




4 DE SEPTIEMBRE DE 2023

Desarrollo y Validación de Modelos
Predictivos de Toxicidad Asociada al
Tratamiento de Pacientes con Cáncer
de Pulmón Basados en Técnicas de
Aprendizaje Automático

PROGRAMA DE DOCTORADO EN INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE SEVILLA

TUTOR: D. JOSÉ CRISTÓBAL RIQUELME SANTOS
DIRECTORES: D. JOSÉ LUIS LÓPEZ GUERRA Y D. ALBERTO MORENO CONDE
Candidato: D. Francisco José Núñez Benjumea



If I have seen further, it is by standing on the shoulders of Giants.

Sir Isaac Newton (1642-1727)

Tabla de Contenidos

Índice de figuras	4
Índice de tablas	5
Acrónimos.....	6
1. Introducción	8
1.1. Toxicidad inducida por radiación en pacientes con cáncer de pulmón	9
1.2. Variables clínicas de predicción de toxicidad inducida por radiación	11
1.3. Revisión de métodos basados en aprendizaje automático para la predicción de toxicidad inducida por radiación.....	15
1.4. Hipótesis y objetivos	24
1.5. Trabajos realizados en el marco de esta investigación	26
2. Metodología para el desarrollo y validación de modelos de predicción de toxicidad asociada al tratamiento de RT	31
2.1. El conjunto de datos.....	32
2.2. Definición de la pregunta clínica	34
2.3. Definición de las variables clínicas objetivo.....	35
2.4. Normalización del conjunto de datos	36
2.5. Curación del conjunto de datos.....	37
2.6. Imputación de valores perdidos	38
2.7. Métodos de selección de características.....	40
2.8. Entrenamiento de los modelos de predicción basados en técnicas de aprendizaje automático	43
2.8.1. Máquina de vector soporte (SVM)	45
2.8.2. k-ésimo vecino más cercano (kNN)	46
2.8.3. Red neuronal artificial (ANN)	47
2.8.4. Modelo Lineal Generalizado (GLM)	48
2.8.5. Näive Bayes (NB).....	49

2.9.	Validación interna	50
2.10.	Validación externa y medidas de desempeño.....	50
3.	Resultados	52
3.1.	AUC y curvas de calibración obtenidas para cada toxicidad	52
3.2.	AUC según método de selección de características.....	54
3.3.	AUC según técnica de aprendizaje automático.....	56
4.	Discusión.....	57
4.1.	AUC y curvas de calibración obtenidas para cada toxicidad	57
4.1.1.	Esofagitis aguda	57
4.1.2.	Tos aguda.....	58
4.1.3.	Disnea aguda	59
4.1.4.	Neumonitis aguda	60
4.1.5.	Disnea crónica	60
4.1.6.	Neumonitis crónica	61
4.1.7.	Curvas de calibración	61
4.2.	AUC según método de selección de características.....	62
4.3.	AUC según método de predicción basado en técnicas de aprendizaje automático	63
4.4.	Cumplimiento de los objetivos planteados	64
4.5.	Limitaciones	66
5.	Conclusiones	70
	Agradecimientos.....	71
	Bibliografía	72
	Anexo A. AUCs obtenidos tras la validación interna	87
	Anexo B. AUCs obtenidos tras la validación externa.....	89

Índice de figuras

FIGURA 1. NOMOGRAMA PARA LA PREDICCIÓN DE SUPERVIVENCIA EN PACIENTES CON CP. FUENTE: ZHENG ET AL., 2019 (18).....	10
FIGURA 2. RESUMEN DE TOXICIDADES INDUCIDAS POR RT REPORTADAS EN LA LITERATURA TRAS TRATAMIENTO DE RT EN ZONA TORÁCICA.	14
FIGURA 3. RELACIÓN RENDIMIENTO-INTERPRETABILIDAD EN MODELOS DE APRENDIZAJE AUTOMÁTICO. ADAPTADO DE YANG ET AL., 2022 (60).....	17
FIGURA 4. FLUJO DE TRABAJO TÍPICO PARA EL DESARROLLO DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE TOXICIDAD INDUCIDA POR RT. FUENTE: ISAKSSON ET AL., 2020 (62).	18
FIGURA 5. METODOLOGÍA PROPUESTA PARA LA IMPLEMENTACIÓN DE MODELOS DE PREDICCIÓN DE TOXICIDAD ASOCIADA AL TRATAMIENTO DE RT BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO DESARROLLADOS A PARTIR DE DATOS OBTENIDOS EN EL ENTORNO ASISTENCIAL.....	31
FIGURA 6. EJEMPLO DE SVM CON CUATRO VECTORES SOPORTE EN EL ESPACIO DE CARACTERÍSTICAS. FUENTE: PARK ET AL. (109).....	45
FIGURA 7. ILUSTRACIÓN DEL ALGORITMO KNN. LA NUEVA MUESTRA ES EL PUNTO VERDE. PARA $k=3$ VECINOS, SE CONSIDERA QUE LA NUEVA MUESTRA PERTENECE A LA CLASE TRIANGULAR, MIENTRAS QUE PARA $k=5$ VECINOS, EL ALGORITMO LA CLASIFICA EN LA CLASE CUADRADA. FUENTE: WIKIPEDIA (112).....	46
FIGURA 8. REPRESENTACIÓN DE UNA RED NEURONAL ARTIFICIAL RECURRENTE CON RETROALIMENTACIÓN HACIA ADELANTE. FUENTE: TOBORE ET AL. (115).	47
FIGURA 9. FUNCIONES DE DENSIDAD DE PROBABILIDAD DE CUATRO CASOS DE USO DE UN CLASIFICADOR NB UTILIZADO PARA DETERMINAR LA ESTABILIDAD DE LOS TÚNELES. FUENTE: LI ET AL. (119).	49
FIGURA 10. CURVAS DE CALIBRACIÓN CORRESPONDIENTES A LOS MEJORES MODELOS DE PREDICCIÓN OBTENIDOS PARA CADA TOXICIDAD.....	54
FIGURA 11. DIAGRAMAS DE CAJA (BOXPLOT) DE LAS AUCs OBTENIDAS POR CADA MÉTODO DE SELECCIÓN DE CARACTERÍSTICAS EN LAS ETAPAS DE VALIDACIÓN INTERNA (FONDO BLANCO) Y VALIDACIÓN EXTERNA (FONDO GRIS). DE IZQUIERDA A DERECHA, LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS SON: BORUTA, CFS, CHI CUADRADO, GOLD STANDARD, IG, MRMR, VOTACIÓN (CLASIFICACIÓN), VOTACIÓN (SUBCONJUNTOS), RELIEF Y RF.	55
FIGURA 12. DIAGRAMAS DE CAJA (BOXPLOT) DE LAS AUCs OBTENIDAS POR CADA ALGORITMO DE CLASIFICACIÓN EN LAS ETAPAS DE VALIDACIÓN INTERNA (FONDO BLANCO) Y VALIDACIÓN EXTERNA (FONDO GRIS). DE IZQUIERDA A DERECHA: ANN, GLM, KNN, NB Y SVM.	56

