de contraste). A su vez, el

tumor es segmentado directamente sobre estas mismas imágenes mapeadas, desde donde se calculan y extraen los *features* para su posterior análisis. El proceso completo se detalla a continuación.

3.2.1 Protocolos de segmentación

Segmentación de la mama de interés

i) Se cargan en el software *3D Slicer* los archivos de secuencia: T1w y DCE-MRI.

ii) Por medio módulo *"MultiVolumeExplorer"*, el multi-volumen correspondiente a la secuencia DCE-MRI se fracciona en los tres *frames* que lo componen, dando lugar a tres secuencias independientes de imágenes concernientes a las adquisiciones: pre-inyección de contraste, post-inyección de contraste a los 2 minutos, y luego a los 7 minutos.

iii) Utilizando el módulo *"Subtract Scalar Volumes"*, y con un orden de interpolación 0 como configuración, se realiza la sustracción entre la secuencia *post-contraste a los 2 min* y la secuencia *pre-contraste*, cuya secuencia resultante se denominará en adelante como *Dynamic-Sustr*.

iv) Se realiza el registro entre las dos secuencias estudiadas: T1w y *Dynamic-Sustr*, aplicando una transformación lineal sobre la primera por medio del módulo *"Transform"*.

v) Sobre la secuencia T1w, y en visualización axial, se segmenta ambas mamas por medio de la herramienta *"Level tracing"* al seleccionar algún pixel contenido en el contorno anterior (externo) de éstas, en todas las imágenes.

vi) En la imagen central de la secuencia se marca un 'punto referencial' en el medio y sobre el contorno anterior cóncavo entre ambas mamas, utilizando el efecto *"Paint"* con el menor diámetro disponible. Este punto queda descrito por las coordenadas , equivalentes al número del pixel en cada dirección de la imagen, cuyo origen $(0,0)$ corresponde a la esquina inferior-derecha de ésta, mientras que la esquina superior-izquierda pertenece a la coordenada $(255\ ,\ 255)$.

vii) Mediante la herramienta *"Scissors"* con forma rectangular, se corta la mitad de la segmentación volumétrica realizada en el punto v) concerniente a la mama sin el tumor.

viii) En la imagen central se procede a trazar una línea a 12^{o} bajo la horizontal, emulando la pared torácica, con el 'punto referencial' como origen. Para ello, se identifica primero la longitud del pixel en ambas direcciones, informado en el módulo *"Volumes"*. Por trigonometría, y para

el ángulo deseado, se calcula la coordenada $(255, y)$ hasta la que se trazará la línea fraccionaria en cuestión:

$$\tan 12° = \frac{y_{pr} - y}{(255 - x_{pr}) \cdot L_x} \quad (3.1)$$

Se demarca esta línea y toda el área bajo ella como una nueva segmentación por medio de la herramienta *"Scissors"*, con forma libre; y se propaga en eje z para que la segmentación quede aplicada a todas las imágenes en su conjunto. Si bien este paso puede ejecutarse para cada paciente, la segmentación que se realiza para la primera de ellas se exporta, en conjunto con el 'punto referencial' del paso vi), y se utiliza para todos los pacientes aplicándole una transformada lineal por medio de *"Transform"*, si fuese necesario, para ubicar dicho punto referencial donde corresponde sobre la mama en las respectivas imágenes de cada paciente.

ix) La segmentación obtenida en viii) se le resta a la segmentación resultante en vii) utilizando *"Logical operators"*.

x) Se afinan detalles de contorno por medio de "*Erase"* , removiendo 'manualmente' zona del pezón, zona del músculo pectoral y pared torácica aun presente.

xi) Por último, se le realiza a esta segmentación una reducción de margen de 4mm configurada en la sección *"Margin"* del editor de segmentación.

Segmentación del parénquima de la mama de interés

xii) Con la herramienta *"Threshold"* se ejecuta un umbral de Otsu [27], restringido a la segmentación de la mama lesionada lograda en xi) como área editable, con un rango de intensidad de pixel entre 0 y 30% del máximo valor registrado para la secuencia en cuestión (T1w), cuyo pixel pertenece al tejido mamario. Se genera así una binarización y, por consiguiente, la segmentación del volumen deseado en el rango de intensidad escogido. El valor mínimo y máximo de pixel según cada secuencia se muestran en el módulo *"Volumes"*.

xiii) En "*Segment Editor*" se crea un nuevo objeto de segmentación, aplicado ahora al volumen proyectado por la secuencia *Dynamic-Sustr*; y mediante el módulo *"Segmentation",* se le copia (*"Copy/move segments"*) la segmentación resultante en xii).

xiv) Finalmente, a la segmentación anterior se le resta la segmentación del tumor que se obtiene a partir del protocolo que se describe a continuación.

Segmentación del tumor

xiv) Sobre la secuencia *Dynamic-Sustr*, en visualización sagital, se aplica un umbral de Otsu con un rango de intensidad de pixel entre el 30 % del máximo valor y el máximo valor registrado.

xv) Con la herramienta *"Level tracing"* se segmenta el tumor cada 2 cortes de imágenes aproximadamente (puede ser menos o más según como varía la forma del tumor), restringiendo el contorno de la demarcación dentro de la segmentación realizada en el punto anterior.

xvi) Con la herramienta "*Paint*", y como una nueva segmentación, se demarca manualmente y de forma arbitraria en estos mismos u otros cortes la zona circundante al tumor. Este paso es requisito para proceder con el siguiente, ya que caracteriza el tejido externo al tumor para la posterior delimitación de éste.

xvii) Las segmentaciones anteriores se utilizan como semilla en el cálculo para el crecimiento y generación de la segmentación volumétrica final del tumor mediante la herramienta *"Grow from sedes"*.

En la Sección 4 (Resultados) se presenta, mediante imágenes, el resultado de los distintos pasos del proceso detallado previamente.

3.2.2 Extracción de *features*

Los *features* de las segmentaciones del tumor y del parénquima en la mama se calculan a partir de la secuencia mapeada de imágenes *Dynamic-Sustr*, utilizando el módulo *Radiomics -3DSlicer*.

Los *features* que se consideran para el estudio de esta tesis comprenden los basados en forma, en estadística de primer orden y estadística de segundo orden. El listado de los *features*, su significado y descripción matemática, se detalla en la documentación de *pyradiomics* previamente señalada.

En la pestaña *Select Input Volume* and *Segmentation* de la interfaz gráfica se selecciona el volumen *Dynamic-Sustr* sobre el cual se obtiene la información cuantificable *radiomics*; y luego se escoge la región o segmentación sobre la que se restringe el cálculo de *features* propiamente tal, de modo que se realiza este mismo procedimiento para el tumor y el parénquima mamario, independientemente.

Para configurar la extracción, en la pestaña *Extraction Customization* se escoge la opción de personalización manual (*Manual Customization*) seguido por la alternativa *All Features* en la sección *Feature Classes*, calculando entonces los *features:* de forma, de primer orden y de segundo orden (gldm, glcm, glrlm, glszm, ngtdm), que posteriormente se someterán a filtrado y selección.

En la pestaña *Settings* se mantiene el valor *Bin Width* en 25 por defecto.

A partir de los archivos generados para todos los pacientes, se conforman 2 matrices con los *features* extraídos, una referente al tumor y la otra al parénquima de la mama de interés, asociándoles un vector *target* que registra en forma binaria la clase pCR (1) y no-pCR (0) de cada paciente correlativamente. Se construye además una tercera matriz referente a la base de datos clínicos de los pacientes comprometidos.

Los archivos mencionados y los códigos elaborados posteriormente para el análisis de datos y desarrollo del modelo de predicción de pCR se encuentran disponibles para observación y utilización[16].

El procedimiento y cálculo realizado por el software fue ejecutado por un procesador 1.6 GHz Intel Core i5 de dos núcleos, con 8.0 GB de RAM.

3.3 Selección de *features*

Del proceso descrito en la sección anterior se obtiene una serie de datos, de los cuales algunos podrían proporcionar información redundante o no contribuir a la finalidad del estudio. Las dos matrices de *features*, junto con la matriz de datos clínicos, se analizan entonces estadísticamente para evaluar la significancia y poder predictivo de cada atributo entre los dos grupos de pacientes con distinta respuesta patológica (pCR/ no-pCR). Como resultado de este análisis estadístico se seleccionan los atributos que se utilizarán en la etapa de *Machine Learning* para el desarrollo de los modelos predictivos.

El procesamiento de datos y análisis estadístico se realiza en lenguaje de programación *Python*, implementando principalmente sus librerías *Pandas*, *SciPy* y *Matplotlib*.

3.3.1 Análisis estadístico de *features*: selección 'manual'

La cohorte de pacientes presenta dos resultados posibles al evaluar la respuesta del tumor a la quimioterapia: pCR (clase 1) y no-pCR (clase 0), conformándose así las dos muestras para el análisis comparativo. Ya que interesa determinar para qué *features* estos dos grupos de pacientes se presentan como independientes según su clase, se aplica un test de hipótesis de comparación de poblaciones. Éste será de tipo paramétrico o no-paramétrico según si estas muestras provienen de una población

[16] https://github.com/mbramirezb/Breast-cancer-pCR-prediction

con distribución normal o no, respectivamente, de modo que se evalúa primero si ambos grupos presentan para cada variable este tipo de distribución.

Análisis de normalidad

En este estudio se utiliza el *Test de Shapiro-Wilk* como prueba de normalidad, dado que el tamaño de muestra analizado es reducido, con un nivel de significancia estadística de $\alpha = 0.1$.

El *Test de Shapiro-Wilk* se aplica a ambas matrices de *features* (tumor y parénquima) y a las variables continuas de la matriz clínica, con la finalidad de contrastar la normalidad del conjunto de datos; esto se realiza de manera independiente para los dos grupos de pacientes según clase (pCR y no-pCR).

Como complemento al análisis de normalidad descrito, se realizan gráficos cuantil - cuantil (*QQ plot*) de cada *feature* para cada clase.

Comparación de poblaciones

A continuación, se aplica un test de hipótesis para comparar ambos grupos de pacientes según clase para cada *feature*.

En el caso de que las variables continuas estudiadas presenten distribución normal, se aplicará el 'test paramétrico *t de Student'*; en cambio, si gran parte de éstos ($\geq 50\%$) no presentan dicho tipo de comportamiento, se aplicará el 'test de *Wilcoxon-Mann-Whitney'* para todos ellos a modo de estandarizar el procedimiento. Por otra parte, en el caso de las variables dicotómicas presentes en la matriz de datos clínicos, se utiliza el 'test exacto de Fisher'.

Los *features* relativos al tumor y parénquima cuya hipótesis H_0 se rechazan, según Sección 2.3.2, son seleccionados para continuar con su análisis estadístico de forma particular, ya que son las posibles variables de entrada para la creación del modelo predictivo de interés en base a *ML*.

Análisis predictivo univariado

Por medio del puntaje AUC, detallado en la sección 2.3.5, se evalúa el poder predictivo de los *features* seleccionados como clasificador binario univariado respecto al resultado pCR (1) o no-pCR (0) que presentan los pacientes de la cohorte o muestra (M), cuyo resultado se denota por AUC_{umbral}^{M}. Para ello, se obtiene la curva ROC a partir de la sensibilidad y especificidad del clasificador al variar

el umbral de discriminación positivo/negativo según el valor de los datos y en función de su clase (0 o 1).

Este resultado es el que permitirá consolidar la selección de *features*. El cálculo de estos parámetros y métricas se implementa en *Python*, empleando como herramientas las funciones *roc_curve* y *roc_auc_score* de la librería *Scikit-learn*; y, al igual que todas las gráficas realizadas en este estudio, las curvas ROC se obtienen mediante *Matplotlib*.

Intervalos de confianza AUC

Un estimador poblacional relevante en el análisis de los datos, utilizado además universalmente para informar resultados de investigación, corresponde al intervalo de confianza (I.C.) cuyo rango de valores describe la variabilidad de la medida de la muestra.

Dado que los pacientes de ambas clases no distribuyen normal estrictamente para todos los *features* según sus valores, y que lo interesante es el poder predictivo que ellos poseen, entonces estos intervalos se construyen para la métrica AUC aplicando la técnica de remuestreo de datos *bootstrap*, Sección 2.3.3, dentro de la misma muestra debido al pequeño tamaño de ésta.

En este trabajo se fija n = 500 como la cantidad de submuestras a obtener, equivalente al tamaño de la 'muestra alternativa' mencionada en el marco teórico, con un tamaño de submuestra equivalente al 80% de la cohorte total de pacientes. Se calcula la métrica AUC de cada *feature* para cada una de estas submuestras de forma independiente, según el resultado pCR/no-pCR registrado originalmente para las pacientes consideradas en ellas respectivamente.

Se calcula posteriormente AUC^{B}_{umbral}, para cada *feature*, como el promedio entre las métricas AUC de todas las submuestras generadas por *bootstrap*, con su respectiva desviación estándar σ. Se fija un nivel de confianza del 90%, para el cual se obtiene el estadístico $z_{\alpha/2} = 1.645$ mediante las tablas estadísticas de 'Distribución Normal Estándar'.

Correlación de variables

Se aplica un test de correlación de variables entre los *features* seleccionados con el fin de descartar los que ofrecen información redundante, evidenciado estadísticamente por medio del coeficiente de correlación r.

Si la totalidad de los *features* evaluados presentan una distribución normal para ambas clases (0 y 1), se aplica el test paramétrico de correlación de *Pearson* en la evaluación; en caso contrario, se

utiliza el test no-paramétrico de *Spearman*. Las variables discretas clínicas son igualmente evaluadas con esta última prueba.

En el proceso se calcula y se grafica, en función de r, la matriz de correlación de las *features* de mama y parénquima mamario por separado, fijando un umbral $r = \pm 0.8$ para aceptar o descartar dependencia entre las variables evaluadas. De existir correlación en algún par, el *feature* con menor valor AUC_{umbral}^{M} es descartado.

Como resultado del análisis estadístico univariado desarrollado en esta sección, se seleccionan "manualmente" los *features* que configurarán la matriz de entrada a los algoritmos de predicción de pCR basados en ML.

No obstante, y de forma paralela, se realiza también una selección automatizada de atributos con el objetivo de comparar y validar el resultado previamente obtenido. Para esta selección se disponen, al igual que para el análisis estadístico previo, las dos matrices originales completas de *features* (tumor y parénquima mamario) y la matriz de datos clínicos.

3.3.2 Selección automática de *features* por ML

Para la selección se utiliza regresión LASSO, desde la librería de software de aprendizaje automático *Scikit-learn* para *Python*. Este modelo se prefiere por sobre otros cuando se espera que la cantidad de variables significativas para la predicción sea pequeña, proporcionando una forma basada en principios para reducir el número de estas variables e identificar las más importantes.

Bajo estas premisas, se entrena el modelo LASSO con los datos de la cohorte completa previamente estandarizados, utilizando la técnica de validación cruzada *k-fold* estratificada con repetición, para elegir intrínsecamente un factor de penalización para la resolución. En la configuración de LASSO se utiliza $k = 3$, con una repetición de 10 veces, y se elige el valor AUC como *scoring* para determinar la importancia de los *features* como predictores del pCR.

3.4 Modelos de predicción de pCR basados en ML

En esta siguiente etapa del estudio, se desarrollan y evalúan distintos modelos uni- y multi- variados basados en ML como predictores de la respuesta patológica del tumor tratado con NACT conforme a

las dos clases registradas: pCR (1) y no-pCR (0). Posteriormente, se comparan sus rendimientos de clasificación según la métrica AUC con el fin de escoger el mejor modelo para el caso evaluado.

Los modelos de clasificación se implementan usando la librería de software de aprendizaje automático *Scikit-learn* para *Python*. Estos algoritmos por utilizar son: regresión logística (LR), k vecinos más cercanos (kNN), árbol de decisión (DT), *random forest* (RF), máquinas de soporte vectorial (SVM) y clasificador bayesiano ingenuo (NB), los cuales fueron definidos en la sección 2.4.2 y cuyos hiperparámetros presentan una configuración predeterminada.

3.4.1 Pre-procesamiento de datos

Se construye la matriz de datos de la cohorte de pacientes con los *features* y variables clínicas seleccionadas a partir del análisis descrito en la sección 3.3.1, la que será utilizada como entrada para entrenar y validar al mismo tiempo los modelos a medida que se construyen.

Para evitar que alguna variable domine por sobre otras la función objetivo de los algoritmos, se escalan los datos que compondrán el set de entrenamiento por método de estandarización (*StandarScaler)* para cada *feature*, expresado mediante:

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

donde x son los datos, μ es la media y σ, la desviación estándar. Estos dos últimos se calculan para el set de entrenamiento que se señalará enseguida, transformando luego estos mismos datos según el ajuste obtenido. Debido al reducido tamaño de la cohorte de pacientes, se utiliza el método de resampleo de validación cruzada *k-fold*, por lo que la estandarización y transformación de datos se realiza en el conjunto formado por los *k-1 fold* destinados a entrenamiento, y luego se aplica este ajuste de parámetros a los datos de testeo equivalentes al *fold k*. Esto se debe realizar en cada iteración del proceso, según como se detalla en la sección siguiente.

De este modo, se espera que el estimador aprenda correctamente y de forma similar de todas las variables según lo esperado.

3.4.2 Construcción de los modelos de predicción de pCR

En el proceso de entrenamiento y validación de cada uno de los modelos señalados se emplea la validación cruzada de *k*-iteraciones para un $k = 3$, equivalente a la cantidad de subconjuntos en que

se divide la muestra, calculando la métrica AUC como principal medición de la predicción de pCR para cada set de testeo *fold-k*. La división de los datos para conformar cada uno de los *k-fold* de realiza de manera estratificada (*Stratified k-fold Cross Validation*) entre las dos clases (PCR/no-PCR).

Este proceso de CV se repite $N = 500$, por medio de la función *RepeatedStratifiedKFold(),* generando cada vez diferentes conjuntos de forma aleatoria al cambiar la semilla *random* en la división interna. Esto permite aumentar, de cierta manera, el tamaño del conjunto de datos, logrando como resultado tener un conjunto de 1500 valores de AUC. Con ello, se determina el intervalo de confianza del 90%, junto con la media y desviación estándar, para la métrica de evaluación AUC de cada modelo predictor.

Paralelamente y de igual manera, se calcula la métrica *Accuracy* de los modelos predictores de pCR generados, como complemento para la distinción entre modelos cuyo rendimiento según AUC sean significativamente equivalentes.

Todo el proceso descrito en la sección se realiza sistemáticamente tanto de manera independiente para cada variable de la matriz de entrada como también para todas las combinaciones posibles entre ellas, configurando así modelos univariados y multivariados cuyos resultados permitirán realizar una comparación en cuanto al rendimiento de ellos y una posterior selección respecto al modelo predictivo buscado según el objetivo de la tesis.