Índice de tablas

TABLA 1. EJEMPLOS DE MODELOS DE PREDICCIÓN DE TOXICIDAD INDUCIDA POR RT BASADOS EN MODELOS DE APRENDIZAJE AUTOMÁTICO.	21
TABLA 2. SUBCONJUNTOS DE DATOS RESULTANTES TRAS EL PREPROCESAMIENTO. ENTRE PARÉNTESIS, NÚMERO DE CASOS CON TOXICIDAD POSITIVA EN CADA SUBCONJUNTO.....	39
TABLA 3. CARACTERÍSTICAS DE LOS ALGORITMOS CLASIFICADORES EMPLEADOS.....	45
TABLA 4. RESUMEN DE LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS IMPLEMENTADOS A PARTIR DE LIBRERÍAS DE R.	42
TABLA 5. MEJORES MODELOS SEGÚN AUC DURANTE VALIDACIÓN INTERNA PARA CADA TOXICIDAD BAJO ESTUDIO.	53
TABLA 6. AUC PROMEDIO Y DESVIACIÓN ESTÁNDAR (ENTRE PARÉNTESIS) DURANTE LAS FASES DE VALIDACIÓN INTERNA Y VALIDACIÓN EXTERNA PARA CADA UNO DE LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS ESTUDIADOS. SE INCLUYE COMPARACIÓN CON EL GOLD STANDARD.....	55
TABLA 7. AUC PROMEDIO Y DESVIACIÓN ESTÁNDAR (ENTRE PARÉNTESIS) OBTENIDOS DURANTE LAS FASES DE VALIDACIÓN INTERNA Y VALIDACIÓN EXTERNA PARA CADA UNO DE LOS ALGORITMOS DE CLASIFICACIÓN EMPLEADOS.....	56

Acrónimos

ALK	Anaplastic lymphoma kinase
ANN	Red neuronal artificial
AUC	Área bajo la curva característica operativa del receptor
CFS	Selección de características basada en la correlación
CP	Cáncer de pulmón
CPNM	Cáncer de pulmón no microcítico
CTCAE	Common terminology criteria for adverse events
CTV	Clinical target volume
CVD	Complejo vagal dorsal
CVRS	Calidad de vida relacionada con la salud
DLCO	Capacidad de difusión del monóxido de carbono
D_{MAX}	Dosis máxima de radiación recibida
D_{MEAN}	Dosis media de radiación recibida
D_{MEDIAN}	Dosis mediana de radiación recibida
EFMI	European Federation of Medical Informatics
EGFR	Epidermal growth factor receptor
EPOC	Enfermedad pulmonar obstructiva crónica
ESTRO	European Society for Radiotherapy and Oncology
FENO	Fracción exhalada de óxido nítrico
FEV1	Volumen espiratorio forzado en un segundo
GLM	Modelo lineal generalizado
GTV	Gross tumor volume
HIS	Sistemas de información hospitalarios
HUVR	Hospital universitario Virgen del Rocío
IG	Ganancia de información
IMC	Índice de masa corporal
ISCI	Instituto de Salud Carlos III
kNN	k-nearest neighbor (k-ésimo vecino más cercano)
KPS	Karnofsky performance status
LOO	Leave one out (dejar uno fuera, técnica de validación interna)
MCC	Coeficiente de correlación de Matthew
mRMR	Mínima redundancia-máxima relevancia
NB	Näive Bayes
NCCN	National comprehensive cancer network
NICE	National institute for health and care excellence
NIH	National institutes of health
NTCP	Probabilidad de complicación de tejido normal
PET	Tomografía por emisión de protones
PFT	Pruebas funcionales pulmonares
PROM	Patient reported outcome measurement

QoL	Calidad de vida
QT	Quimioterapia
RF	Random forest
RT	Radioterapia
RWHD	Real-world health data
SNP	Polimorfismo de nucleótido simple
SVM	Máquina de vector soporte
TNM	Sistema de estadificación del cáncer
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
VIH	Virus de inmunodeficiencia humana

1. Introducción

Según el informe GLOBOCAN 2020 (1), el cáncer de pulmón (CP) es el segundo cáncer más diagnosticado en el mundo, después del cáncer de mama femenino, y representa el 11.4% de los nuevos casos de cáncer diagnosticados, pero sigue siendo la principal causa de morbilidad y mortalidad por cáncer en el mundo (18% del total de muertes por cáncer). Las tasas de incidencia y mortalidad son aproximadamente 2 veces más altas en los hombres que en las mujeres, y de 3 a 4 veces más altas en los países desarrollados que en los países en vías de desarrollo (1). La supervivencia del CP localmente avanzado a los 5 años del diagnóstico oscila entre el 10% y el 50% (2), muy por detrás de otros tipos de cáncer con tasas de incidencia comparables, como el cáncer de mama femenino (85-90%) o el cáncer de próstata (80-95%) (3), a pesar de la evidencia sobre la eficacia de la implantación de programas de cribado de CP basados en la tomografía computarizada de baja dosis para los individuos de alto riesgo (1).

El Instituto Nacional del Cáncer, perteneciente al Instituto Nacional de Salud (NIH) de Estados Unidos, define la radioterapia (RT) como el uso de radiación de alta energía generada mediante rayos X, rayos gamma, neutrones, protones y otras fuentes para eliminar las células cancerosas y reducir el tamaño y la actividad de los tumores (4). Utilizada como opción terapéutica primaria, secundaria, concomitante o adyuvante, se prescribe habitualmente como parte del tratamiento activo (5) según las guías clínicas más extendidas y consensuadas, como son las elaboradas por la National Comprehensive Cancer Network (NCCN) (6,7) de la Sociedad Americana de Oncología y las elaboradas por el National Institute for Health and Care Excellence (NICE) (8) del Ministerio de Sanidad Británico para el manejo de pacientes con CP.

1.1. Toxicidad inducida por radiación en pacientes con cáncer de pulmón

A pesar de los avances realizados en torno a la mejora de la precisión y eficacia de esta terapia (9), los órganos y tejidos que rodean al tumor pueden verse afectados por el haz de radiación de alta energía, generando así una respuesta inflamatoria en los volúmenes afectados que puede desencadenar la aparición de una serie de eventos adversos (es decir, lo que se conoce como toxicidad inducida por RT). Estos acontecimientos parecen depender de múltiples factores, como la dosis de radiación administrada, la localización de la irradiación, el perfil celular y molecular (10,11) y los hábitos de vida de los pacientes (12,13).

La gestión de las posibilidades de aparición de estas toxicidades está reconocida como uno de los factores más limitantes durante la planificación de la RT (14,15), impidiendo que se pueda administrar una mayor dosis de radiación para facilitar una mejor respuesta tumoral, lo que se traduciría en una potencial mejora de la supervivencia global del paciente. Además, las toxicidades inducidas por la RT pueden convertirse en afecciones crónicas en función de su evolución y persistencia a lo largo del tiempo, lo que perjudicaría la calidad de vida relacionada con la salud (CVRS) del paciente a largo plazo. Por lo tanto, tanto los pacientes como los oncólogos radioterápicos a menudo tienen que mantener un equilibrio entre la CVRS y la supervivencia en un contexto de toma de decisiones compartida, siendo la CVRS en el momento del diagnóstico y las expectativas de supervivencia los principales determinantes en las decisiones de los pacientes (16).

En este contexto, disponer de una predicción personalizada y precisa de la posible aparición de toxicidades inducidas por la RT y de las expectativas de supervivencia sería de gran ayuda para que los pacientes y los oncólogos radioterápicos comprendieran mejor las implicaciones de sus decisiones sobre las opciones de tratamiento disponibles. De las directrices de la NCCN se puede extraer un conjunto básico de variables clínicas (por ejemplo: edad, sexo, estadificación del tumor, número de ganglios afectados y decisiones terapéuticas, entre otras) para realizar una estimación cuantitativa de la supervivencia a 1, 2 y 5 años de los pacientes con CP. Además, en Internet hay disponibles varias "calculadoras de supervivencia"

validadas para uso clínico, también conocidas como nomogramas, basadas principalmente en estas variables (17,18).

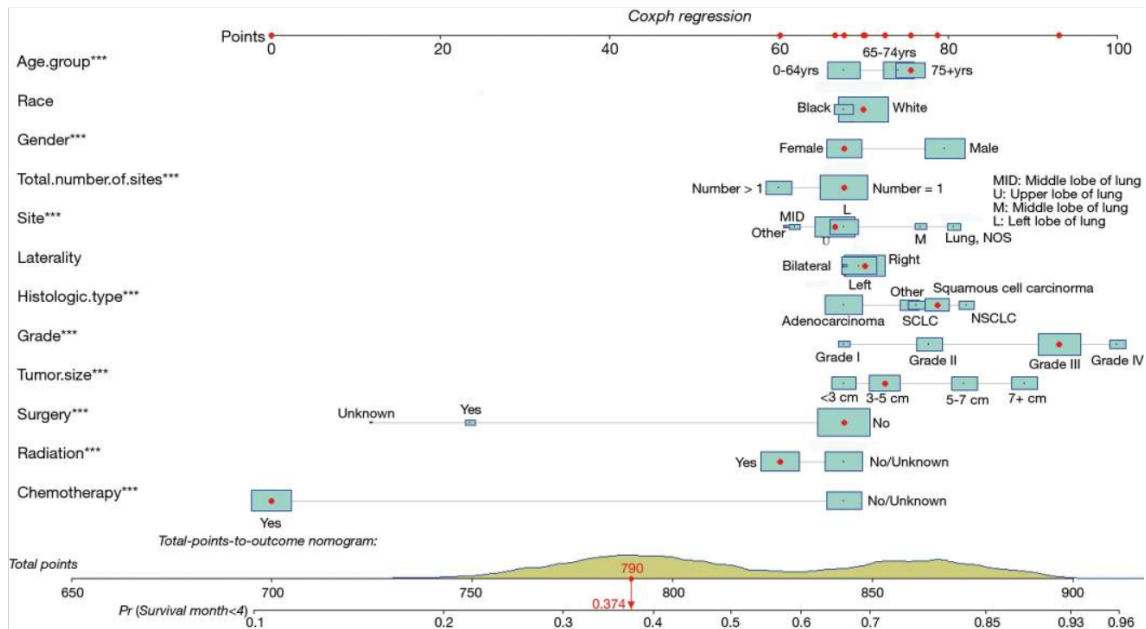


FIGURA 1. NOMOGRAMA PARA LA PREDICCIÓN DE SUPERVIVENCIA EN PACIENTES CON CP. FUENTE: ZHENG ET AL., 2019 (19).

Sin embargo, en la literatura existen experiencias basadas en métodos de aprendizaje automático y de minería de datos (DM) en las que se obtienen mejoras en la precisión de la predicción de la supervivencia añadiendo otras variables relevantes que no se tienen en cuenta en las guías clínicas (20,21), proporcionando un pronóstico más personalizado y preciso sobre los posibles resultados del tratamiento oncológico. Este pronóstico también sería de utilidad para informar el proceso de toma de decisiones compartidas entre el paciente y el oncólogo radioterápico. Este pronóstico sobre la supervivencia podría complementarse con un método similar para extraer predicciones personalizadas y precisas sobre la posible aparición de toxicidades inducidas por la RT y su impacto en la CVRS de los pacientes, con el fin de proporcionar a los pacientes y a los oncólogos radioterápicos una comprensión completa de las implicaciones para el equilibrio vital del paciente relacionadas con las opciones de tratamiento disponibles.

1.2. Variables clínicas de predicción de toxicidad inducida por radiación

En la literatura se han descrito ampliamente algunos predictores clínicos de la toxicidad inducida por la RT. Los seres humanos responden a la RT de manera individualmente variable (22), lo que penaliza la precisión de los modelos de predicción generalizados cuando se aplican a casos específicos. A continuación, se ofrece una revisión ilustrativa de los predictores más comúnmente reportados en los sistemas anatómicos comúnmente afectados en pacientes con CP sometidos a RT, con la idea de visibilizar la heterogeneidad de las causas subyacentes para la aparición de las toxicidades asociadas al tratamiento radioterápico.