3.4.3 Comparación de los modelos

A diferencia del test de comparación descrito en la sección 2.3.2, el conjunto de muestras en este caso es pareado debido al uso de una misma matriz de entrada para cada modelo de clasificación. Ya que, además, las observaciones no son independientes debido a la utilización de la técnica de validación cruzada, se debe introducir una correlación entre las observaciones debido a esta superposición de los conjuntos de datos de entrenamiento.

Entre los modelos se destacan y comparan particularmente los univariados entre sí, y algunos multivariados: el modelo conformado por todos los *features* seleccionados referente al tumor (T), el conformado por los *features* referentes al parénquima mamario (P), y los modelos conformados en forma combinada entre sí de estos *features* (T + P) y en conjunto con los datos clínicos (T + P + C).

Por otra parte, se comparan también los modelos con mejor rendimiento al considerar las combinaciones de *features*, referentes a tumor y parénquima mamario, de: forma y primer orden (Sh + f.O.), y a la combinación de éstos con los datos clínicos (Sh + f.O. + clínicos). La combinación de éstos con los *features* texturales se encuentran dentro del análisis mencionado en el párrafo anterior.

Cabe señalar, que en los casos que se presente equivalencia estadística entre modelos con la misma configuración de *features* y distinto algoritmo clasificador, al comparar la métrica AUC mediante análisis bayesiano, se selecciona para el registro en las tablas de resultados (Capítulo 4) los que presenten un mayor valor para la métrica *Accuracy*.

Se utiliza, en primer lugar, *Frequentist correlated t-test* para comparar el rendimiento de los modelos mencionados respecto a un modelo de clasificación aleatoria en base a la métrica AUC, con $\alpha = 0.05$. El objetivo es discriminar si resultan equivalentes o no.

Se realiza luego un análisis bayesiano mediante *Bayesian correlated t-test* para comparar entre sí los modelos mencionados y estimar la probabilidad de que el rendimiento de éstos sea igual o diferente, e identificar cuál predice mejor.

Se establece $rope$ escogiendo la desviación estándar de la métrica AUC de los modelos predictivos como el parámetro estándar; de este modo, el intervalo de *rope* se fija como $[-0.1 \cdot \sigma_{AUC}\ , 0.1 \cdot \sigma_{AUC}]$.

Finalmente, se contrastan también los resultados de AUC entre los algoritmos predictivos de pCR conformados por todos los *features* seleccionados manualmente y los seleccionados automáticamente mediante LASSO. Escogido el mejor modelo para cada una de estas dos configuraciones (selección) de *features*, se evalúa la importancia o los pesos que estos atributos tienen en su entrenamiento y ajuste.

CAPÍTULO 4

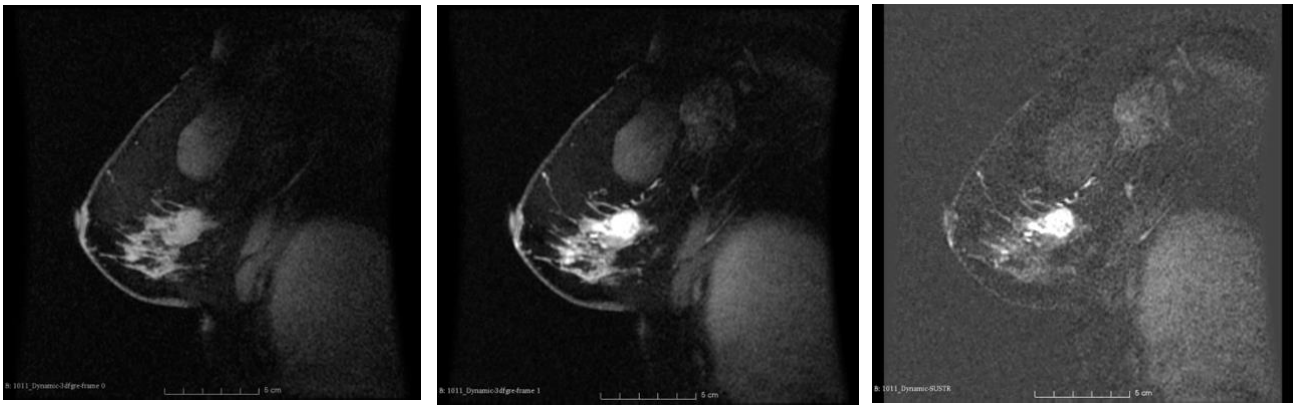
Resultados

Se muestran a continuación los resultados del estudio conforme a la metodología señalada en el Capítulo 3. En la primera sección se detalla, por medio de imágenes, el proceso de segmentación de las regiones de interés sobre las series correspondientes de MRI; mientras que en las siguientes secciones se muestran los resultados, propiamente tal, del análisis estadístico de los datos, de la selección de *features* para la construcción de los modelos predictivos de pCR basados en ML, y del desempeño de estos mismos junto con su posterior confrontación.

4.1 Procesamiento de imágenes

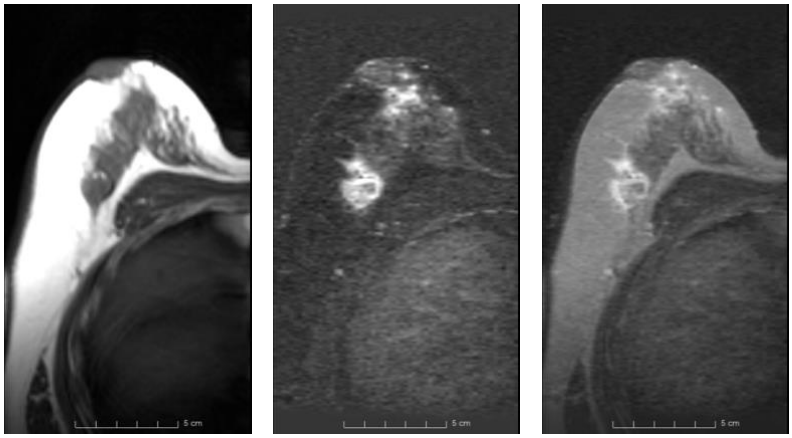
Tal como se detalla en el protocolo de segmentación de la Sección 3.2.1, de la operación aritmética de sustracción realizada entre los volúmenes post-contraste 2 min y pre-contraste que conforman la serie DCE-MRI se obtiene la serie *Dynamic-Sustr*. En la Figura 4.1 a. y b. se observan las imágenes de un mismo corte en plano sagital de la mama de una paciente adquiridas en los tiempos mencionados, mientras que la tercera imagen expuesta corresponde al resultado de la sustracción entre las dos anteriores, donde se manifiesta con mayor acentuación la forma del tumor.

Para el mismo caso de paciente, en la Figura 4.2 se observa la imagen de un corte en plano axial de la secuencia ponderada en T1, junto a su análoga perteneciente a la serie *Dynamic-Sustr*, mientras que la tercera imagen corresponde a la superposición registrada de ambas. La visualización en el software de la superposición entre las dos secuencias señaladas posibilitó verificar y/o realizar un registro adecuado entre ellas, siguiendo como principal referente el contorno interior de la piel de la mama y el músculo pectoral.



a. b. c.

Figura 4.1: Imagen de corte sagital de una mama lesionada de un paciente con pCR en los tiempos de: a.) 'pre-contraste' y b.) 'post-contraste 2 min', de la secuencia DCE-MRI. c.) Imagen del corte sagital equivalente (misma coordenada) en la serie *Dynamic-Sustr*.



a. b. c.

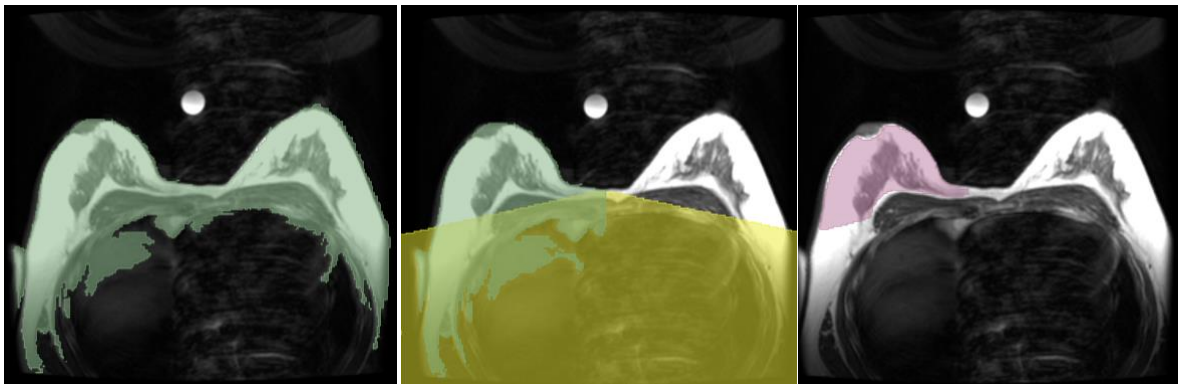
Figura 4.2: Sección de un corte axial de la secuencia ponderada en T1; b.) imagen del corte equivalente al de a., perteneciente a la serie generada *Dynamic-Sustr*; c.) superposición de ambas imágenes, con visibilidad del 50% para cada una.

4.1.1 Segmentación de volúmenes de interés y extracción de *features*

En primer lugar, se realizó la segmentación de la mama lesionada en las imágenes de la secuencia ponderada en T1 con el fin de delimitar la región de trabajo donde se perfilaron posteriormente las segmentaciones correspondientes al parénquima y tumor.

En la Figura 4.3 se aprecia el proceso de esta primera segmentación según protocolo, por medio de imágenes para la misma paciente retratada previamente. En a. se observa la segmentación aproximada de ambas mamas sobre un corte axial de la secuencia en cuestión, y cuya demarcación

se realizó en todas las imágenes por medio de un rastreo de contorno de nivel, priorizando su exactitud respecto al contorno externo de ambas mamas. Esta segmentación se corta a lo largo de su eje central y se descarta la sección referente a la mama contralateral, como se muestra en la imagen b. Sobre esta misma se observa en amarillo el área bajo las líneas referenciales utilizada para suprimir la zona sin interés de la segmentación de la mama lesionada. En c. se observa, finalmente, el resultado luego de otros ajustes como: remoción del pezón, afinación manual principalmente en la zona de musculatura pectoral, y la reducción del margen del segmento de 4mm.



a. b. c.

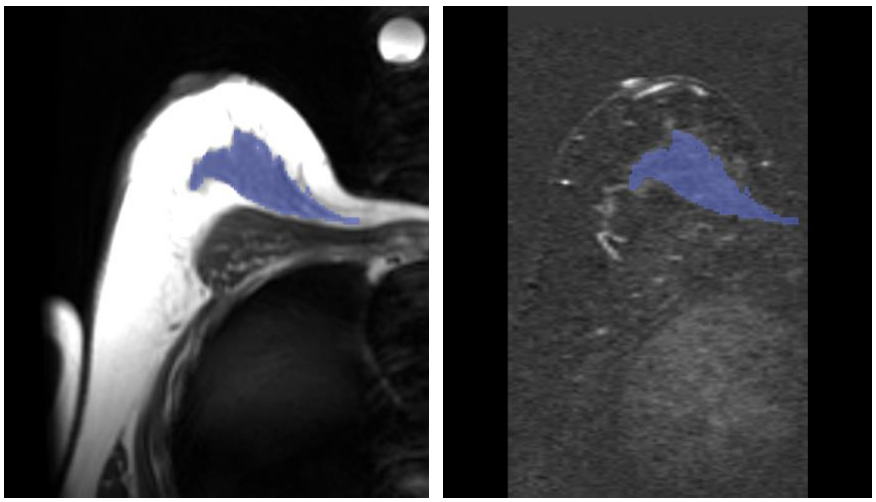
Figura 4.3: Proceso de segmentación de la mama lesionada en la secuencia ponderada T1: a.) Demarcación de ambas mamas por rastreo de contorno de nivel; b.) área bajo la línea referencial en 12º respecto a la horizontal, emulando la pared torácica; c) segmentación final de la mama de interés.

A continuación, se realizó la segmentación del parénquima de la mama de interés. En las imágenes de la Figura 4.4 se observa el resultado del proceso en el mismo corte axial presentado anteriormente, tanto en la secuencia ponderada en T1 como en la serie *Dynamic-Sustr*. Esta segmentación se obtuvo al aplicar el umbral de corte igual al 30% del valor máximo de pixel de las imágenes en su conjunto (volumen), supeditado a la segmentación volumétrica de la mama lograda previamente; y posteriormente se asignó a la serie *Dynamic-Sustr* para extraer sus *features*.

Por último, se segmenta el tumor según el protocolo establecido para ello. Las imágenes en la Figura 4.5, que corresponden a un corte de plano sagital de la serie *Dynamic-Sustr* para la misma paciente retratada, muestran parte de este proceso. Dado que el tumor se destaca como zonas blancas brillantes principalmente, exceptuando los tejidos necróticos que éste pudiese contener, en a. se observa el resultado de aplicar el umbral de corte igual al 30% del valor máximo de pixel de todas las imágenes de la serie en cuestión como primer paso. En b. se aprecia en color lila la segmentación del

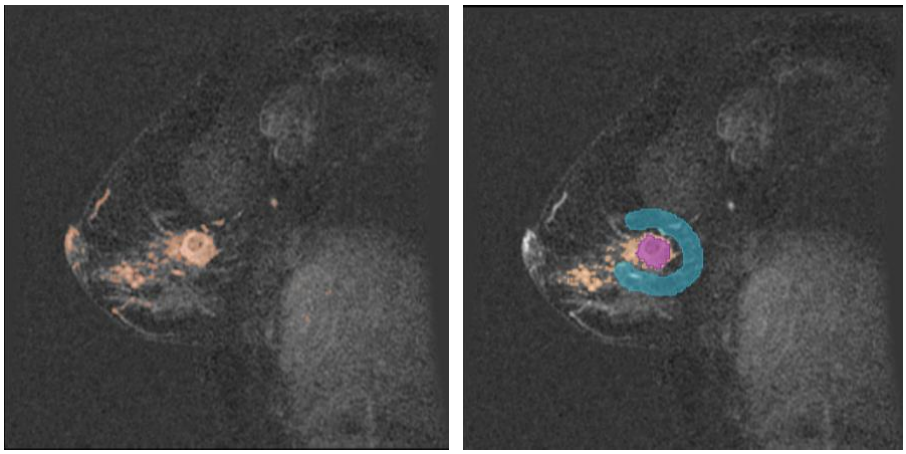
tumor realizada sobre dicha imagen mediante un contorno de nivel restringido a la zona demarcada previamente en a. y en color celeste la zona circundante, utilizándose como semilla de crecimiento para el cálculo y demarcación del volumen del tumor.

En la Figura 4.6 se muestran las imágenes con la segmentación final del tumor de la paciente aludida en cada plano de visualización y su proyección volumétrica, extraídas de la interfaz del *software*.



a. b.

Figura 4.4: a.) Segmentación de la parénquima sobre la secuencia ponderada en T1 mediante umbral de corte; b.) misma segmentación aplicada ahora directamente sobre la serie *Dynamic-Sustr*.



a. b.

Figura 4.5: Corte sagital de la serie *Dynamic-Sustr* de la mama lesionada; a.) Segmentación volumétrica en la mama a partir de un valor umbral definido, b.) Segmentación superficial del tumor mediante un contorno de nivel del valor de pixel.

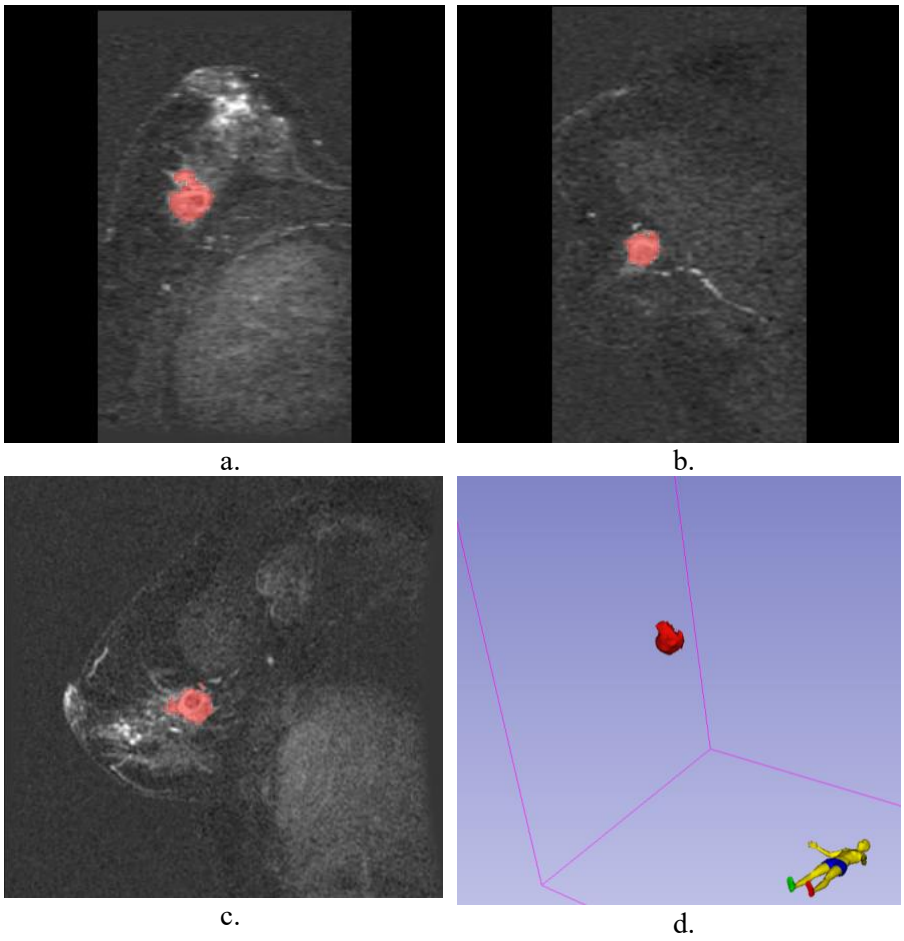


Figura 4.6: Segmentación final del tumor para la paciente retratada con pCR; a.) corte axial, b.) corte coronal, c.) corte sagital, y d.) proyección volumétrica.

4.2 Selección de *features*

4.2.1 Selección manual

La selección de *features* significativos para la generación del modelo de predicción de interés se realizó a partir del análisis estadístico de los datos, seguido de un análisis del poder predictivo de éstos por medio de AUC, y cuyos resultados se exponen a continuación.