En relación con las vías respiratorias, una de las toxicidades más estudiadas es la aparición de neumonitis por radiación. Esta toxicidad se produce en el 5-20% de los pacientes sometidos a RT y limita la dosis máxima de radiación que se puede administrar, reduciendo por tanto la probabilidad de control tumoral. Esta toxicidad puede provocar disnea, fibrosis pulmonar y deterioro de la calidad de vida (23), entre otras comorbilidades. Los predictores clínicos reportados con mayor frecuencia en los casos de neumonitis inducida por RT son, entre otros (24–27), volumen espiratorio forzado en un segundo (FEV1), dosis media de radiación recibida (D_{mean}) en la zona pulmonar, volúmenes relativos de la dosis pulmonar, capacidad de difusión del monóxido de carbono (DLCO), fracción exhalada de óxido nítrico (FeNO), probabilidad de complicación del tejido normal (NTCP), aparición de esofagitis aguda inducida por la radiación, y la existencia de enfermedad pulmonar obstructiva crónica (EPOC) de forma basal.

Otra toxicidad respiratoria comúnmente reportada después de la aplicación de la RT es la disnea inducida por radiación, que ocurre en aproximadamente el 10-20% de los pacientes con CP (28). Los predictores clínicos de esta toxicidad encontrados en la literatura revisada (28–30) son: comorbilidad cardíaca basal, FEV1 basal, Estado funcional del paciente (*performance status*), historial de tabaquismo, edad del paciente, D_{mean} en la zona pulmonar, disnea basal, enfermedad psiquiátrica basal, y volúmenes relativos de la dosis pulmonar, entre otros.

En cuanto a la aparición de dolor en la pared torácica y/o fractura de costillas inducidos por la RT, estudios recientes (31,32) han destacado la capacidad predictiva de los

volúmenes relativos de dosis pulmonar y la localización del tumor, aunque todavía es necesario investigar más sobre esta toxicidad específica.

En el caso de las toxicidades inducidas por la RT relacionadas con el tracto digestivo, se han documentado en la bibliografía factores clínicos predictivos relacionados con la aparición de esofagitis, disfagia y náuseas, que se exponen a continuación.

Los síntomas de esofagitis aguda incluyen disfagia, náuseas, anorexia, dolor al tragar (odinofagia) y ardor de estómago (pirosis). Si estos síntomas empeoran, pueden provocar deshidratación, desnutrición y pérdida de peso (33). La esofagitis se puede considerar crónica cuando su duración es superior a 3 meses. En promedio, los casos reportados tienen una duración de pocos meses, llegando en ocasiones a perdurar durante 1 año o más tiempo.

El debut de la esofagitis inducida por la RT se ha correlacionado en la bibliografía (34–37) con las siguientes variables: dosis máxima (D_{max}) recibida por el esófago, D_{mean} recibida por el esófago, dosis mediana (D_{median}) recibida por el esófago, volúmenes relativos de la dosis del esófago, existencia de tratamiento de quimioterapia concomitante, y género del paciente, entre otros.

Respecto a la disfagia, comúnmente está asociada con la desnutrición, la deshidratación, la capacidad de aspiración y la presencia de reflujo gastroesofágico, y se ha observado que su desarrollo afecta negativamente al bienestar percibido por los pacientes en las mediciones de la CVRS (38,39). Su aparición y gravedad se han correlacionado con el índice de masa corporal (IMC), el uso de una sonda de alimentación parenteral, el estadio del tumor y el régimen de tratamiento, entre otros factores (39,40).

Las náuseas y los vómitos también se han descrito como una toxicidad asociada a la RT frecuente, aunque menos documentada, que empeora la CVRS de los pacientes. Las variables predictoras de esta toxicidad se han comunicado en una revisión narrativa (41), destacando las posibles correlaciones con la edad del paciente, el estadio del tumor, la existencia de tratamiento de quimioterapia concomitante y los volúmenes relativos de dosis recibidas en otras estructuras anatómicas situadas fuera del área torácica, con limitado interés para este trabajo. Sin embargo, esta revisión narrativa concluye que se necesitan llevar

a cabo más investigaciones para arrojar luz sobre cuáles son los factores predictivos asociados a la aparición de esta toxicidad inducida por RT.

Un conjunto diferente de toxicidades inducidas por la RT que han sido ampliamente reportadas son las relacionadas con el sistema cardiovascular. Las toxicidades cardíacas tras la exposición a la RT incluyen la enfermedad arterial coronaria, la enfermedad valvular, la cardiomiopatía, la enfermedad pericárdica y las arritmias (42), entre otras. Se ha informado de los siguientes factores predictivos de la cardiotoxicidad: D_{mean} cardíaca, los volúmenes relativos de la dosis cardíaca, el estilo de vida basal del paciente (niveles de actividad física, hábito tabáquico y obesidad), algunos biomarcadores, técnica de RT usada en el tratamiento y la localización del tumor (43–47). Sin embargo, parece que no hay unanimidad en la comunidad investigadora acerca de la capacidad de predicción de los biomarcadores (47) y ciertas variables dosimétricas (48), por lo que en el futuro deberán realizarse más investigaciones sobre este tema para alcanzar el consenso.

La piel es otro órgano expuesto a la toxicidad inducida por la RT en los pacientes con CP. La dermatitis por radiación puede afectar a la CVRS y comprometer la continuidad de la aplicación de tratamientos antitumorales. Esta toxicidad puede progresar desde el eritema a la descamación seca, a la descamación húmeda e incluso a la ulceración, tras lo cual pueden desarrollarse cambios cutáneos permanentes que afectan a la cosmética (49). En la literatura se han descrito varios factores predictivos de dermatitis por radiación como, por ejemplo: la localización del tumor, el IMC, el hipofraccionamiento en la administración del tratamiento, la técnica de RT usada durante el tratamiento, los volúmenes relativos de dosis recibidas en la piel, el hábito tabáquico y la administración de dosis de refuerzo (36,49–51), entre otros.

Por último, pero no por ello menos importante, la fatiga o astenia es uno de los síntomas más comunes y angustiosos en los pacientes con cáncer sometidos a tratamientos RT, y ha sido identificada como una causa importante de deterioro de la CVRS en esta población (52). La RT puede inducir el desencadenante de factores sistémicos activos, una reducción de los niveles de hemoglobina, una disminución de la eficiencia neuromuscular y un aumento de la fatiga muscular, contribuyendo así a la aparición de este síntoma (53). Algunas de las posibles variables predictoras de la fatiga inducida por la RT que se han reportado en la literatura son: factores psicológicos, higiene del sueño, IMC, biomarcadores,

existencia de tratamiento de quimioterapia concurrente, edad del paciente, diabetes concurrente y presencia de dolor (54–58). Sin embargo, los mecanismos subyacentes que conducen al desarrollo de la fatiga siguen siendo poco conocidos (53).

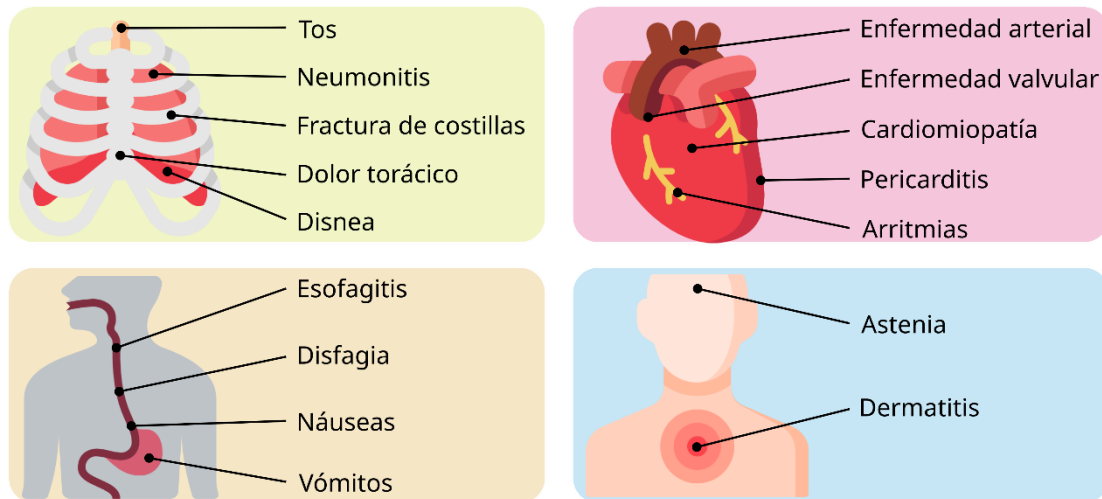


FIGURA 2. RESUMEN DE TOXICIDADES INDUCIDAS POR RT REPORTADAS EN LA LITERATURA TRAS TRATAMIENTO DE RT EN ZONA TORÁCICA.

Cabe señalar que la mayor parte de la bibliografía revisada reflejaba un alto nivel de variabilidad en los regímenes de tratamiento, las poblaciones estudiadas, la definición de las toxicidades agudas o crónicas y la precisión de los resultados comunicados, lo que lastra en gran parte la calidad de las conclusiones obtenidas.

Es en este contexto donde una metodología sólida, basada en métodos de ciencia de datos para impulsar el descubrimiento o constatación de factores predictivos de toxicidades inducidas por la RT, y que tenga la capacidad de ser personalizada a cada paciente en función de sus características físicas, biológicas, moleculares y de comportamiento, tiene el potencial de proporcionar evidencias fiables y basadas en datos obtenidos a través de las interacciones médico-paciente para, en última instancia, facilitar el proceso de toma de decisiones terapéuticas compartidas e informadas entre los pacientes y oncólogos radioterápicos.

1.3. Revisión de métodos basados en aprendizaje automático para la predicción de toxicidad inducida por radiación

Cuando se trata de la aplicación de métodos basados en aprendizaje automático para el desarrollo de modelos predictivos de toxicidad inducida por RT, el primer paso es delimitar qué variables tendría sentido emplear, desde un punto de vista teórico, como potenciales predictores del *endpoint* clínico que deseamos predecir. Lambin *et al.* (59) propuso los siguientes conjuntos de información a tener en cuenta en esta tarea basándose en el conocimiento actual de los factores que contribuyen a la aparición de toxicidades inducidas por la RT:

- a) Características clínicas: estado funcional del paciente, grado y estadio del tumor, resultados de los análisis de sangre, cuestionarios del paciente.
- b) Características del tratamiento: distribución espacial y temporal de la dosis planificada, tratamiento de quimioterapia concurrente.
- c) Características de imagen: tamaño y volumen del tumor, capacidad de captación metabólica (este concepto se incluye dentro de la disciplina conocida como radiómica)
- d) Características moleculares: radiosensibilidad intrínseca, hipoxia, proliferación celular y reacción del tejido normal.

La inclusión de otros factores relacionados con la planificación del tratamiento de RT, tanto dosimétricos como no dosimétricos, también se ha sugerido como un conjunto de variables de interés para mejorar la precisión de los modelos de predicción de toxicidad asociada a la RT (60).

Respecto a qué metodología seguir para desarrollar este tipo de modelos de predicción, Kang *et al.* (61) propuso un conjunto de principios rectores que deberían tenerse en cuenta. Estos principios se obtuvieron a partir de una revisión sistemática, y tenían como objetivo facilitar la adopción de una metodología para la implementación de modelos de aprendizaje automático en el contexto de la oncología radioterápica que estuvieran respaldados por una metodología sólida. Los principios se resumen a continuación:

- i. Considerar las variables de predicción dosimétricas y las no dosimétricas.

- ii. Curar manualmente el conjunto de datos antes de realizar cualquier análisis.
- iii. Implementar un método automático de selección de características para una primera selección de potenciales variables de predicción.
- iv. Considerar los sesgos que pudiera estar introduciendo en el modelo la potencial multicolinealidad o correlación entre las variables de predicción.
- v. Utilizar correctamente la técnica de validación cruzada para obtener un valor realista de la precisión del modelo y evaluar su capacidad de generalización.
- vi. Evaluar la capacidad de generalización del modelo desarrollado mediante su uso en conjuntos de datos externos siempre que sea posible.
- vii. Implementar múltiples modelos y comparar los resultados obtenidos.