Test de Normalidad

Como resultado de aplicar el test estadístico de *Shapiro Wilks* con un nivel de confianza del 90% a la matriz completa de *features*, tanto del tumor como del parénquima, y de forma independiente para ambas clases (pCR y no-pCR), se obtiene que menos del 50% de un total de 107 *features* extraídos presentan un comportamiento con distribución normal (*p-value* > 0.1), tal como se señala en el gráfico de la Figura 4.7.

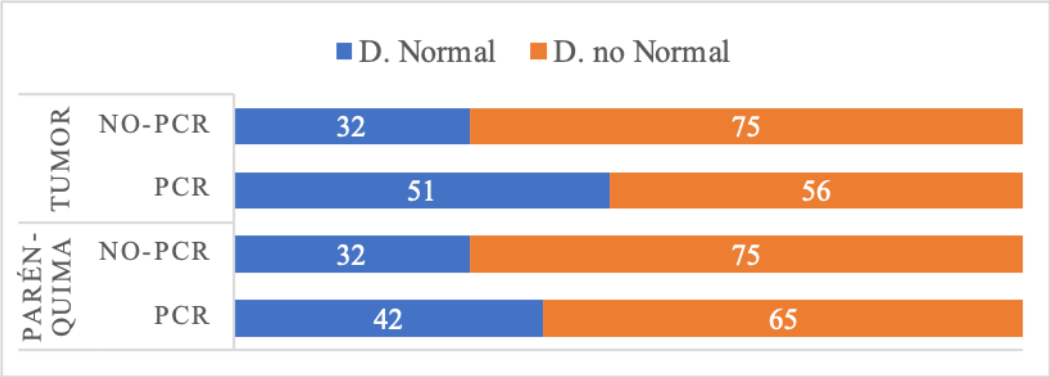


Figura 4.7: Cantidad de *features* que distribuyen normal y no-normal conforme al test *Shapiro Wilks* para ambas clases de pacientes (pCR y no-pCR) según la región segmentada.

En el caso de la matriz de datos clínicos se realiza el análisis del test a las únicas 2 variables continuas registradas y definidas en la Sección 3.1: *Edad* y *LD Tumor baseline*, para los cuales se acepta como resultado la hipótesis nula de provenir de una distribución normal.

A modo de complementar la aplicación de la prueba y corroborar sus resultados, se construyeron los gráficos cuantil-cuantil *QQ* para cada *feature*. En la Figura 4.8 se observa, como ejemplo, los gráficos QQ para la variable *Kurtosis,* tanto del tumor como del parénquima, según clase pCR y no-pCR. Éstos se corresponden con los resultados obtenidos en el test de normalidad realizado.

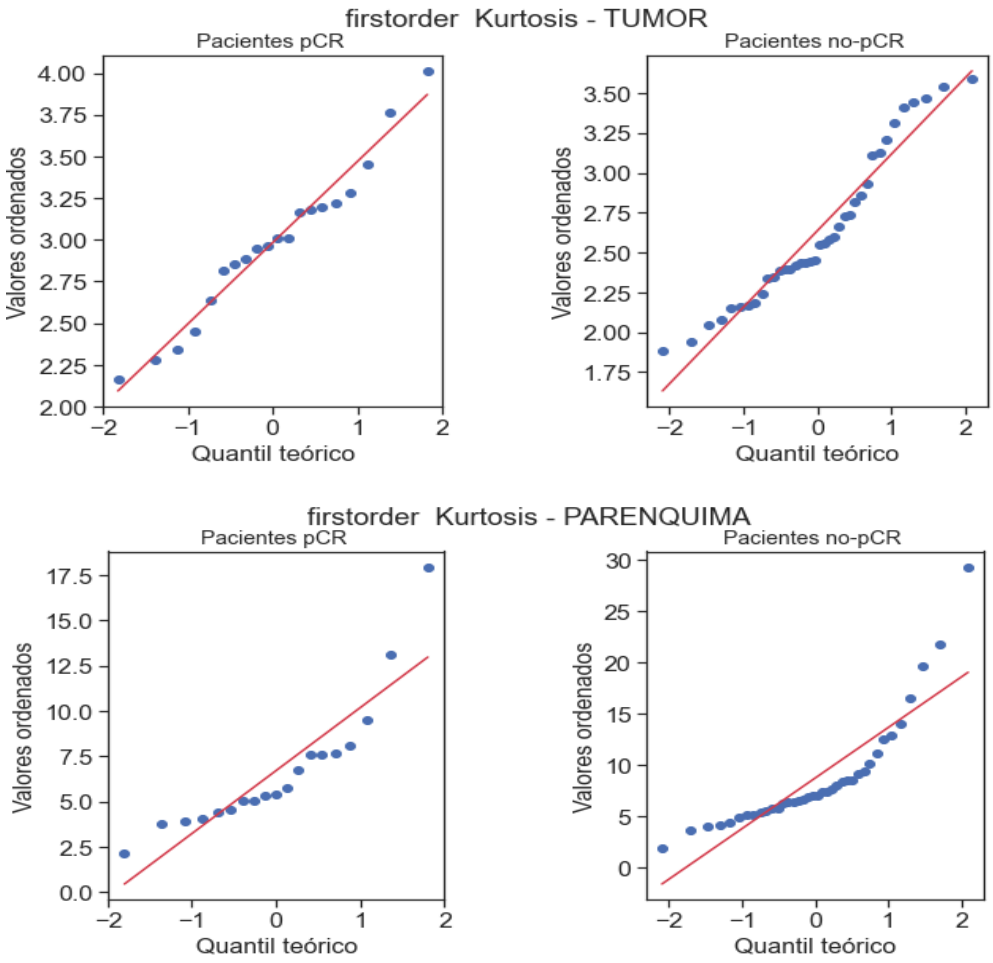


Figura 4.8: Gráficos QQ para el *feature Kurtosis* según volumen segmentado y clase. El gráfico superior derecho es el único que presenta una distribución normal en sus datos, con un $p - value = 0.217$ (> 0.1).

Cabe recordar que la importancia de evaluar la normalidad en la distribución de datos es por el tipo de prueba de *comparación de población*, paramétrica o no paramétrica, que se aplica en la siguiente etapa.

Test de hipótesis de comparación de poblaciones

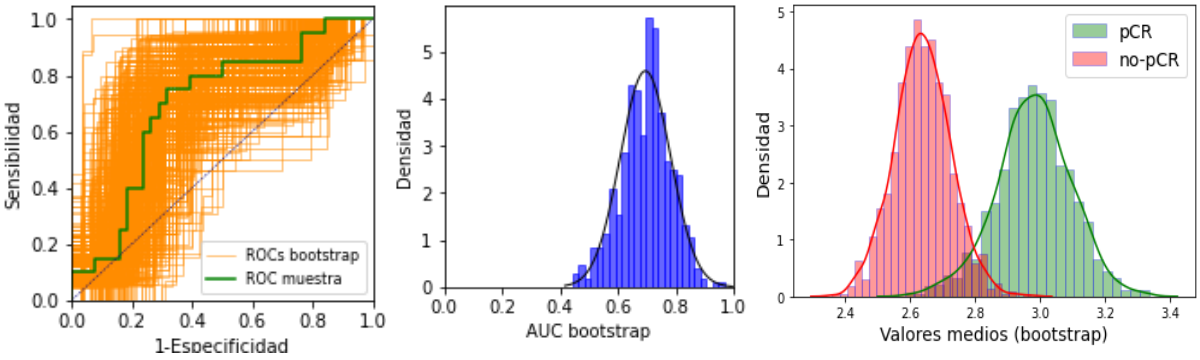
Debido a la distribución no normal de la mayor parte de los *features*, se hizo imperativo aplicar el test no paramétrico de *Wilcoxon-Mann-Whitney* para la evaluación de independencia entre las dos muestras según clase, con un nivel de significancia de $\alpha = 0.1$ en el contraste de hipótesis. Este mismo fue aplicado a las variables continuas de la matriz de información clínica; y para el caso de

las variables dicotómicas se utilizó el test exacto de Fisher, igualmente evaluado con un nivel de significancia del 10%.

En la Tabla A.1 del Apéndice-A se detallan los *features* del tumor, del parénquima y los atributos clínicos que presentan una diferencia estadística significativa tal ($p - value < 0.1$), que entre los pacientes con pCR y no-pCR es posible asumir independencia entre sus poblaciones. En el caso del tumor son 3 los *features* que cumplen dicha condición, mientras que para la parénquima y datos clínicos son 38 y 2, respectivamente; siendo éstos los seleccionados como primera aproximación.

Análisis predictivo AUC e Intervalos de confianza

Se determinan posteriormente para cada *feature* seleccionado el valor AUC^{M}_{umbral}, a partir de la muestra original de la cohorte de pacientes, y AUC^{B}_{umbral}, obtenido por remuestreo *bootstrap*, por medio de su curva ROC.



a. b. c.

Figura 4.9: a.) Curvas ROC de predicción estadística de pCR según *feature Kurtosis* del tumor a partir de la muestra original y de las submuestras generadas por *bootstrap*. b.) Distribución de AUC, relativa a las curvas ROC en a. c) Distribución de las medias del *feature* en las submuestras por *bootstrap* para ambas clases de forma independiente.

A modo de ejemplo, se retrata el análisis para el *feature* de *Kurtosis* del tumor en los gráficos de la Figura 4.9. En a. se observa la curva ROC (verde) trazada a partir de la muestra original y en naranjo las distintas curvas ROC correspondientes a las submuestras generadas por la técnica de *bootstrap* de la sección, utilizada para construir los intervalos de confianza de la métrica AUC para los *features* de interés. En b., se observa la distribución que presentan los valores AUC obtenidos a partir de dichas curvas ROC, calculando la media (AUC^{B}_{umbral}) y desviación estándar que permiten

definir el I.C. al 90% de confianza para esta métrica. En c. de la figura se muestran las distribuciones de los valores medios del *feature*, mediante el método de *bootstrap*, que presentan las pacientes según clase (pCR/no-pCR), manifestando la independencia entre estas dos poblaciones como resultado de la prueba de *Wilcoxon-Mann-Whitney*.

En la misma Tabla A.1 del Apéndice-A se presentan para los *features* seleccionados los resultados del análisis predictivo univariado: AUC^{M}_{umbral}, AUC^{B}_{umbral}, con su media y desviación estándar, y $I.C.\ AUC_{umbral}$, que presenta el intervalo de confianza de la métrica al 90%, según Ec. 2.12.

En la Figura B.1 del Apéndice-B se exponen los gráficos equivalentes a los de la Figura 4.9 para cada uno de los *features* seleccionados al finalizar el análisis estadístico completo.

Correlación de las variables seleccionadas

Se evaluó la existencia de correlación entre los atributos pre-seleccionados mediante el test de correlación de *Spearman* de modo independiente para los del tumor y del parénquima. Los coeficientes de correlación obtenidos para el tumor fueron todos menores que el umbral 0.8 establecido en la Sección 3.3.1, por lo tanto, ninguno es excluido para la siguiente etapa. Por otra parte, de los 40 *features* que se tenían para el parénquima, cuya matriz de correlación se adjunta gráficamente en la Figura B.2 del Apéndice-B, solo 6 fueron escogidos. La discriminación entre los *features* correlacionados con umbral sobre 0.8 se realizó considerando los resultados de la métrica AUC^{M}_{umbral}, prevaleciendo el con mayor valor.

En la Figura 4.10 se observa gráficamente los resultados del valor medio AUC en conjunto con su intervalo de confianza del 90% para los *features* del tumor, del parénquima, y variable continua clínica que se seleccionaron para la siguiente etapa. En la Figura 4.11 se grafica la matriz de correlación con estos mismos *features* analizados y finalmente seleccionados para: el tumor y el parénquima mamario.

En definitiva, y adicionando el atributo HR, fueron 11 atributos los seleccionados manualmente mediante estadística, a partir de los cuales se configuró la matriz de entrada para los algoritmos de clasificación basados en ML descritos en la Sección 2.4.2.

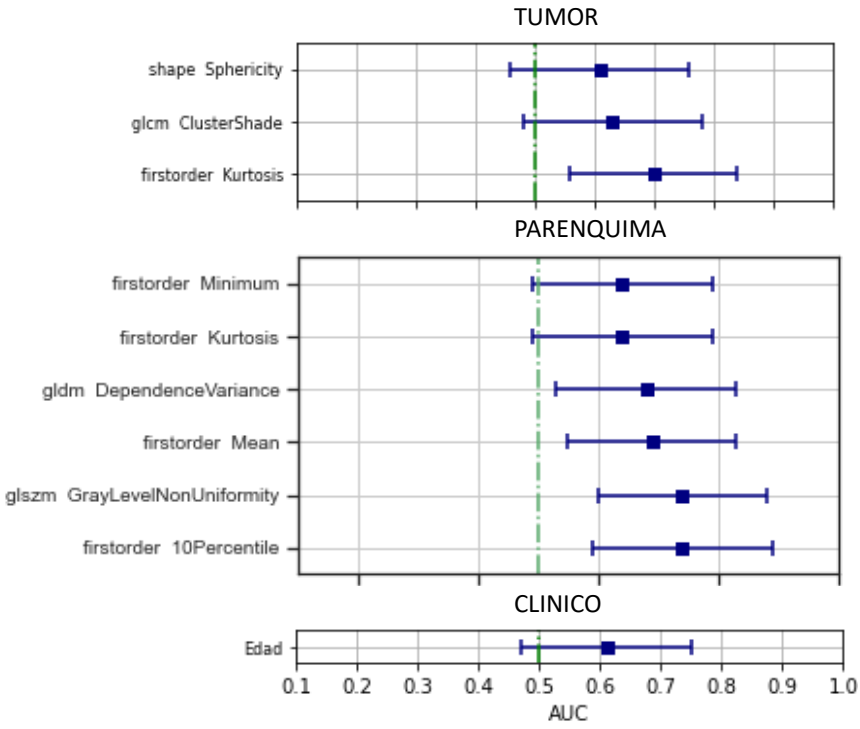


Figura 4.10: Valor medio e I.C. de la métrica AUC, por técnica de *bootstrap*, para los *features* preseleccionados. La línea segmentada en verde, AUC = 0.5, representa un clasificador aleatorio.

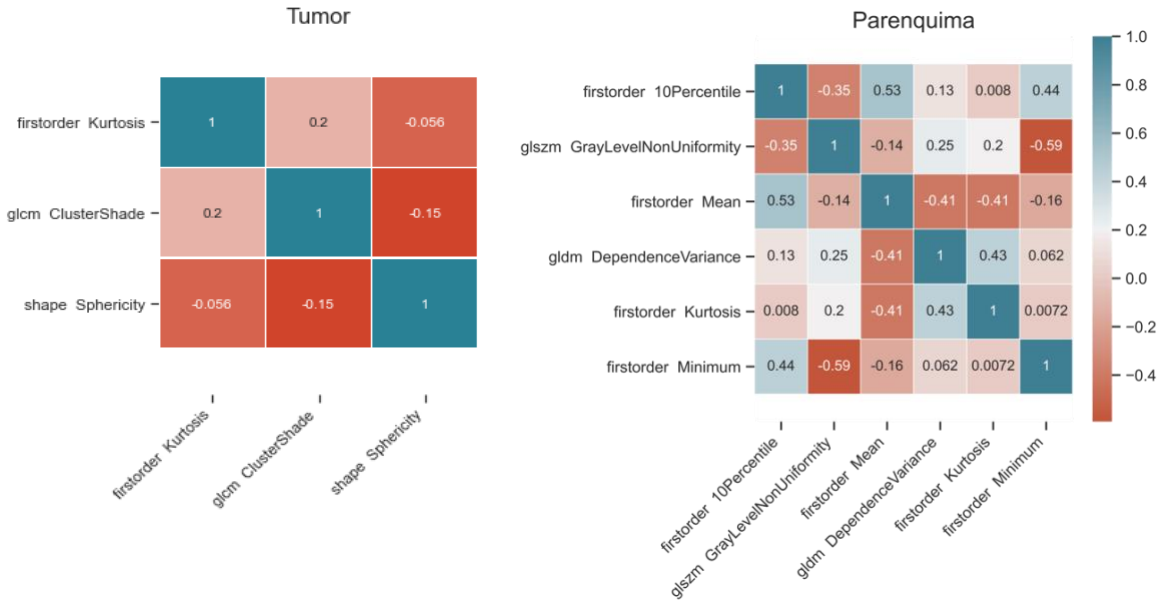


Figura 4.11: Matriz de correlación de los *features* de tumor y parénquima seleccionados finalmente como atributos para la generación de modelos predictivos en base a ML.

4.2.2 Selección automática

Utilizando las herramientas *GridSearchCV* y *Pipeline* de la librería *Scikit-learn*, se rastreó y estableció en forma sistemática y automática el valor del hiperparámetro de ajuste que el algoritmo LASSO incorpora, evaluando sucesivamente en función de éste el rendimiento del modelo como predictor de pCR mediante validación cruzada. El resultado de los *features* identificados como significativos, según los coeficientes de importancia para la estimación asignados a todos los atributos disponibles, avala la selección de *features* realizada previamente de forma manual dada la coincidencia en todos ellos. No obstante, con la selección manual se seleccionaron además otros *features* que incluyen incluso algunos clínicos.

Features seleccionados

Como síntesis, en la Tabla 4.1 se registran los *features* seleccionados 'manualmente', por medio del análisis estadístico llevado a cabo rigurosamente en la sección anterior, y los seleccionados de forma automática por medio del algoritmo operador LASSO. En el primer caso son 11 los atributos escogidos, mientras que en el segundo caso son 6. El nombre de cada *feature* se acompaña de una "T" o "P" según corresponda a información del tumor o del parénquima; y fueron etiquetados numéricamente para facilitar su referencia en el posterior análisis.

La diferencia entre los *features* seleccionados manualmente mediante análisis estadístico y automáticamente mediante LASSO, como puede deducirse de la Tabla A.1 del Apéndice-A, se debe principalmente al umbral de corte para la significancia estadística en las pruebas de hipótesis, fijado en $\alpha = 0.1$. Con una mayor exigencia en el nivel de significancia, por ejemplo $\alpha = 0.05$, entonces los atributos: '*glcm ClusterShade T*', '*firstorder Kurtosis P*', '*firstorder Minimum P*', la edad del paciente, y también '*shape Sphericity T*' que sí fue seleccionado automáticamente, no hubiesen sido identificados como indicadores significativos para la predicción deseada.

Selección manual	Selección automática
1. firstorder Kurtosis T	1. firstorder Kurtosis T
2. glcm ClusterShade T	
3. shape Sphericity T	3. shape Sphericity T
4. firstorder 10Percentile P	4. firstorder 10Percentile P
5. glszm GrayLevelNonUniformity P	5. glszm GrayLevelNonUniformity P
6. firstorder Mean P	6. firstorder Mean P
7. gldm DependenceVariance P	7. gldm DependenceVariance P
8. firstorder Kurtosis P	
9. firstorder Minimum P	
10. Edad	
11. HR	

Tabla 4.1: *Features* seleccionados: 'manualmente' mediante análisis estadístico y por algoritmo de aprendizaje automático LASSO.