Cabe destacar la existencia de iniciativas como la declaración TRIPOD (62) a la hora de intentar homogeneizar la forma en la que los modelos de predicción multivariables deben ser presentados a la comunidad científica. Esta declaración recoge el consenso de expertos en el desarrollo de modelos de predicción, metodólogos profesionales de la salud y editores de revistas científicas sobre las buenas prácticas que deben tenerse en cuenta a la hora de informar sobre el desarrollo y validación de modelos predictivos aplicados en el ámbito clínico-asistencial. Aun así, tal y como se ha evidenciado en una reciente revisión sistemática sobre la aplicación de esta declaración en el contexto oncológico (63), en la práctica aún dista bastante de lo que sería mínimamente deseable para poder extrapolar evidencias robustas a través de meta-análisis de resultados individuales.

También se ha destacado la necesidad de generar confianza entre los usuarios finales (es decir, los oncólogos radioterápicos) sobre el uso de modelos de predicción basados en técnicas de aprendizaje automático con la intención de facilitar su adopción en la rutina asistencial. En este sentido, la comunidad investigadora está invirtiendo grandes esfuerzos en optimizar el equilibrio entre la interpretabilidad y el rendimiento ofrecidos por los modelos predictivos basados en técnicas de aprendizaje automático dado que, en general, resulta que cuanto menos interpretable es el modelo, mejor es su rendimiento (mayor precisión en la predicción ofrece). Esto lleva como consecuencia al desarrollo de modelos predictivos llamados de “caja negra” que, a pesar de ofrecer un rendimiento superior que, en ocasiones, incluso mejora la habilidad de los expertos en la materia, no son capaces de ofrecer un razonamiento lógico lineal al por qué de las predicciones que realizan, siendo este un factor

limitante para la generación de confianza en estos modelos por parte de los oncólogos radioterápicos.

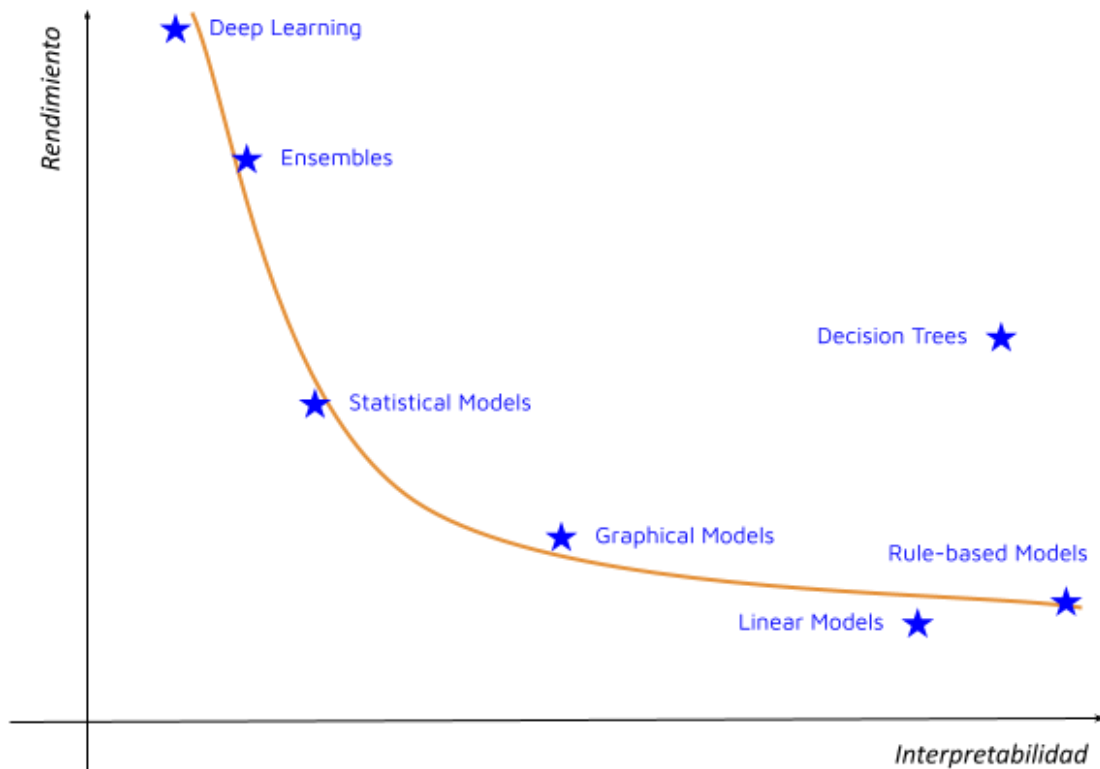


FIGURA 3. RELACIÓN RENDIMIENTO-INTERPRETABILIDAD EN MODELOS DE APRENDIZAJE AUTOMÁTICO. ADAPTADO DE YANG ET AL., 2022 (64).

Actualmente, existe una creciente comunidad investigadora enfocada en el desarrollo de la *Explainable AI* (Inteligencia artificial interpretable) en la que se están produciendo avances significativos en metodologías encaminadas a proporcionar un mayor grado de interpretabilidad a los modelos predictivos basados en aprendizaje automático a través de estrategias que aprovechan las características propias de los modelos predictivos (por ejemplo, haciendo explícitos las variables y el peso otorgado a cada una a la hora de realizar una predicción), complementándola con técnicas de análisis de sensibilidad que, tras producir una perturbación controlada a la entrada del modelo, pueden ofrecer una explicación comprensible de los cambios producidos en la predicción obtenida (65).

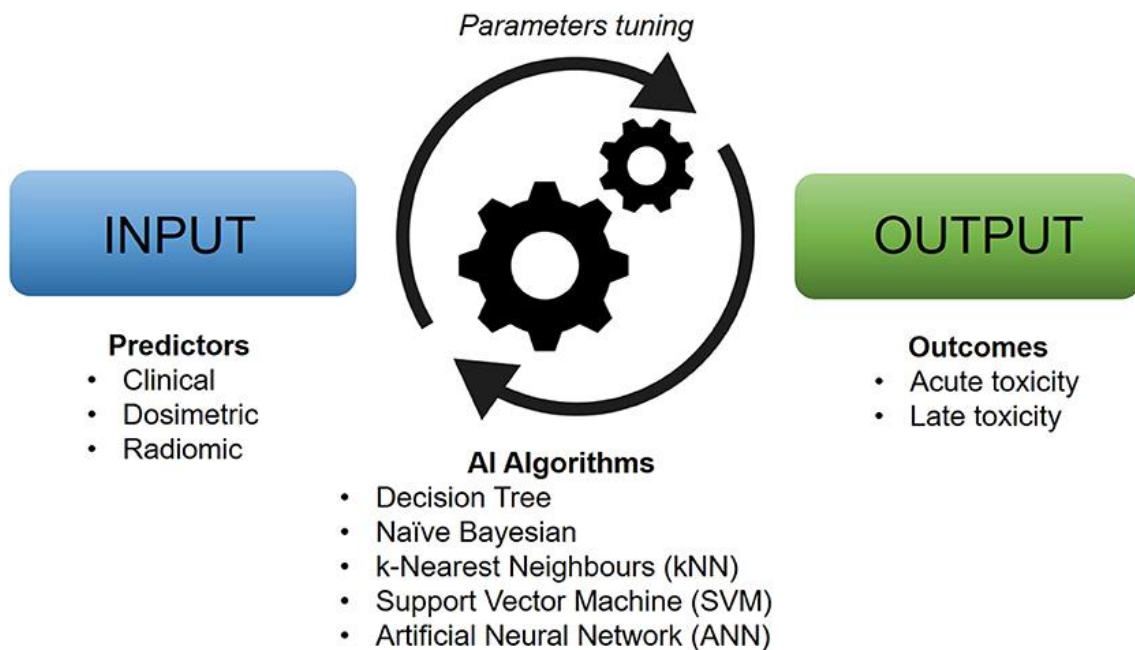


FIGURA 4. FLUJO DE TRABAJO TÍPICO PARA EL DESARROLLO DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE TOXICIDAD INDUCIDA POR RT. FUENTE: ISAKSSON *ET AL.*, 2020 (66).

Isaksson *et al.* (66) publicó recientemente una revisión general de las metodologías para elaborar modelos de predicción de toxicidad inducida por RT basadas en técnicas de aprendizaje automático. En esta revisión, los autores explican el funcionamiento de los algoritmos de aprendizaje automático, e incluyen un enfoque para el cálculo automático de los hiperparámetros del algoritmo mediante el modelado de la relación entre los datos de entrada y salida. Sin embargo, los autores hacen hincapié en la falta de comparabilidad entre los modelos de predicción encontrados en la literatura debido, por una parte, a la heterogeneidad de las metodologías empleadas para su implementación y, por otra parte, a la ausencia en estas publicaciones de métricas de validación del rendimiento comparables, problema que sigue sin resolverse hasta la fecha. Esta afirmación también es compartida por otros investigadores (67,68), añadiendo a estos problemas la falta de interpretabilidad de los modelos predictivos más avanzados como otra de las barreras que dificultan la aplicabilidad clínica y la integración en la rutina asistencial de estos métodos de predicción.

La aplicación de métodos de aprendizaje automático para predecir la toxicidad inducida por la RT ha sido un reto para la comunidad investigadora durante más de una

década. Ya en el año 2007, Dehing-Oberije *et al.* (69) realizó una primera aproximación a la aplicación de técnicas de aprendizaje automático para construir un modelo de predicción de la toxicidad pulmonar basado en variables dosimétricas, demográficas y clínicas, logrando un área bajo la curva característica operativa del receptor (AUC) igual a 0.67 con un modelo multivariante derivado de una cohorte de 376 pacientes con CP que incluía como variables la escala de nivel funcional del paciente de la OMS, el FEV1, el hábito tabáquico, la dosis total de radiación recibida y variables dosimétricas. Un hallazgo interesante en este trabajo fue que las variables que representaban las características de los pacientes tenían un mayor poder predictivo que los parámetros dosimétricos.

Oh *et al.* (70) desarrolló en el año 2009 un enfoque basado en técnicas de aprendizaje automático para predecir la neumonitis por radiación que incluía varios métodos de selección de características y algoritmos de clasificación utilizando para ello una población de 209 pacientes con CPNM en la que 48 sujetos fueron diagnosticados con neumonitis por radiación. El mejor rendimiento en términos del coeficiente de correlación de Matthew (MCC) (71) se consiguió con la combinación de un método de selección de características basado en la correlación y un clasificador basado en una máquina de vector soporte (SVM), que obtuvo un MCC=0.41 (equivalente a un AUC=0.72) con un modelo que incluía 5 variables que representaban parámetros clínicos, patológicos y dosimétricos.

En el año 2011, Pella *et al.* (72) exploraron la aplicación de un enfoque basado en técnicas de aprendizaje automático para predecir la aparición de toxicidades agudas en el intestino y la vejiga en un conjunto de 321 pacientes con cáncer de próstata. Tras extraer manualmente las características relevantes basándose en la experiencia clínica del equipo investigador, compararon el rendimiento de un modelo SVM frente a una red neuronal artificial (ANN) optimizada mediante un algoritmo genético, obteniendo finalmente resultados similares (AUC=0.70 y AUC=0.69, respectivamente).

Ya en 2018, Lee *et al.* (73) publicó un enfoque de aplicación de técnicas de aprendizaje automático novedoso que incluía información genética (polimorfismos de nucleótido simple, SNP). Aplicaron un algoritmo de *Random Forest* (RF) precondicionado para predecir la toxicidad crónica genitourinaria en función de un conjunto de variables clínicas y genéticas. La selección de características se realizó en base a un análisis univariante realizado sobre cada

variable del conjunto original, y los valores que presentaban alguna correlación estadísticamente significativa se desecharon para el modelo de predicción. Siguiendo esta estrategia, lograron una precisión de $AUC=0.70$ para la predicción de flujo urinario débil. En este trabajo, es interesante la inclusión de información genética en el modelo con el objetivo de explorar el papel de varios SNPs en la aparición de toxicidad crónica.

Otro trabajo que incluye variables dosimétricas y no dosimétricas en la predicción de la aparición de neumonitis por radiación fue publicado recientemente por Cui *et al.* (74). Su enfoque demostró el potencial de la combinación de métodos tradicionales de aprendizaje automático y técnicas de aprendizaje profundo cuando se trata de conjuntos de datos limitados para modelar las toxicidades inducidas por RT. Con una combinación de características clínicas seleccionadas a mano y variables dosimétricas y ómicas seleccionadas automáticamente en base al cálculo de variables latentes, sumando 24 características en total, su modelo de predicción alcanzó una precisión de $AUC=0.83$.