4.3 Predicción de pCR en base a modelos *ML*

4.3.1 Modelos predictivos univariados

En primera instancia, se evaluaron modelos univariados para la predicción deseada. Mediante la técnica de validación cruzada *k-fold* estratificada y repetitiva, según parámetros establecidos previamente, se obtienen la media, desviación estándar e intervalos de confianza del 90% para AUC de cada modelo, cuyos resultados se visualizan gráficamente en la Figura 4.12. Como complemento para la evaluación del rendimiento de los modelos predictores se estimó también la métrica *Accuracy* (*Acc*) en todos ellos como factor discriminador en caso de que se obtenga un mismo valor AUC en los modelos evaluados según interés.

Con el fin de establecer el intervalo *rope* para comparar los modelos por *Bayesian correlated t-test* en los casos que se hace mención más adelante, y cumpliendo con la condición señalada en la Sección 3.4.3, se calculó la media para la distribución de los valores de desviación estándar que presentan los resultados AUC, siendo igual a $\mu_{desv\ std} = 0.13$. En consecuencia, se fijó *rope* en $[-0.01; 0.01]$.

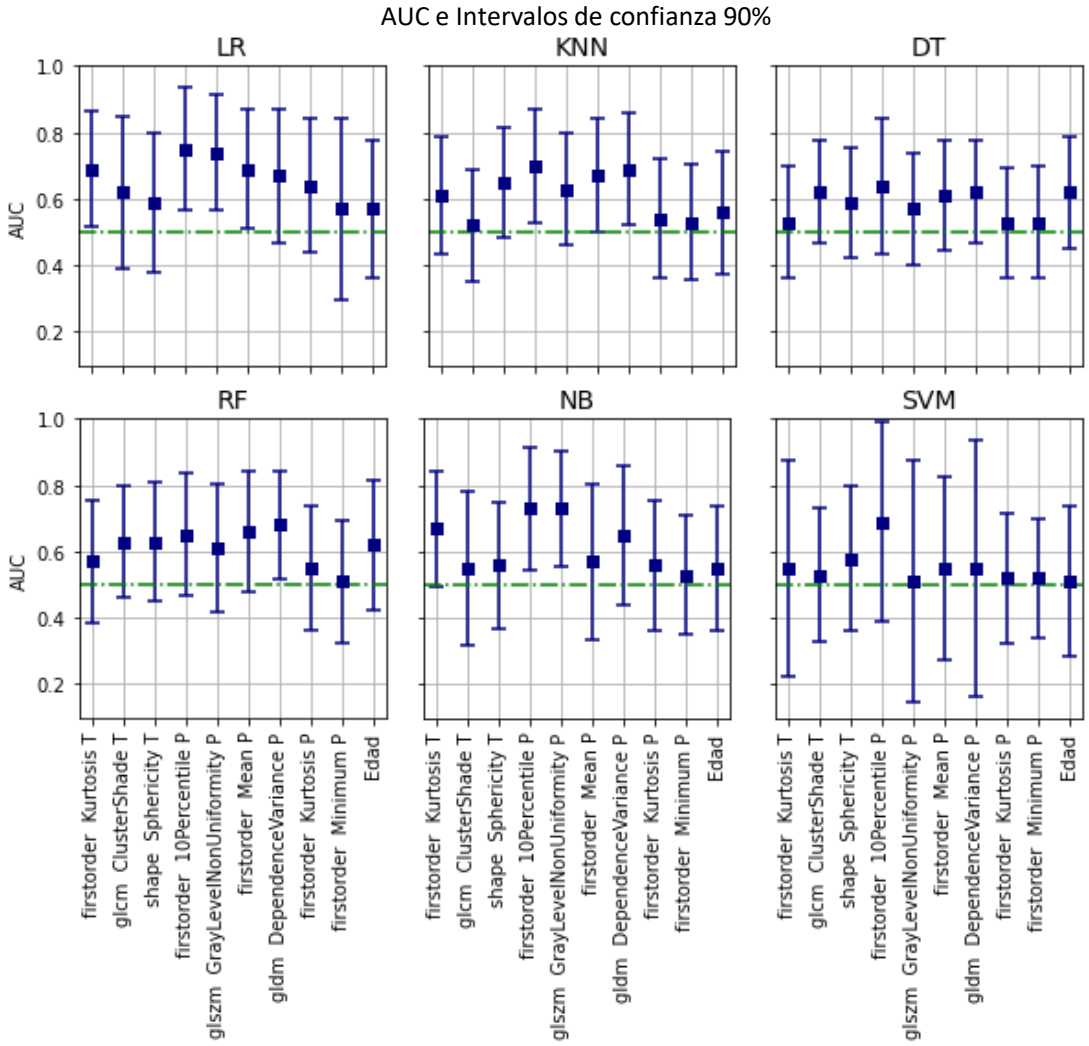


Figura 4.12: Resultados del rendimiento AUC de los distintos modelos univariados automatizados de predicción de pCR. Se identifican, además, los intervalos de confianza respectivos.

De los gráficos, se observa que el mejor modelo univariado se compone del *feature* '*firstorder 10Percentile P*' y clasificador LR, con un $AUC = 0.75\ (0.11)$ y $Acc = 0.73\ (0.07)$. El modelo conformado por el mismo algoritmo LR y por el *feature 'glszm GrayLevelNonUniformity P'* como predictor presenta un rendimiento equivalente por análisis comparativo, con un valor $AUC = 0.74\ (0.10)$; sin embargo, logra un $Acc = 0.67\ (0.06)$ apreciablemente menor. Muy similares en el rendimiento son los dos modelos univariados que consideran estos mismos dos *features* respectivamente, pero que emplean el algoritmo NB para su predicción, con un resultado $AUC = 0.73\ (0.11)$ en ambos casos. Los valores de *Acc* de estos últimos son iguales al de los modelos

señalados con LR congruentes con el *feature* respectivo. A su vez, estos cuatro modelos que se mencionan se diferencian significativamente de un modelo predictor aleatorio $AUC = 0.50$.

En caso contrario, los modelos con *features* del parénquima '*firstorder Kurtosis P*' y '*firstorder Minimum P*', indistintamente del algoritmo clasificador, presentan un rendimiento bajo equivalente a un predictor aleatorio. Lo mismo ocurre para el atributo de la edad, donde únicamente los modelos con DT y RF alcanzan apenas un $AUC = 0.62\ (0.10)$.

Por otra parte, al considerar los modelos en base a *features* del tumor, los dos mejores evaluados fueron los con algoritmos LR y NB, y mismo atributo '*firstorder Kurtosis T*'. Éstos obtuvieron rendimientos $AUC = 0.69\ (0.11)$ y $AUC = 0.67\ (0.11)$, respectivamente, logrando una diferencia estadística respecto a un modelo aleatorio. Como complemento, ambos alcanzaron un *Accuracy* equivalente entre sí, con $Acc = 0.63\ (0.07)$ y $Acc = 0.62\ (0.06)$, respectivamente.

4.3.2 Modelos multivariados predictivos

Con posterioridad, estos mismos clasificadores fueron evaluados como modelos multivariados para la predicción deseada, considerando todas las combinaciones posibles entre los 11 *features* de la matriz de entrada. Debido a la gran extensión de los resultados y sus gráficas, se presentarán los más característicos y significativos para el posterior análisis comparativo y selección del mejor modelo.

Por practicidad, en las tablas y gráficos de resultados que se exponen en adelante, los *features* y sus combinaciones son representados por los números con los que fueron etiquetados en la Tabla 4.1.

De todos los modelos generados como resultado mediante *ML*, en la Tabla 4.2 se exponen aquellos con mejor rendimiento al analizar los que consideran combinaciones de *features* únicamente de forma (Sh) y de primer orden (f.O), y en conjunto con los dos atributos clínicos. Con esto se pretende observar el rendimiento que logran los modelos para la predicción de pCR sin considerar los *features* de textura referentes al tumor y al parénquima, los cuales se informan más adelante.

Combinación tipo de *features*	Combinación *Features*	Clasificador ML	AUC	(σ)	I. C. AUC	*Acc*	(σ)
Sh + f.O	(1, 3, 4, 6, 8, 9)	RF	0.74	(0.10)	[0.58; 0.90]	0.69	(0.07)
Sh + f. O + 'edad'	(1, 3, 4, 6, 8, 9, 10)	LR	0.74	(0.10)	[0.58; 0.90]	0.69	(0.09)
Sh + f. O + clínicos	(1, 3, 4, 6, 8, 9, 10, 11)	LR	0.75	(0.10)	[0.59; 0.91]	0.73	(0.09)

Tabla 4.2: Modelos con mejor rendimiento para la predicción de pCR, conforme al clasificador y a la combinación de *features*, según los tipos de atributos que se consideran como matriz de entrada.

Para las mismas combinaciones de *features* señaladas en la tabla, se obtuvo en los modelos ajustados mediante DT y SVM un rendimiento $AUC \leq 0.65$ para los tres casos, mientras que en los con KNN y NB se logró $0.65 \leq AUC \leq 0.68$. Los modelos con clasificadores RF y LR fueron los que alcanzaron un mayor rendimiento para la predicción, con $AUC > 0.73$. En detalle, para la combinación de *features* 'Sh + f.O' se registró el modelo con algoritmo RF como el mejor; sin embargo, con LR se obtuvo un rendimiento equivalente $AUC = 0.73\ (0.10)$ y $Acc = 0.71\ (0.08)$. El modelo con clasificador RF y que incluye el atributo 'edad' para la predicción logra un rendimiento $AUC = 0.71\ (0.11)$ y $Acc = 0.67\ (0.08)$; mientras que, en el último caso, el modelo también basado en RF obtiene un $AUC = 0.73\ (0.10)$ y $Acc = 0.69\ (0.08)$.

Como consecuencia de la similitud en los resultados del rendimiento de los modelos basados en RF y LR para las tres combinaciones de *features* descritas, en la Figura 4.13 se presenta gráficamente el análisis bayesiano para la comparación respectiva entre estos modelos, con *rope* definido previamente.

De los gráficos se observa con mayor evidencia una diferencia estadística mayor entre los modelos RF y LR que consideran los *features* de forma, de primer orden y la edad para la predicción; en tanto que los modelos ajustados únicamente mediante los *features* de forma y primer orden se establecen como equivalentes estadísticamente dado que la media de la distribución posterior, obtenida por la diferencia de las distribuciones de AUC entre ambos modelos, se encuentra dentro de los límites del intervalo $rope = \pm 0.1$, y el pequeño desfase observado se encuentra dentro de la desviación estándar referida a esta misma distribución.

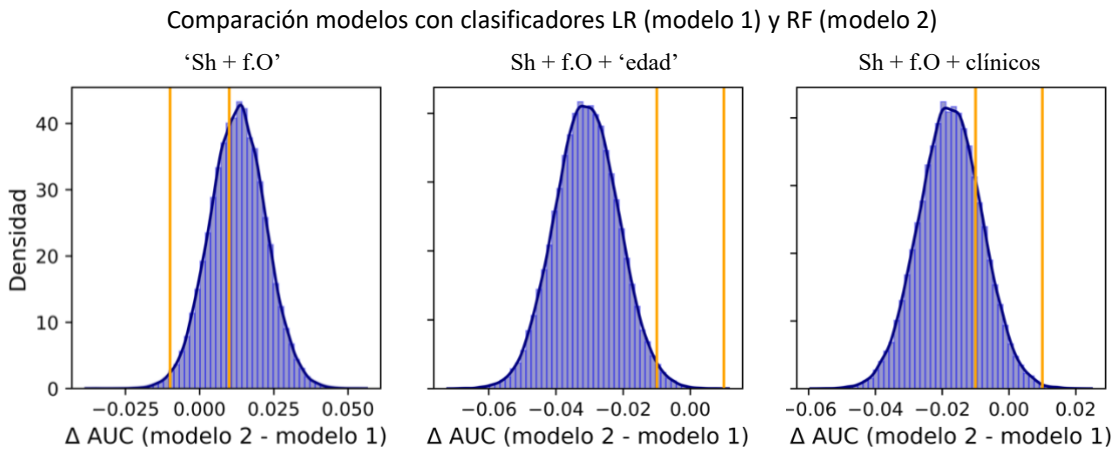


Figura 4.13: Comparación mediante análisis Bayesiano entre modelos con clasificadores LR (modelo 1) y RF (modelo 2), los cuales se configuran con las mismas combinaciones de *features*, respectivamente, en correspondencia con la Tabla 4.2.

A los modelos presentados en la tabla anterior (Tabla 4.2) se les aplicó el *Frequentist correlated t-test* para compararlos respecto a un predictor aleatorio $AUC = 0.50$, obteniendo una diferencia estadísticamente significativa ($p - value < 0{,}05$) en todos los casos.

En esta misma tabla, se aprecia como el modelo predictivo que incorpora los *features* seleccionados de forma, de primer orden y la edad presenta un rendimiento idéntico, con $AUC = 0.74\ (0.10)$ y $Acc = 0.69$, al modelo que considera los mismos *features* a excepción de la edad. Si bien entre ambos difiere el algoritmo clasificador registrado (LR y RF, respectivamente), el modelo que no incluye la edad como atributo logra un rendimiento equivalente al predecir mediante LR y RF, como resultado de la comparación entre modelos por análisis Bayesiano. La edad entonces, que se seleccionó inicialmente presentando un $p - value = 0.099$ y $AUC^{M}_{umbral} = 0.61$, es un atributo que, con esta evidencia y respaldada por los resultados de los modelos multivariados que se presentan más adelante, demostró no tener una influencia significativa en la predicción de pCR para lograr un mejor rendimiento.

En otro ámbito, de todos los modelos generados como predictores de pCR mediante *ML*, en la Figura 4.14 se muestran gráficamente los resultados de los modelos multivariados que involucran únicamente los *features* del tumor.

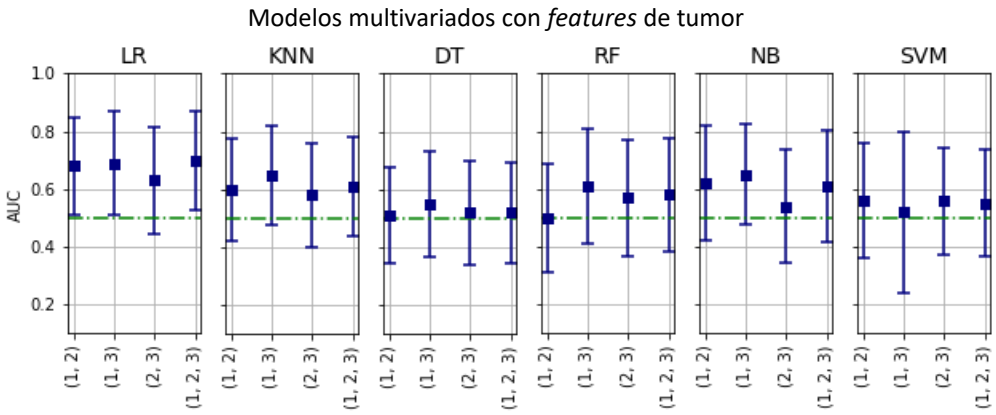


Figura 4.14: Resultados para la métrica AUC de los distintos modelos, generados por las combinaciones de los *features* únicamente del tumor y por los distintos algoritmos de clasificación para la predicción de pCR.

De los gráficos presentados, se observa que el mejor modelo predictor de pCR se logra al considerar los 3 *features* combinados para el clasificador LR, con un rendimiento de $AUC = 0.70\ (0.11)$ y $Acc = 0.67\ (0.08)$. No obstante, la combinación de *features* $(1, 3)$ para este mismo algoritmo LR presenta un resultado equivalente por comparación, con $AUC = 0.69\ (0.11)$ y $Acc = 0.67\ (0.08)$. Esta combinación, dada por '*firstorder Kurtosis T*' y '*shape Sphericity T*', se corresponde con los dos únicos *features* seleccionados automáticamente mediante LASSO.

Los resultados para los modelos con las combinaciones de *features* únicamente del parénquima se observan gráficamente en la Figura 4.15, mientras que los resultados generales, de todos los modelos que combinan todos los atributos de la matriz de entrada y los distintos algoritmos de predicción, se pueden observar gráfica y directamente en los archivos del código *Python* ya mencionados (p. 40, pie de página). Éstos no se exponen en este informe debido a su gran extensión; no obstante, en las Tablas 4.2, 4.3 y 4.4 se precisan los resultados de algunos modelos específicos según las consideraciones señaladas en la metodología, Sección 3.4.3.

De los modelos que combinan *features* del parénquima, dos equivalentes son los que logran el mejor rendimiento, con $AUC = 0.80\ (0.10)$ y $Acc = 0.74\ (0.08)$ en ambos. Uno de ellos se conforma por el algoritmo LR y combinación de *features* $(4, 5, 7)$, concordantes con '*firstorder 10Percentile P*', '*glszm GrayLevelNonUniformity P*' y '*gldm DependenceVariance P*', respectivamente; y el otro, por el algoritmo NB y combinación $(4, 5)$. Como modelo equivalente se tiene al conformado por el algoritmo LR y combinación $(4, 5)$ de *features*, con un rendimiento $AUC = 0.79\ (0.09)$ y $Acc = 0.75\ (0.07)$. Por otra parte, el predictor que alcanza el mayor rendimiento para

la combinación de todos los *features* del parénquima, descrita por $(4, 5, 6, 7, 8, 9)$, corresponde al algoritmo LR, con $AUC = 0.75\ (0.10)$ y $Acc = 0.70\ (0.09)$.

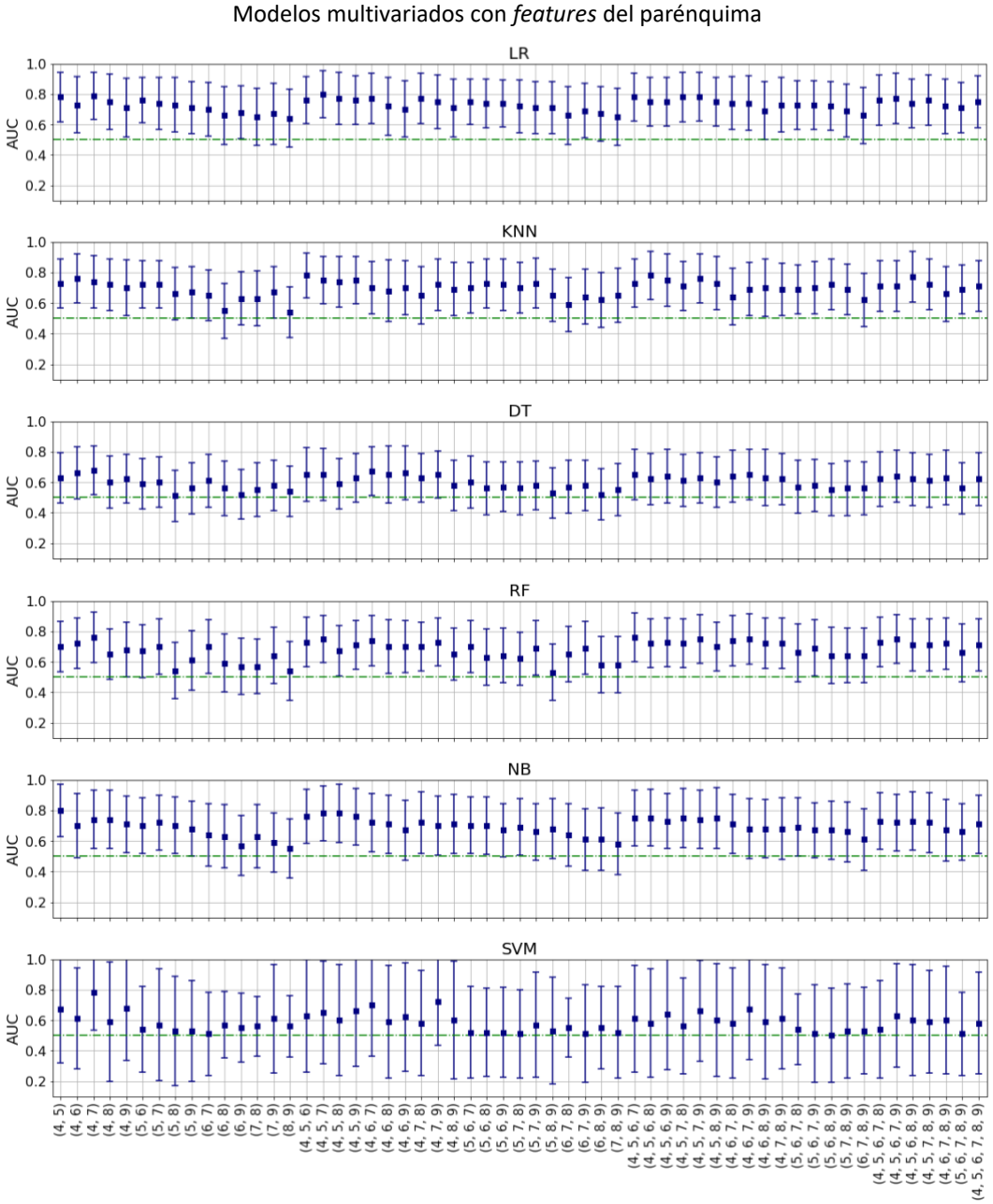


Figura 4.15: Rendimiento de los modelos multivariados basados en las combinaciones de *features* únicamente del parénquima, según los algoritmos de ML utilizados como clasificador.

Pese a que no se exponen los gráficos con los resultados de todos los modelos generados, como se explicó previamente, cabe mencionar que gran parte de aquellos basados en los clasificadores SVM y DT, para las distintas combinaciones de *features*, lograron un rendimiento bajo para la predicción de pCR, con un $AUC \leq 0.65$. Sin embargo, la mayoría de los modelos basados en los clasificadores LR, KNN y RF consiguieron un rendimiento promedio de $AUC \geq 0.7$, algunos de los cuales incluso obtuvieron $AUC \geq 0.8$.

Recapitulando, en la Tabla 4.3 se registra el algoritmo clasificador mejor evaluado, junto al rendimiento AUC y métrica *Acc*, para los modelos compuestos por la combinación de: los 3 *features* seleccionados referentes al tumor ('T'), los *features* seleccionados únicamente del parénquima ('P'), los *features* mencionados conjuntamente ('T + P'), y estos mismos en totalidad con la información clínica seleccionada ('T + P + C'). Se observa que el algoritmo LR es el que mejor se ajusta y predice pCR para las combinaciones de atributos señaladas. Todos ellos ofrecen superioridad en rendimiento respecto a un predictor aleatorio ($AUC = 0.50$), como resultado de la comparación estadística por *Frequentist correlated t-test* ($p < 0.05$), siendo el modelo que integra los 11 *features* el mejor evaluado entre los cuatro.

Información (cantidad *features*)	Clasificador ML	AUC	(σ)	I. C. AUC	*Acc*	(σ)
T (3)	LR	0.70	(0.11)	[0.53; 0.87]	0.67	(0.08)
P (6)	LR	0.75	(0.10)	[0.58; 0.92]	0.70	(0.09)
T + P (9)	LR	0.81	(0.09)	[0.66; 0.95]	0.73	(0.09)
T + P + C (11)	LR	0.83	(0.08)	[0.70; 0.97]	0.76	(0.08)

Tabla 4.3: Algoritmos ML con mejor rendimiento en modelos de predicción multivariado de pCR conformados por los atributos seleccionados, agrupados de modo independiente y combinado según correspondan al tumor (T), parénquima (P) y clínico (C).

Cabe mencionar que también existe una diferencia estadística, evaluada por análisis bayesiano mediante *rope*, respecto a los modelos generados por las mismas combinaciones de *features* según el agrupamiento señalado en la tabla y los demás algoritmos de clasificación. En forma generalizada, existe una mayor diferencia respecto a los modelos en base a SVM y DT, y en menor magnitud respecto a los ajustados mediante KNN y RF. Así, en el caso de los *features* del tumor ('T') le siguen los modelos en base a: KNN con rendimiento $AUC = 0.61\ (0.10)$ y $Acc = 0.64\ (0.08)$, y NB con $AUC = 0.61\ (0.12)$ y $Acc = 0.63\ (0.09)$, ambos estadísticamente equivalentes. En el caso de los *features* del parénquima (`P`) se presentan en segundo lugar los modelos basados en KNN, NB y RF

con un rendimiento equivalente $AUC = 0.71\ (0.10)$ y $Acc = 0.69\ (0.09)$. En el caso de la combinación de la totalidad de *features* 'T + P', el modelo en base a RF se posiciona en segundo lugar por rendimiento, con $AUC = 0.76\ (0.10)$ y $Acc = 0.72\ (0.08)$. En el último caso, donde se considera la combinación de todos los *features* seleccionados manualmente ('T + P + C'), los modelos ajustados mediante RF, con rendimiento $AUC = 0.78\ (0.10)$ y $Acc = 0.72\ (0.08)$, y mediante KNN, con $AUC = 0.77\ (0.09)$ y $Acc = 0.72\ (0.08)$, se posicionan a continuación del basado en LR, siendo ambos significativamente equivalentes por estadística.

Como ya se había señalado previamente, al entrenar y ajustar los algoritmos con todas las combinaciones posibles de los *features* referentes al parénquima, a diferencia de los *features* del tumor, el modelo que tiene mejor rendimiento no es el que considera a todos ellos, sino que corresponde a la combinación $(4, 5, 7)$. Esto mismo sucede al entrenar los clasificadores con los *features* del tumor y parénquima en conjunto ('T + P') y al integrar los datos clínicos ('T + P + C'). Estos resultados se presentan en la Tabla 4.4, registrando el algoritmo predictor y la combinación de *features* del mejor modelo al considerar los mismos agrupamientos de atributos ('Información') que en la tabla anterior.

Información	Combinación *Features*	Clasificador ML	AUC	(σ)	I. C. AUC	*Acc*	(σ)
T	(1, 2, 3)	LR	0.70	(0.11)	[0.53; 0.87]	0.67	(0.08)
P	(4, 5, 7)	LR	0.80	(0.10)	[0.64; 0.95]	0.70	(0.09)
T + P	(1, 2, 3, 4, 7)	LR	0.86	(0.08)	[0.73; 0.99]	0.73	(0.09)
T + P + C	(1, 2, 3, 4, 5, 7, 10, 11)	LR	0.87	(0.08)	[0.74; 1.00]	0.79	(0.08)

Tabla 4.4: Mejor modelo predictivo multivariado, determinado por la combinación de *features* y algoritmo clasificador, según agrupamiento de atributos para el entrenamiento y ajuste.

Predicción de pCR mediante *features* por selección manual vs selección automatizada

En la Figura 4.16 se muestran gráficamente los resultados del rendimiento AUC para los modelos construidos con los distintos clasificadores estudiados y la matriz de entrada completa para ambos casos de selección de *features*: los seleccionados manualmente mediante análisis estadístico y los seleccionados de forma automatizada mediante LASSO, con el objetivo de evaluar sus diferencias. Se recuerda que la combinación de *features*, en el primer caso, comprende los atributos del 1 al 11 según su etiquetado, y para el segundo caso corresponde a $(1, 3, 4, 5, 6, 7)$. Para ambos casos de

selección, el clasificador LR ofrece el mejor rendimiento en la predicción de pCR, con $AUC = 0.83\ (0.08)$ y $AUC = 0.84\ (0.08)$, respectivamente. Cabe destacar que, para estas mismas dos configuraciones de *features*, los modelos con clasificador RF logran un rendimiento que los posiciona en segundo lugar, con: $AUC = 0.79\ (0.10)$ y $Acc = 0.73\ (0.08)$, y $AUC = 0.78\ (0.10)$ y $Acc = 0.72\ (0.08)$, en el mismo orden correlativo anterior, siendo equivalentes entre sí por significancia estadística en ambas métricas.

Los dos modelos mencionados con mejor rendimiento, según técnica de selección de *features* y clasificador LR, son comparados mediante *Frequentist correlated t-test* entre sí, obteniendo un $p - value = 0.996$, y respecto a un predictor aleatorio, obteniendo $p - value < 0.001$ en ambos. A su vez, se comparan también entre sí mediante *Bayesian correlated t-test*, cuyo resultado se muestra gráficamente en la Figura 4.17. Se observa, de manera evidente, la equivalencia estadísticamente significativa que existe entre ellos (diferencia menor o igual al 1% según *rope*), con un valor exacto de $\Delta AUC = 0.001$ para la media de la distribución posterior.

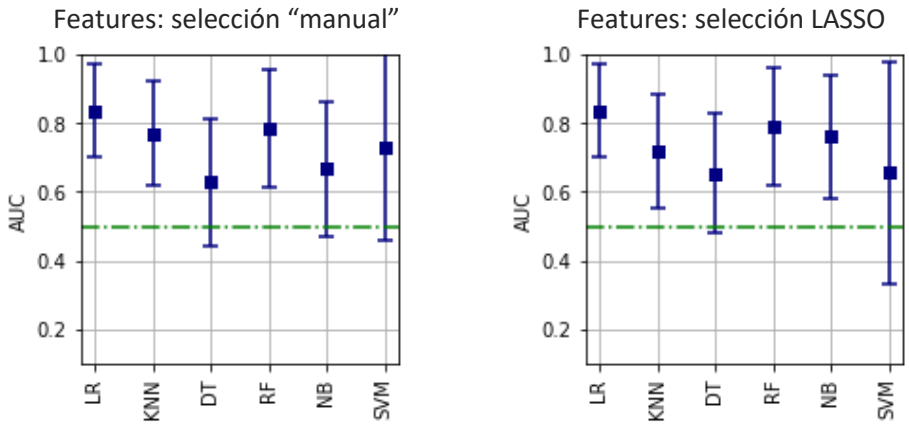


Figura 4.16: Rendimiento de los modelos de predicción de pCR construidos con la matriz de entrada completa de *features* seleccionados manualmente, mediante estadística inferencial, y de manera automatizada, mediante LASSO.

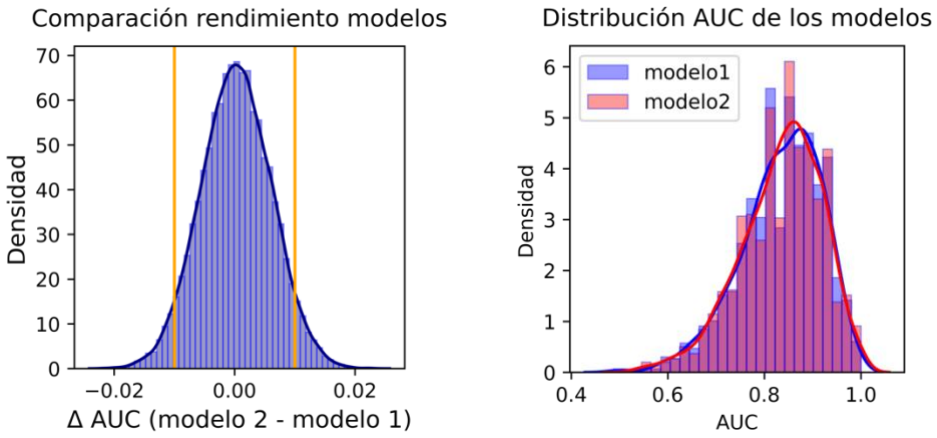


Figura 4.17: Comparación del rendimiento entre el modelo 1: *features* por selección manual y el modelo 2: *features* por selección LASSO, ambos con clasificador LR, mediante el *Bayesian correlated t-test* con $rope = [-0.01 ; 0.01]$.

En la Figura 14.18 se caracterizan las distribuciones, mediante un gráfico *boxplot*, de los coeficientes que estos dos modelos descritos le otorgan a los *features* involucrados respectivamente en su aprendizaje y entrenamiento como predictor de pCR. Éstos indican la importancia que tienen en la predicción de pCR al entrenar y ajustar de forma conjunta el modelo, y su signo se correlaciona con el sentido del valor predictivo que entrega, es decir, si es directo o inverso. Se observa de los gráficos que todos los *features* que poseen un coeficiente con valor absoluto mayor o igual a 0.5 son coincidentes entre ambos modelos, a excepción del atributo por selección manual HR correspondiente al estado hormonal. Así, el *feature* '*firstorder 10Percentile P*' es el que posee mayor relevancia en ambos modelos, seguido de '*gldm DependenceVariance P*', con media entre 0.8 y 1.0 en sentido inverso, y de los *features* '*firstorder Kurtosis T*' y '*shape Sphericity T*', cuyas medias se encuentran entre 0.5 y 0.8. Con una importancia equivalente para la predicción de pCR entre los dos modelos se encuentran los *features* '*glszm GrayLevelNonUniformity P*' y '*firstorder Mean P*', respectivamente. Los atributos restantes, que corresponden a los de selección manual, poseen todos una media con valor absoluto menor a 0.5 en sus coeficientes de importancia. Esto se condice con el hecho abordado anteriormente, en que si el nivel de significancia en las pruebas para la selección de *features* hubiese sido de 5%, entonces éstos no hubiesen sido escogidos.

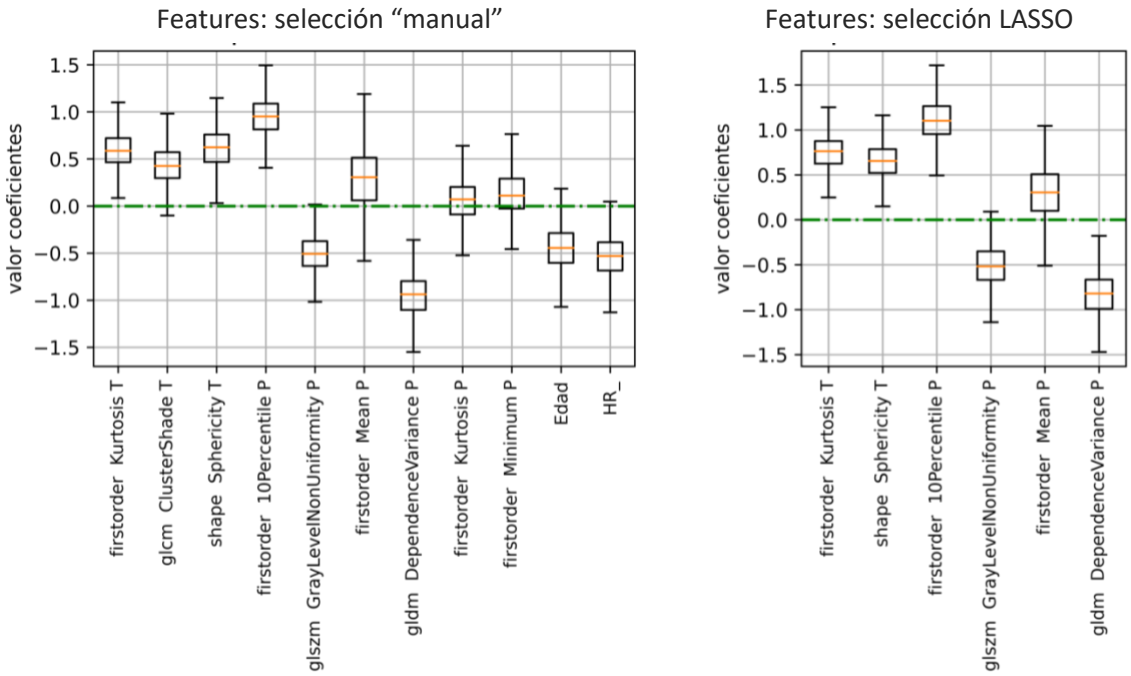


Figura 4.18: Distribución de los valores del peso de los coeficientes de los *features* para ambos modelos, según técnica de selección de *features* y clasificador LR, finalizado su entrenamiento.

Mejor modelo

En la Tabla 4.5 se resume las características propias del modelo seleccionado como el mejor para la predicción de pCR en pacientes con cáncer de mama tratadas con NACT, según los resultados tanto del rendimiento mismo en la predicción, como del análisis comparativo entre modelos equivalentes. La prueba *Frequentist correlated* t-test ratifica al modelo señalado como un buen modelo predictor, mientras que la prueba *Bayesian correlated t-test* lo determinó equivalente al modelo con el mismo algoritmo clasificador y configuración de *features,* pero que incluye la edad como atributo, indicado en la Tabla 4.4. En la discusión del próximo capítulo de la tesis se explica la razón por la que finalmente se elige, de entre los dos modelos mencionados, al que no incluye la edad como atributo.

Luego, en la Figura 4.19, se presenta el gráfico de la curva ROC para este modelo seleccionado, obtenido mediante validación cruzada estratificada y repetida conforme a los parámetros establecidos en el estudio llevado a cabo.

Features del modelo	Clasificador	AUC (σ)	*Acc* (σ)	p-value Freq. correlated test
1. firstorder Kurtosis T 2. glcm ClusterShade T 3. shape Sphericity T 4. firstorder 10Percentile P 5. glszm GrayLevelNonUniformity P 7. gldm DependenceVariance P 11. HR	LR	0.87 (0.08)	0.78 (0.08)	<0.001

Tabla 4.5: Modelo con mayor rendimiento AUC para la predicción de pCR a NACT en pacientes con cáncer de mama.

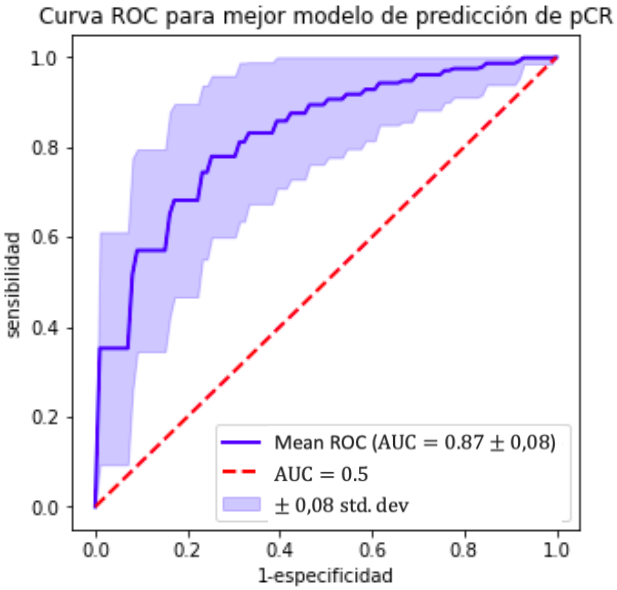


Figura 4.19: Curva ROC, con su desviación estándar, para el modelo seleccionado como el mejor entre los evaluados para la predicción de pCR del tumor a NACT en pacientes con cáncer de mama.

En síntesis, y analizando los resultados, los modelos univariados mediante *ML* lograron un rendimiento $AUC \leq 0.7$ en los casos con *feature* del tumor, y $AUC \leq 0.75$ en los del parénquima mamario, mientras que con los atributos clínicos seleccionados se alcanzó un $AUC \leq 0.67$.

Del análisis multivariado posterior, el rendimiento que logran los modelos conformados por todos los *features* de forma y primer orden de ambos ROIs es similar, con $AUC \leq 0.74$, obteniendo el mayor valor los modelos con clasificador RF y LR. Si se incluyen también todos los *features* texturales (2º orden) para la predicción, el rendimiento de los modelo aumenta hasta $AUC \leq 0.81$; mientras que la incorporación adicional de los dos atributos clínicos permite alcanzar un $AUC \leq 0.83$.

No considerar la combinación estricta de todos los *features* seleccionados mediante el análisis estadístico realizado, permitió mejorar el desempeño en la predicción de algunos de los modelos detallados. Así, en el modelo con LR que considera los *features* del parénquima mamario se eleva el AUC de 0.75 a 0.80 para la combinación $(4, 5, 7)$, correspondientes a '*firstorder 10Percentile P*', '*glszm GrayLevelNonUniformity P*' y '*gldm DependenceVariance P*'; mientras que para los *features* del tumor, el modelo con mayor rendimiento sí se obtuvo con la combinación de los únicos 3 seleccionados y clasificador LR. El modelo que predice pCR a partir de ambos conjuntos de *features* (T+P) mediante LR aumenta su rendimiento de 0.81 a 0.86 al considerar la combinación $(1, 2, 3, 4, 7)$ en vez de los 9 *features* en conjunto. En el último caso en que se consideran todos los tipos de atributos (T+P+C) se logra aumentar el AUC desde 0.83, obtenido con los 11 *features*, a 0.87 para la combinación $(1, 2, 3, 4, 5, 7, 11)$ que constituye al mejor modelo escogido.

Cabe destacar que el clasificador LR es el que mostró tener el mejor desempeño en todos los modelos señalados, no obstante, algunos de éstos tienen equivalencia estadísticamente significativa con los basados en RF y NB para la misma combinación de *features*, según fueron señalados en el capítulo.

CAPÍTULO 5

Discusión

En su publicación [28], Ashirbani *et al.* evidencian el impacto que tienen los parámetros del scanner de resonancia magnética en los *features* o información radiómica del tejido fibroglandular y del tumor en pacientes con cáncer de mama. En su investigación, las segmentaciones de estos ROIs se realizaron en las secuencias T1w no saturada en grasa y T1w saturada en grasa post-contraste, respectivamente, y se registraron posteriormente en la secuencia DCE-MRI para extraer los *features* directamente del subconjunto de secuencias post-contraste, sin normalizar. Dado que los parámetros influyentes fueron: el fabricante del scanner, la intensidad del campo magnético de este mismo y el grosor del corte, es que para esta tesis se seleccionaron desde la colección *"ISPY-1"* del repositorio TCIA a las pacientes que presentan la misma información para estos parámetros, de un mismo centro médico, registrada en la Tabla 3.1. Esta selección se realizó justamente para evitar errores por diferencias en las imágenes debido a estos parámetros, a pesar de que la secuencia de imágenes desde la cual se extrajeron los *features* corresponde a una sustracción entre secuencias DCE-MRI con la finalidad de normalizarla y generar un modelo predictivo independiente de estos mismos.

Del proceso de segmentación de las áreas de interés (tumor y parénquima mamario), se lograron establecer protocolos convenientes a modo de estandarizar indicadores y criterios para las demarcaciones, lo que permitió su reproducibilidad en las distintas pacientes para su posterior análisis.

Bajo las premisas de reproducibilidad y replicabilidad del procedimiento, los umbrales de corte escogidos en el proceso, que condicionan la selección de los pixeles y/o vóxeles para las segmentaciones, y la secuencia de imágenes *Dynamic-Sustr*, desde donde se obtienen los *features*, corresponden a valores relativos respecto a su propia escala de grises, según se detalla en los protocolos de la Sección 3.2.1. Esto permitiría aplicar los protocolos desarrollados a alguna otra

cohorte de pacientes con la misma enfermedad para generar nuevos modelos de predicción de pCR a NACT, en el caso que registren la información pCR o no-pCR, o incluso utilizar el mismo modelo entrenado y ajustado en esta tesis sobre otra cohorte externa para predecir la respuesta en las pacientes. En este último caso, y teniendo el registro del estado de respuesta del tumor al tratamiento, el modelo predictivo aplicado puede testearse al calcular el 'error de generalización', mencionado en la Sección 2.4.3, mediante estos datos invisibles o no observados.

Si bien parte de estos protocolos se respaldan con las referencias señaladas en la metodología, no se contó con la asistencia de un médico radiólogo especialista para la aprobación de éstos, ni tampoco para verificar y corregir las segmentaciones realizadas en las imágenes para este estudio. Se menciona esta desventaja, ya que en la bibliografía reseñada que involucra segmentación de un ROI sí se cuenta con la participación de un médico o técnico en imagenología para el reconocimiento y evaluación de la lesión, encargado de validar y/o corregir manualmente los detalles bajo una metodología de segmentación semiautomatizada y supervisada. Es fundamental que el tumor y parénquima mamario sean representados lo más preciso posible mediante los pixeles y vóxeles seleccionados en la segmentación, ya que algunos *features* radiómicos podrían presentar mayor sensibilidad frente a las variaciones en ésta y, como consecuencia, diferir los *features* seleccionados debido al AUC en la predicción de pCR que éstos logren. Por ello, estudios [29, 30] en pacientes con cáncer de mama tienen como principal objetivo optimizar las técnicas de segmentación en imágenes MRI mediante algoritmos automatizados e inteligencia artificial (IA). En particular, en la investigación reseñada de Zhang J. *et al.* (2023) se desarrolla un asistente IA para segmentar automáticamente tumores de mama, teniendo como pasos previos: segmentación de mamas y distinción automática entre mama normal y anormal (lesionada). En éste, la información radiómica se obtiene de las secuencias normalizadas DCE-MRI: pre-contraste y postcontraste en distintas fases, siendo el modelo entrenado en un conjunto muy grande de datos con un transformador espaciotemporal especialmente diseñado para fines de diagnóstico.

Del análisis para la selección de *features*, los atributos radiómicos significativos para la predicción univariada de pCR a NACT son coherentes con estudios similares del tema [31–33], los que presentan algunas diferencias en la metodología respecto a la segmentación y a la extracción de *features* según secuencias DCE-MRI. El tumor canceroso manifiesta robustez respecto a su morfología, resultando como significativos estadísticamente en este estudio los *features* referentes a la esfericidad en su forma, a la curtosis de la distribución de intensidad de los vóxeles que lo componen, y la asimetría y uniformidad en su textura respecto a la media de la matriz de coocurrencia del nivel de gris. Por otro

lado, para el parénquima mamario se seleccionaron 4 *features* de primer orden y 2 texturales, que representan la evaluación de los vóxeles en su conjunto como una suma de los conglomerados que lo componen.

La edad, que se seleccionó inicialmente presentando un $p - value = 0.099$ y $AUC^{M}_{umbral} = 0.61$, es un atributo que demostró no tener una mayor influencia en los modelos de predicción de pCR para lograr un mejor rendimiento, deducción que es respaldada por otros autores como Sasanpour [34].

Velasco *et al.* [35], en su estudio para evaluar esquemas de quimioterapia y determinar la respuesta pCR a NACT de la lesión analizó resultados en función del subtipo molecular e histología, concluyendo que pacientes con fenotipo basal (HR-) y HER2+/ER- logran pCR en una tasa mayor que los pacientes con Luminal A y B. Este resultado, congruente con algunas de las publicaciones previamente reseñadas que incorporan datos clínicos, valida la selección del estado hormonal (HR) como atributo significativo para los modelos de predicción de pCR en este estudio, ya que logra discriminar entre ambas clases según su estado positivo o negativo. De su análisis estadístico para la predicción univariada se obtuvo un AUC negativo, lo que se aprecia también en el gráfico '*Features*: selección "manual" ' de la Figura 4.18 al presentar un coeficiente de importancia estimado de -0.5 (negativo). Así, si se tiene HR+, entonces el modelo ideal lo predice como clase 0 equivalente a non-pCR, y a HR- como clase 1 o pCR. En términos de grupos moleculares, y de acuerdo a la clasificación detallada en la Sección 2.1.1, este atributo permite discriminar entre ambas clases según pacientes que pertenecen a los grupos 1 y 2 de los que pertenecen a los grupos 3 y 4.

Aunque no se haya evidenciado tan claramente un único modelo como el mejor entre todos los generados debido a la equivalencia estadísticamente significativa que se presenta entre algunos que combinan de manera diversa los *features* y algoritmos, sí se logró identificar finalmente a uno por sobre los demás, detallado en la Tabla 4.5, cumpliendo así el objetivo general de esta tesis. Los atributos que integran este modelo concuerdan justamente con los que presentaron un coeficiente de importancia con valor absoluto ≥ 0.4 en el modelo constituido por LR y todos los atributos seleccionados 'manualmente' (Figura 14.18). Y aunque este mejor modelo no incluye todos los *features* seleccionados mediante el análisis estadístico, sí integra atributos clínicos y radiómicos de primer y segundo orden, demostrando una mejora en el rendimiento de la predicción respecto a los modelos analizados de forma particular según el tipo de *features* que éstos introduzcan.

El valor $AUC = 0.87\ (0.08)$ en la métrica de evaluación del rendimiento para el mejor modelo escogido (Tabla 4.5) es comparable con resultados obtenidos en otras investigaciones [36, 37]. Ambos reportes reseñados resumen estudios para la predicción de respuesta tumoral (pCR definido bajo distintos criterios) al tratamiento de quimioterapia neoadyuvante en cáncer de mama basada en información radiómica obtenida de imágenes MRI. El primero expone los resultados AUC de validación para los modelos multivariados que se detallan, los que oscilan entre 0.53 y 0.94; además, se destaca que la regresión logística es el modelo elegido con mayor frecuencia y que no se identificaron *features* idénticos en el análisis multivariado para la configuración de los modelos. En particular, al comparar el modelo desarrollado (Tabla 4.5) con uno de los modelos publicado por W. Li *et al* en su estudio [38], el cual combina únicamente *features* extraídos de imágenes MRI de otra serie de ensayos I-SPY[17] logrando un $AUC = 0.81\ (0.76, 0.86)$, se observa que se logró un rendimiento mayor para la predicción de pCR en esta tesis. Del mismo modo, en un rango de valores AUC más acotado y elevado, entre 0.72 y 0.97, se encuentran los resultados para el rendimiento de algunos modelos multivariados que utilizan un aprendizaje profundo (*deep learning*) para la predicción de pCR en cáncer de mama, detallados en el reporte de Khan N *et al.* [39].

La elección entre algoritmos clasificadores y la generación de modelos de predicción no solo deben considerar el alto rendimiento en el aprendizaje, sino que también una alta replicabilidad, como ya se ha declarado, facilitando la reproducción de resultados y reduciendo una posible sobre búsqueda y ajuste. Para ello, R. Bouckaert y E. Frank [40] sostienen que la validación cruzada repetida ha demostrado ser un buen método por sobre otros, con un resultado óptimo de 10 iteraciones (*fold*) repetido 10 veces en su estudio con distintas cohortes de tamaño entre 100 y 1000 pacientes. En esta tesis se hizo una exploración de prueba con distintas combinaciones de valores para establecer los parámetros, fijados finamente en 3 - *fold* (iteraciones) y 500 repeticiones, que garantizaran replicabilidad. Por el contrario, en el caso de los hiperparámetros implicados en los algoritmos utilizados no se realizó ningún tipo de exploración o rastreo para fijar sus valores, sino que se utilizaron las configuraciones predeterminadas que de igual manera eran consistentes para el aprendizaje de interés. Este hecho constituye una oportunidad de mejora para los modelos desarrollados, ya que optimizar el valor de éstos de acuerdo a los requerimientos particulares en cada clasificador podría incrementar significativamente el rendimiento de los modelos para la predicción de pCR.

[17] https://www.ispytrials.org/

Se presentaron limitaciones en este estudio llevado a cabo, algunas de las cuales podrían representar una mayor repercusión en los resultados obtenidos, tanto en la selección de *features* como en el rendimiento de los modelos de predicción.

Las secuencias T1w de la colección '*ISPY1'* se adquirieron como herramienta para un diagnóstico general de la enfermedad y para la localización del tumor, y el espaciado que se tiene entre las imágenes de corte axial es considerable (10 mm), de modo que la interpolación y proyección que se realiza para definir la segmentación volumétrica del parénquima mamario, y que posteriormente se registra sobre la secuencia *Dynamic-Sustr*, podría no ser tan exacta con respecto a su forma real. Lo ideal para optimizar esta segmentación sería contar con cortes menos espaciados, como es en el caso de la adquisición de la secuencia DCE-MRI en plano sagital; sin embargo, esto le implica al centro médico invertir mayor tiempo y recursos, situación de la cual se debe realizar balance de costo-beneficio. Otra opción sería profundizar y estudiar el método para que la interpolación entre las imágenes represente lo más preciso posible la morfología de esta región de interés.

Otras limitaciones presentes son: el tamaño reducido de la cohorte en estudio, de modo que no es posible separar un set de datos para el testeo del modelo final seleccionado restringiendo la cantidad de ejemplos de los que aprende el algoritmo para entrenar y ajustar su predicción a futuros datos, y la proporción entre las clases pCR y no-pCR, con 20 y 39 pacientes respectivamente. Aunque la razón entre ambas clases (1:2) no se considera desbalanceada [41], al no contar con una cohorte numerosa, lo ideal es que fuesen lo más equitativas posible ($\cong 1:1$).

Si bien estos dos aspectos no fueron un impedimento para llevar a cabo el estudio y considerar los resultados como aceptables, ya que se cumplieron los requerimientos de razón entre clases positiva/negativa ya señalado y de cantidad de *features* seleccionados significativos acorde a la cantidad de muestras [18], y aunque el método de remuestreo de validación cruzada estratificada repetida ha demostrado ser una herramienta óptima para compensar este tipo de inconvenientes en la ejecución del entrenamiento y validación de modelos, siempre será preferible contar con un mayor tamaño de cohorte y lo más balanceado posible entre clases. Un método alternativo para abordar el problema presente de cohortes reducidas, como trabajo futuro, es utilizar la técnica de sobremuestreo de minorías sintéticas (SMOTE) o SMOTE modificado [42] para sintetizar datos de la clase minoritaria y así aumentar su tamaño estableciendo mayor balance entre clases.

Por otra parte, contar con una cohorte inicial numerosa [43] permitiría adicionar al estudio el desarrollo de modelos basados en redes neuronales, como los señalados en el reporte [39].

A modo de complementar el trabajo realizado, y tal como se comentó inicialmente en la discusión, faltó considerar una cohorte externa para poder determinar el 'error generalizado' del mejor modelo

seleccionado, permitiendo ratificarlo como un buen predictor de pCR a NACT para otras cohortes de pacientes con cáncer de mama con datos desconocidos o invisibles para éste.

CAPÍTULO 6

Conclusión y trabajo futuro

El estudio de la información radiómica de imágenes médicas y su aplicación mediante el uso de aprendizaje automático en la predicción de la respuesta tumoral al tratamiento de quimioterapia neoadyuvante en pacientes con cáncer de mama se ha venido realizando desde hace algún tiempo con énfasis en la optimización de los procesos que permitan lograr modelos con mayor rendimiento, adecuados al interés, y de utilidad para el cuerpo médico y el paciente como herramienta para la toma de decisiones clínicas y personales.

Con el fin de establecer un procedimiento genérico, semiautomatizado y no supervisado de segmentación de las áreas de interés, se logró desarrollar y definir en esta tesis protocolos que permitieron su replicabilidad en las 59 pacientes de la cohorte, con potencial aplicación en otras con la misma enfermedad.

La información radiómica del tumor y del parénquima mamario se obtuvo de la secuencia *Dynamic-Sustr*, resultante de la sustracción entre la primera secuencia adquirida post-contraste y la secuencia base (pre-contraste). Fueron 11 los atributos seleccionados a partir del análisis estadístico univariado, que incluyen *features* de distinto tipo (forma, primer y segundo orden) del tumor y del parénquima mamario, además de clínicos.

El modelo univariado semiautomatizado que logró el mejor rendimiento se compone del *feature* '*firstorder 10Percentile P*' y el clasificador de regresión logística LR, con un $AUC = 0.75\ (0.11)$ y $Acc = 0.73\ (0.07)$. Por otra parte, de todos los modelos multivariados generados a partir de las combinaciones posibles entre los 11 atributos y 6 algoritmos supervisados (LR, DT, RF, NB, KNN y SVM), se seleccionó finalmente como mejor modelo al compuesto por el clasificador LR y por los *features*: '*firstorder Kurtosis T*', '*glcm ClusterShade T*', '*shape Sphericity T*', '*firstorder*

10Percentile P', '*glszm GrayLevelNonUniformity P*', '*gldm DependenceVariance P*' y HR, con un rendimiento de $AUC = 0.87\ (0.08)$ y $Acc = 0.78\ (0.08)$.

El clasificador LR fue el que tuvo mejor desempeño en los modelos analizados de manera particular, presentando algunos de ellos una equivalencia estadísticamente significativa en el rendimiento AUC respecto a los conformados por los clasificadores NB y RF, principalmente.

Con este resultado se cumple el objetivo y se establece un modelo radiómico-clínico semiautomatizado con potencial para predecir la respuesta tumoral a NACT, válido para esta cohorte de pacientes reducido dado que no ha demostrado su aplicabilidad en nuevos pacientes. Su nivel de rendimiento, que es comparable respecto a otros resultados de modelos similares (mas no iguales) en la bibliografía, podría mejorar aún más al solventar las limitaciones que éste presenta.

Para ello, y como trabajo futuro, se propone realizar el estudio con una cohorte de mayor tamaño (sobre 200 datos) para dividirla y conformar un set de datos para entrenamiento/validación y un set de datos para testeo, en una proporción 4:1 respetando la razón entre clases; al mismo tiempo, que dicha cohorte se encuentre equilibrada entre clases o cumpla con un proporción de 1:2 entre ellas. Se propone también, al contar con una numerosa cohorte, generar un modelo basado en redes neuronales como algoritmo predictor para su comparación.

Para un modelo más generalizado, se propone también incorporar información de pacientes con la enfermedad en distintos estadíos, integrando así esta variable como atributo y lograr un modelo predictor de pCR a NACT más universal.

Bibliografía

1. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J.R., Cardoso, F., Siesling, S., Soerjomataram, I.: Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast*. 66, 15–23 (2022). https://doi.org/10.1016/j.breast.2022.08.010

2. Chhikara, B.S.; Parang, K. Global Cancer Statistics 2022: The Trends Projection Analysis. *Chem. Biol. Lett*. 10, 451 (2023). URN:NBN:sciencein.cbl.2023.v10.451

3. Documento: Plan Nacional de Cáncer 2018-2028. *Ministerio de Salud, Gobierno de Chile* (2018). https://www.minsal.cl/.../.. Plan-nacional-de-cancer-web.pdf

4. Román Guindo, A., Martí Álvarez, C., Hardisson Hernáez, D., de Santiago García, F.J., Sánchez Méndez, J.I.: Evaluation of pathological response to neoadjuvant chemotherapy in the breast and axilla according to molecular phenotypes of breast cancer. *Revista de Senologia y Patologia Mamaria*. 29, 120–124 (2016). https://doi.org/10.1016/j.senol.2016.05.002

5. Spring, L.M., Fell, G., Arfe, A., Sharma, C., Greenup, R., Reynolds, K.L., Smith, B.L., Alexander, B., Moy, B., Isakoff, S.J., Parmigiani, G., Trippa, L., Bardia, A.: Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis. *Clinical Cancer Research*. 26, 2838–2848 (2020). https://doi.org/10.1158/1078-0432.CCR-19-3492

6. Khalifa, F., Soliman, A., El-Baz, A., Abou El-Ghar, M., El-Diasty, T., Gimel'Farb, G., Ouseph, R., Dwyer, A.C.: Models and methods for analyzing DCE-MRI: A review. *Med Phys*. 41, (2014). https://doi.org/10.1118/1.4898202

7. Chao You, Weijun Peng and Yajia Gu: Association Between Background Parenchymal Enhancement and Pathologic Complete Remission Throughout the Neoadjuvant Chemotherapy in Breast Cancer Patients. *Elsevier Enhanced Reader*. 10 (5), 786-792 (2017). https://dx.doi.org/10.1016/j.tranon.2017.07.005

8. O'flynn, E.A., Collins, D., D'Arcy, J., Schmidt, M., De Souza, N.M.: Multi-parametric MRI in the early prediction of response to neo-adjuvant chemotherapy in breast cancer: Value of non-modelled parameters. *Eur J Radiol*. 85, 837–842 (2016). https://doi.org/10.1016/j.ejrad.2016.02.006

9. Aghaei, F., Tan, M., Hollingsworth, A.B., Zheng, B.: Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy. *Journal of Magnetic Resonance Imaging*. 44, 1099–1106 (2016). https://doi.org/10.1002/jmri.25276

10. Liu, Z., Li, Z., Qu, J., Zhang, R., Zhou, X., Li, L., Sun, K., Tang, Z., Jiang, H., Li, H., Xiong, Q., Ding, Y., Zhao, X., Wang, K., Liu, Z., Tian, J.: Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. *Clinical Cancer Research*. 25, 3538–3547 (2019). https://doi.org/10.1158/1078-0432.CCR-18-3190

11. WHO: World Health Statistics 2023 Monitoring health for the SDGs Sustainable Development Goals. *Geneva: World Health Organization* (2023). Licence: CCBY-NC-SA 3.0 IGO. ISBN 978-92-4-007432-3

12. Sims, A.H., Howell, A., Howell, S.J., Clarke, R.B.: Origins of breast cancer subtypes and therapeutic implications. *Nature Clinical Practice Oncology*. 4 (9), 516-526 (2007). https://doi.org/10.1038/ncponc0908

13. Von Minckwitz, G., Untch, M., Blohmer, J.U., Costa, S.D., Eidtmann, H., Fasching, P.A., Gerber, B., Eiermann, W., Hilfrich, J., Huober, J., Jackisch, C., Kaufmann, M., Konecny, G.E., Denkert, C., Nekljudova, V., Mehta, K., Loibl, S.: Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *Journal of Clinical Oncology*. 30, 1796–1804 (2012). https://doi.org/10.1200/JCO.2011.38.8595

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

15. Frost, J.: Hypothesis Testing – an intuitive guide for making data driven decision. *Statistics By Jim Publishing* (2020).

16. Haukoos, J.S., Lewis, R.J.: Advanced statistics: Bootstrapping confidence intervals for statistics with "difficult" distributions. *Academic Emergency Medicine*. 12, 360–365 (2005). https://doi.org/10.1197/j.aem.2004.11.018

17. Bradley, A.E.: The use of the area under the (ROC) curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 30, 1145-1159 (1997). https://doi.org/10.1016/S0031-3203(96)00142-2

18. Cruz, J.A., Wishart, D.S.: Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*. 2, 59-77 (2006). https://doi.org/10.1177/11769351060020

19. Kuhn, M., Johnson, K.: Applied predictive modeling. *Springer* (2013). https://doi.org/10.1007/978-1-4614-6849-3

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

21. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *Journal of Machine Learning Research*. 18, 1-36 (2017). jmlr.org/papers/v18/16-305.html

22. Kruschke, J.K.: Rejecting or Accepting Parameter Values in Bayesian Estimation. *Adv Methods Pract Psychol Sci*. 1, 270–280 (2018). https://doi.org/10.1177/2515245918771304

23. Hylton, N.M., Esserman, L.J., Gatsonis, S.C., Pisano, E., Weatherall, P., Schnall, M., Rosen, M., Lehman, C., Polin, S., Schmidt, R., Newstead, G., Morris, E., Bolan, P., Garwood, M., Bernreuter, W.: American College of Radiology Imaging Network ACRIN 6657 Contrast-Enhanced Breast MRI and MRS for Evaluation of Patients Undergoing Neoadjuvant Treatment for Locally Advanced Breast Cancer. (2010)

24. TCIA: I-SPY1 Data Sharing Dictionary- TCIA Data Collection, modification/review actions.

25. Chitalia, R., Pati, S., Bhalerao, M., Thakur, S., Jahani, N., Belenky, J. V., McDonald, E.S., Gibbs, J., Newitt, D., Hylton, N., Kontos, D., & Bakas, S.: Expert tumor annotations and radiomic features for the ISPY1/ACRIN 6657 trial data collection [Data set]. *The Cancer Imaging* (2021). https://doi.org/10.7937/TCIA.XC7A-QT20

26. Aghaei, F., Tan, M., Hollingsworth, A.B., Qian, W., Liu, H., Zheng, B.: Computer-aided breast MR image feature analysis for prediction of tumor response to chemotherapy. *Med Phys*. 42, 6520–6528 (2015). https://doi.org/10.1118/1.4933198

27. Gonzalez, R.C., Woods, R.E.: Digital image processing. Prentice Hall (2002)

28. Saha, A., Yu, X., Sahoo, D., Mazurowski, M.A.: Effects of MRI scanner parameters on breast cancer radiomics. *Expert Syst Appl*. 87, 384–391 (2017). https://doi.org/10.1016/j.eswa.2017.06.029

29. Wu, S., Weinstein, S.P., Conant, E.F., Schnall, M.D., Kontos, D.: Automated chest wall line detection for whole-breast segmentation in sagittal breast MR images. *Med Phys*. 40 (2013). https://doi.org/10.1118/1.4793255

30. Zhang, J., Cui, Z., Shi, Z., Jiang, Y., Zhang, Z., Dai, X., Yang, Z., Gu, Y., Zhou, L., Han, C., Huang, X., Ke, C., Li, S., Xu, Z., Gao, F., Zhou, L., Wang, R., Liu, J., Zhang, J., Ding, Z., Sun, K., Li, Z., Liu, Z., Shen, D.: A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework. *Patterns*. 4 (2023). https://doi.org/10.1016/j.patter.2023.100826

31. Zeng, Q., Ke, M., Zhong, L., Zhou, Y., Zhu, X., He, C., Liu, L.: Radiomics Based on Dynamic Contrast-Enhanced MRI to Early Predict Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Therapy. *Acad Radiol*. 30, 1638–1647 (2023). https://doi.org/10.1016/j.acra.2022.11.006

32. Cain, E.H., Saha, A., Harowicz, M.R., Marks, J.R., Marcom, P.K., Mazurowski, M.A.: Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast Cancer Res Treat*. 173, 455–463 (2019). https://doi.org/10.1007/s10549-018-4990-9

33. Li, Y., Fan, Y., Xu, D., Li, Y., Zhong, Z., Pan, H., Huang, B., Xie, X., Yang, Y., Liu, B.: Deep learning radiomic analysis of DCE-MRI combined with clinical characteristics predicts pathological complete response to neoadjuvant chemotherapy in breast cancer. *Front Oncol*. 12 (2023). https://doi.org/10.3389/fonc.2022.1041142

34. Sasanpour, P., Sandoughdaran, S., Mosavi-Jarrahi, A., Malekzadeh, M.: Predictors of pathological complete response to neoadjuvant chemotherapy in Iranian breast cancer

patients. *Asian Pacific Journal of Cancer Prevention*. 19, 2423–2427 (2018). https://doi.org/10.22034/APJCP.2018.19.9.2423

35. Velasco M., Martínez S., Cerda P., Estival A., Fernández M. y Lianes P.: Quimioterapia neoadyuvante en el cáncer de mama localmente avanzado. *Elsevier Rev. Senología y Patología Mamaria*. 25(1), 14-21 (2012). https://doi.org/10.1016/S0214-1582(12)70004-X

36. Granzier, R.W.Y., van Nijnatten, T.J.A., Woodruff, H.C., Smidt, M.L., Lobbes, M.B.I.: Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. *European Journal of Radiology*. 121 (2019). https://doi.org/10.1016/j.ejrad.2019.108736

37. Pesapane, F., Agazzi, G.M., Rotili, A., Ferrari, F., Cardillo, A., Penco, S., Dominelli, V., D'Ecclesiis, O., Vignati, S., Raimondi, S., Bozzini, A., Pizzamiglio, M., Petralia, G., Nicosia, L., Cassano, E.: Prediction of the Pathological Response to Neoadjuvant Chemotherapy in Breast Cancer Patients With MRI-Radiomics: A Systematic Review and Meta-analysis. *Current Problems in Cancer*. 46 (5) (2022). https://doi.org/10.1016/j.currproblcancer.2022.100883

38. Li, W., Newitt, D.C., Gibbs, J., Wilmes, L.J., Jones, E.F., Arasu, V.A., Strand, F., Onishi, N., Nguyen, A.A.T., Kornak, J., Joe, B.N., Price, E.R., Ojeda-Fournier, H., Eghtedari, M., Zamora, K.W., Woodard, S.A., Umphrey, H., Bernreuter, W., Nelson, M., Church, A.L., Bolan, P., Kuritza, T., Ward, K., Morley, K., Wolverton, D., Fountain, K., Lopez-Paniagua, D., Hardesty, L., Brandt, K., McDonald, E.S., Rosen, M., Kontos, D., Abe, H., Sheth, D., Crane, E.P., Dillis, C., Sheth, P., Hovanessian-Larsen, L., Bang, D.H., Porter, B., Oh, K.Y., Jafarian, N., Tudorica, A., Niell, B.L., Drukteinis, J., Newell, M.S., Cohen, M.A., Giurescu, M., Berman, E., Lehman, C., Partridge, S.C., Fitzpatrick, K.A., Borders, M.H., Yang, W.T., Dogan, B., Goudreau, S., Chenevert, T., Yau, C., DeMichele, A., Berry, D., Esserman, L.J., Hylton, N.M.: Predicting breast cancer response to neoadjuvant treatment using multi-feature MRI: results from the I-SPY 2 TRIAL. *NPJ Breast Cancer*. 6, (2020). https://doi.org/10.1038/s41523-020-00203-7

39. Khan, N., Adam, R., Huang, P., Maldjian, T., Duong, T.Q.: Deep Learning Prediction of Pathologic Complete Response in Breast Cancer Using MRI and Other Clinical Data: A Systematic Review. Tomography. 2784-2795 (2022). https://doi.org/10.3390/tomography8060232

40. Bouckaert, R.R., Frank, E.: Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Dai, H., Srikant, R., Zhang, C. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2004. *Lecture Notes in Computer Science 3056, Springer*. https://doi.org/10.1007/978-3-540-24775-3_3

41. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 5 (4), 221-232 (2016). https://doi.org/10.1007/s13748-016-0094-0

42. Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16, 321–357 (2002). https://doi.org/10.1613/jair.953

43. Alwosheel, A., van Cranenburgh, S., Chorus, C.G.: Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling. 28, 167–182 (2018). https://doi.org/10.1016/j.jocm.2018.07.002

APÉNDICES

Apéndice A

A.1 Resultados análisis estadístico de *features* pre-seleccionados

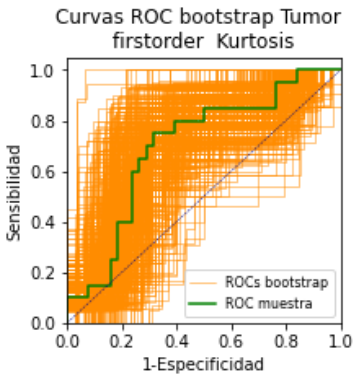
FEATURE / ATRIBUTO	*p-value* W-M-W	$AUC^{M}_{directo}$	$AUC^{B}_{directo}$	(σ)	[I.C.]
TUMOR					
firstorder Kurtosis	0,007	0,70	0,70	(0,09)	[0,55; 0,84]
glcm ClusterShade	0,052	0,63	0,64	(0,09)	[0,49; 0,79]
shape Sphericity	0,079	0,61	0,62	(0,09)	[0,47; 0,76]
PARÉNQUIMA					
firstorder 10Percentile	0,001	0,75	0,75	(0,09)	[0,61; 0,89]
glszm GrayLevelNonUniformity	0,001	0,74	0,75	(0,08)	[0,62; 0,87]
firstorder Mean	0,010	0,69	0,7	(0,08)	[0,56; 0,83]
glrlm GrayLevelNonUniformity	0,011	0,69	0,69	(0,09)	[0,53; 0,84]
firstorder Median	0,011	0,69	0,68	(0,09)	[0,54; 0,83]
gldm GrayLevelNonUniformity	0,015	0,68	0,68	(0,09)	[0,53; 0,82]
glszm SizeZoneNonUniformity	0,016	0,68	0,68	(0,09)	[0,53; 0,83]
gldm DependenceVariance	0,017	0,67	0,68	(0,09)	[0,54; 0,82]
firstorder 90Percentile	0,022	0,67	0,66	(0,08)	[0,53; 0,80]
firstorder RootMeanSquared	0,023	0,66	0,66	(0,09)	[0,52; 0,81]
shape SurfaceArea	0,025	0,66	0,67	(0,09)	[0,53; 0,81]
glcm Idmn	0,028	0,66	0,66	(0,09)	[0,50; 0,81]
glrlm LongRunHighGrayLevelEmphasis	0,033	0,65	0,65	(0,10)	[0,50; 0,81]
glrlm RunLengthNonUniformity	0,036	0,65	0,65	(0,09)	[0,50; 0,80]
shape VoxelVolume	0,037	0,65	0,65	(0,09)	[0,49; 0,80]
shape MeshVolume	0,038	0,65	0,65	(0,09)	[0,50; 0,80]
gldm DependenceNonUniformity	0,038	0,65	0,65	(0,09)	[0,50; 0,80]
firstorder Kurtosis	0,040	0,64	0,65	(0,10)	[0,48; 0,81]
ngtdm Contrast	0,044	0,64	0,64	(0,09)	[0,49; 0,78]
glrlm RunVariance	0,044	0,64	0,64	(0,10)	[0,48; 0,80]
glrlm RunPercentage	0,044	0,64	0,64	(0,09)	[0,48; 0,79]
gldm LargeDependenceEmphasis	0,051	0,63	0,64	(0,09)	[0,48; 0,79]

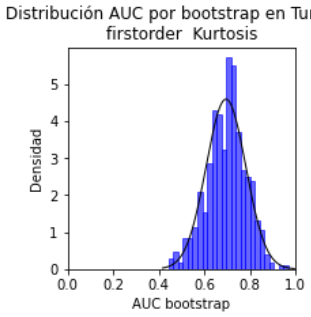
ngtdm Coarseness	0,051	0,63	0,63	(0,09)	[0,49; 0,78]
firstorder Minimum	0,052	0,63	0,63	(0,09)	[0,49; 0,78]
glcm Idn	0,053	0,63	0,63	(0,09)	[0,47; 0,78]
glrlm LongRunEmphasis	0,055	0,63	0,63	(0,09)	[0,47; 0,78]
glszm LargeAreaHighGrayLevelEmphasis	0,063	0,63	0,62	(0,09)	[0,47; 0,77]
ngtdm Strength	0,065	0,62	0,62	(0,08)	[0,48; 0,76]
shape LeastAxisLength	0,069	0,62	0,63	(0,09)	[0,48; 0,78]
glszm GrayLevelNonUniformityNormalized	0,069	0,62	0,62	(0,09)	[0,47; 0,76]
shape Maximum2DDiameterSlice	0,074	0,62	0,62	(0,09)	[0,46; 0,77]
glrlm RunLengthNonUniformityNormalized	0,079	0,62	0,61	(0,09)	[0,47; 0,76]
ngtdm Busyness	0,079	0,62	0,62	(0,08)	[0,49; 0,75]
glszm LargeAreaEmphasis	0,081	0,61	0,62	(0,09)	[0,47; 0,77]
gldm DependenceNonUniformityNormalized	0,081	0,62	0,61	(0,09)	[0,47; 0,76]
glszm ZoneVariance	0,081	0,61	0,62	(0,09)	[0,47; 0,76]
glrlm ShortRunEmphasis	0,084	0,61	0,61	(0,09)	[0,47; 0,76]
firstorder Skewness	0,098	0,61	0,6	(0,10)	[0,44; 0,76]
CLÍNICO					
Edad	0,099	0,61	0,61	(0,08)	[0.46, 0.74]
HR	0,025	0,67	-	-	-

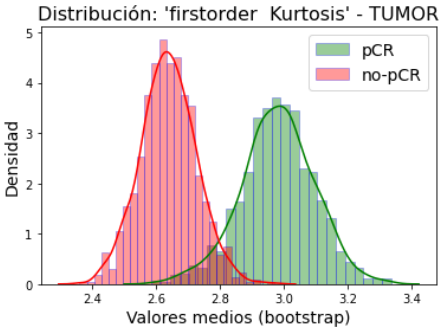
Tabla A.1: *Features* que presentan significancia estadística entre pacientes pCR y no-pCR como resultado del test de hipótesis Wilcoxon-Mann-Whitney (M-W-M), siendo seleccionados como candidatos para componer los modelos de predicción de pCR basado en ML.

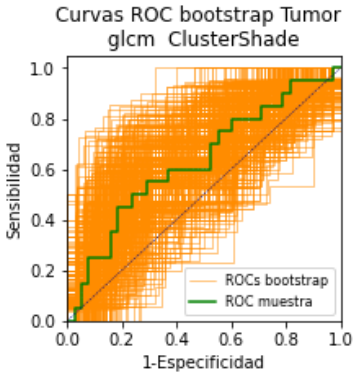
Apéndice B

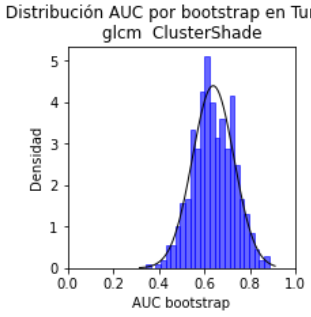
B.1 Análisis predictivo de pCR por modelo estadístico univariado

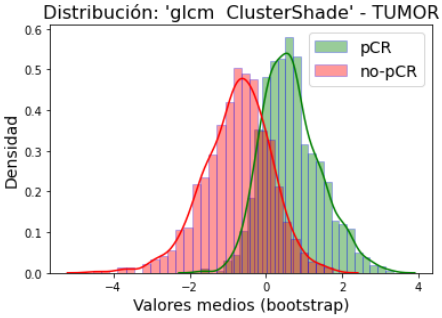


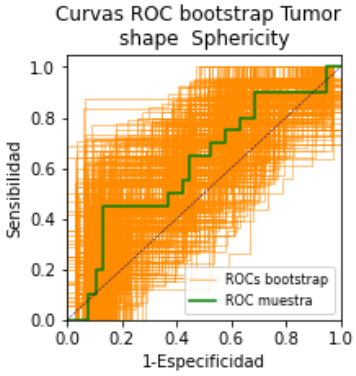


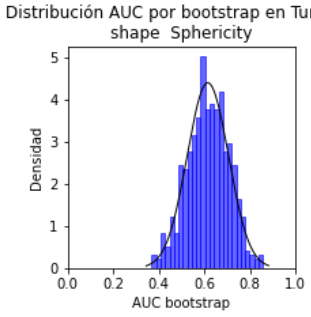


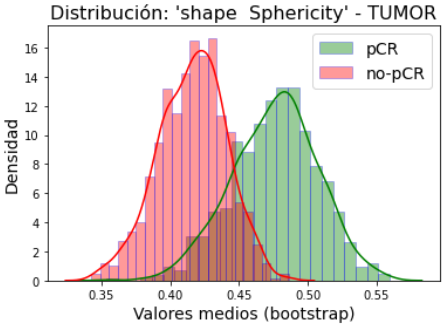


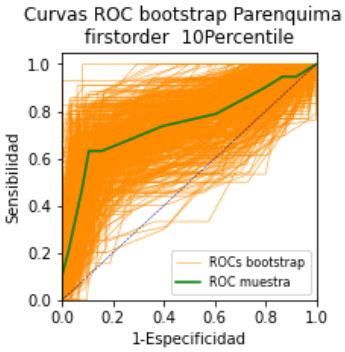


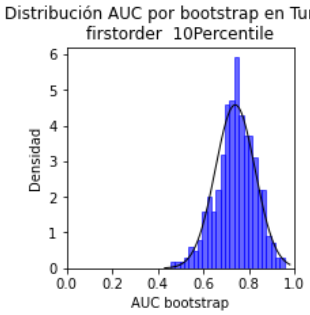


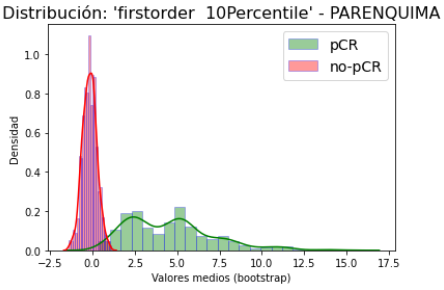


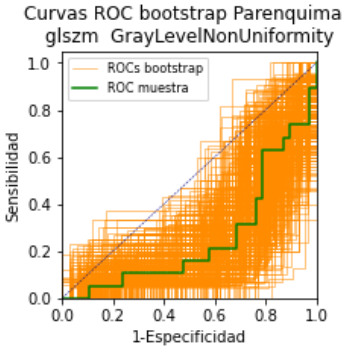


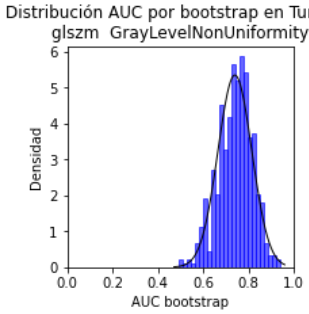


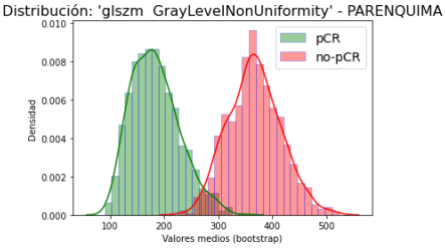

Curvas ROC bootstrap Parenquima
firstorder 10Percentile
Sensibilidad
1-Especificidad
ROCs bootstrap
ROC muestra

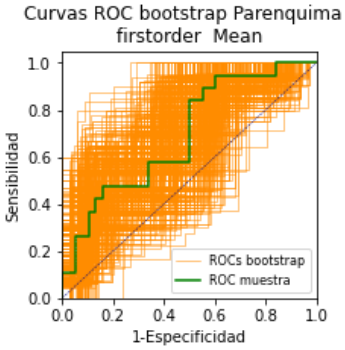

Distribución AUC por bootstrap en Tumor:
firstorder 10Percentile
Densidad
AUC bootstrap

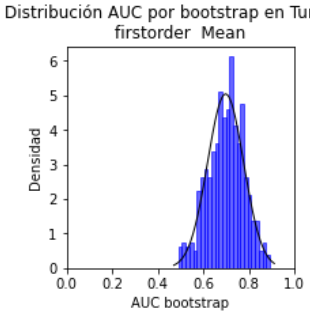

Distribución: 'firstorder 10Percentile' - PARENQUIMA
pCR
no-pCR
Densidad
Valores medios (bootstrap)

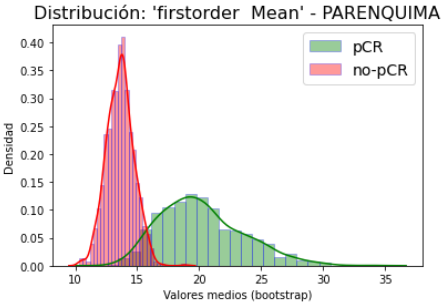

Curvas ROC bootstrap Parenquima
glszm GrayLevelNonUniformity
Sensibilidad
1-Especificidad
ROCs bootstrap
ROC muestra

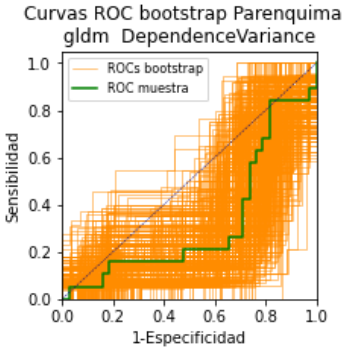

Distribución AUC por bootstrap en Tumor:
glszm GrayLevelNonUniformity
Densidad
AUC bootstrap

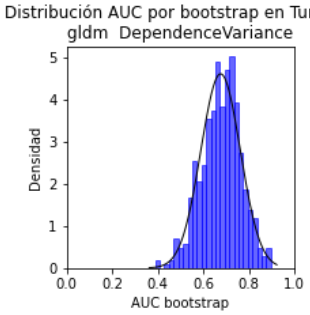

Distribución: 'glszm GrayLevelNonUniformity' - PARENQUIMA
pCR
no-pCR
Densidad
Valores medios (bootstrap)

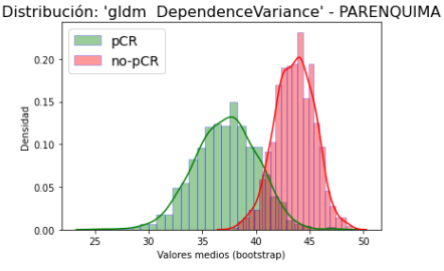

Curvas ROC bootstrap Parenquima
firstorder Mean
Sensibilidad
1-Especificidad
ROCs bootstrap
ROC muestra


Distribución AUC por bootstrap en Tumor:
firstorder Mean
Densidad
AUC bootstrap


Distribución: 'firstorder Mean' - PARENQUIMA
pCR
no-pCR
Densidad
Valores medios (bootstrap)


Curvas ROC bootstrap Parenquima
gldm DependenceVariance
Sensibilidad
1-Especificidad
ROCs bootstrap
ROC muestra


Distribución AUC por bootstrap en Tumor:
gldm DependenceVariance
Densidad
AUC bootstrap


Distribución: 'gldm DependenceVariance' - PARENQUIMA
pCR
no-pCR
Densidad
Valores medios (bootstrap)

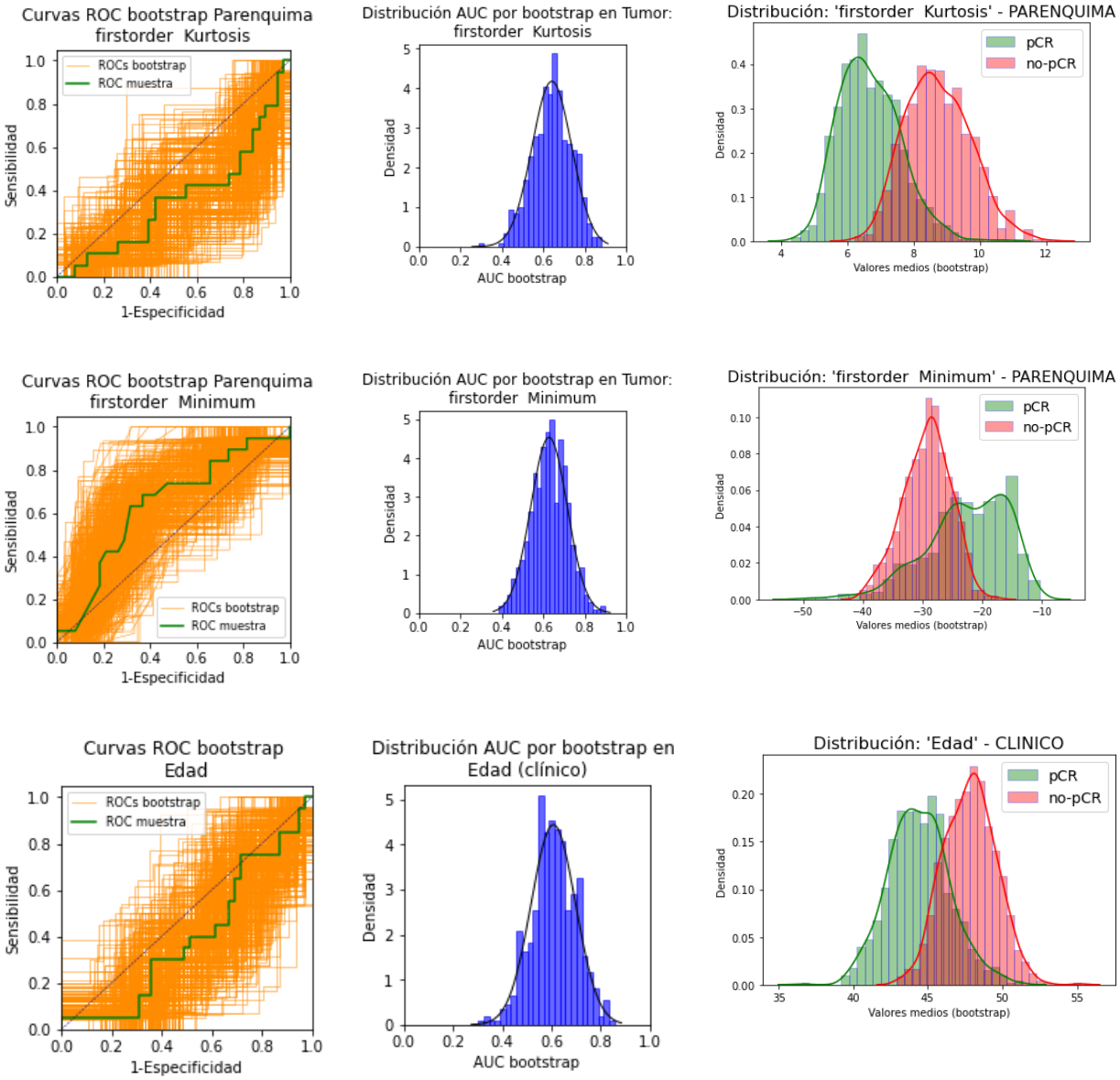


Figura B.1: Para cada *feature* seleccionado al finalizar el análisis estadístico, se observa: a.) curvas ROC de predicción estadística directa de pCR; la curva verde corresponde a la muestra original, y las naranjas, a las submuestras generadas por *Bootstrap*; b.) distribución de valores métrica AUC obtenidas de las curvas ROC en a; c) distribución de la media de los valores del *feature* de las submuestras por *bootstrap* para ambas clases.

B.2 Correlación de variables pre-seleccionadas

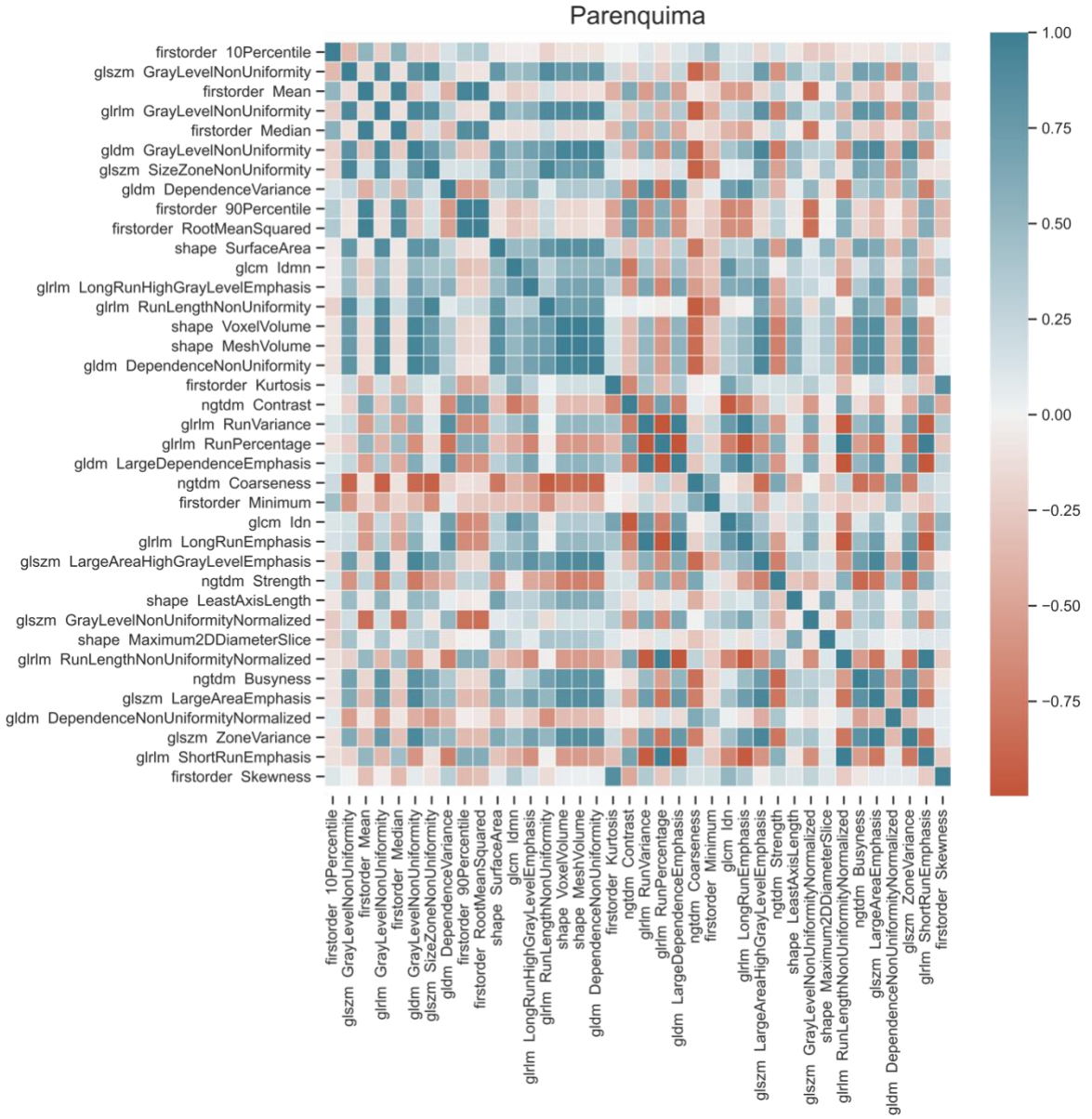


Figura B.2: Matriz de correlación de los *features* pre-seleccionados del parénquima mamario.