Otro estudio sobre el uso de técnicas de aprendizaje automático para predecir la neumonitis por radiación (75) en una cohorte de 203 pacientes con CP no microcítico (CPNM) empleó variables clínicas y dosimétricas en un esfuerzo por construir un modelo predictivo que proporcionara a los oncólogos radioterápicos valores de corte interpretables que facilitaran la planificación del tratamiento de RT. Aunque el modelo desarrollado no alcanzó una precisión destacable ($AUC=0.66$), es notable el esfuerzo realizado para proporcionar un modelo de aprendizaje automático interpretable para los usuarios finales.

TABLA 1. EJEMPLOS DE MODELOS DE PREDICCIÓN DE TOXICIDAD INDUCIDA POR RT BASADOS EN MODELOS DE APRENDIZAJE AUTOMÁTICO.

Autor	Toxicidad	Población	Selección de características	Método ML	Validación	Variables	Métricas
Dehing-Oberije, 2007 (69)	Toxicidad pulmonar aguda (CTCAE 3.0)	376 pacientes de CP tratados con quimio-RT	Selección hacia delante y eliminación hacia atrás	Modelo basado en análisis univariante	Partición 70/30 (100 repeticiones)	V20, D _{mean} pulmón, OMS-estado funcional, FEV1, tabaquismo, dosis total de radiación	AUC
Hun-Oh, 2009 (70)	Neumonitis	209 pacientes con CPNM (160 variables)	SVM-RFE, CFS, Chi-cuadrado, ganancia de información	SVM, J48 DT, RF, NB	10-fold CV (10 repeticiones)	COMSI, estado funcional, seguimiento, D20_lungMC, MOH10_heartMC	MCC, AUC
Pella, 2011 (72)	Toxicidad aguda en intestino y vejiga (RTOG/EORTC)	321 pacientes de cáncer de próstata (13 variables)	Manual, basado en experiencia clínica	RNA optimizada con AG, SVM	Entrenamiento con el 70-90% de las muestras. Sin repetición	Edad, riesgo de recurrencia, TURP, Hormonoterapia, tratamiento, días de RT, IGRT, DR30, DR60, DB50, PTV, volumen de vejiga, volumen del recto	AUC
Lee, 2018 (73)	Frecuencia urinaria, urgencia, nocturia y flujo débil (IPSS)	324 pacientes con cáncer de próstata (SNPs y 14 variables)	<i>P</i> -values. Chi-cuadrado (categóricas) y regresión logística (contínuas)	RF pre-condicionada	5-fold CV (100 repeticiones)	SNPs, estadio, ADT, tipo de RT, tabaquismo, hipertensión, diabetes, puntuación Gleason, α -bloqueantes, edad, PSA, dosis total, volumen de la próstata, IPSS basal, etnia.	AUC
Cui, 2019 (74)	Neumonitis (CTCAE 3.0)	106 pacientes de CPNM (230 variables)	RF, SVM, MLP, VAE, VAE+MLP, manual	MLP, RF, SVM, VAE+MLP	5-fold CV (10 repeticiones)	D _{mean} pulmón, Citoquinas, miRNA, SNPs	AUC
Luna, 2019 (75)	Neumonitis	203 pacientes de CPNM (32 variables)	DT	RF, CART, RUSBoost, regresión logística, SVM	5-fold CV (20 repeticiones)	D _{max} esófago, V20 pulmón, D _{mean} pulmón, tabaquismo, dosis total, V5 pulmón	AUC
Yu, 2021 (76)	Fibrosis pulmonar (CTCAE 3.0)	185 pacientes de CPNM (42 variables)	SVM (restringida a 3 variables)	SVM balanceada, GLM	5-fold CV (5 repeticiones)	CCL4, D _{mean} pulmón, quimioterapia	PPV, NPV, AUC, PRAUC

CTCAE: *Common Terminology Criteria for Adverse Events*; CP: *Cáncer de pulmón*; RT: *Radioterapia*; V20: *Volumen de tejido sano que recibe 20 Gy de radiación*; OMS: *Organización Mundial de la Salud*; FEV1: *Volumen espiratorio forzado en 1 segundo*; AUC: *Área bajo la curva característica operativa del receptor*; CPNM: *Cáncer de pulmón no microcítico*; SVM: *Máquina vector soporte*; RFE: *Eliminación recursiva de características*; CFS: *Selección de características basada en la correlación*; DT: *Árboles de decisión*; RF: *Random Forest*; NB: *Näive Bayes*; CV: *Validación cruzada*; COMSI: *Centro de masa de la localización del tumor en la dirección superior inferior*; D20_lungMC: *Dosis mínima al 20% del volumen más radiado*; MOH10_heartMC: *Dosis media al 10% del volumen más radiado*; MCC: *Coefficiente de correlación de Matthew*; RTOG/EORTC: *Radiation Therapy Oncology Group/European Organization for Research and Treatment of Cancer*; RNA: *Red neuronal artificial*; AG: *Algoritmo genético*; TURP: *Resección de próstata trans-uretral previa a la RT*; IGRT: *RT guiada por imagen*; DR30: *Dosis recibida por el 30% del volumen rectal*; DR60: *Dosis recibida por el 60% del volumen rectal*; DB50: *Dosis recibida por el 50% del volumen de la vejiga*; PTV: *Volumen objetivo en planificación*; IPSS: *International Prostate Symptom Score*; SNP: *Polimorfismo de nucleótido simple*; GWA: *Asociación del genoma completo*; PCA: *Análisis de componentes principales*; ADT: *Terapia supresora de andrógenos*; MLP: *Perceptrón multicapa*; VAE: *Autocodificador variacional*; GLM: *Modelo lineal generalizado*; CCL4: *Ligandos de quimioquinas 4*; PPV: *Valor predictivo positivo*; NPV: *Valor predictivo negativo*; PRAUC: *Curva de balance entre precisión y acierto*.

En un estudio reciente para predecir la fibrosis pulmonar inducida por la RT, Yu *et al.* (76) emplearon una SVM ponderada para extraer las 3 características con mayor poder de predicción sobre un conjunto de datos compuesto por 185 pacientes con CPNM siendo cada paciente representado con 42 variables que incluían información detallada sobre los niveles de citoquinas en plasma además de parámetros clínicos y dosimétricos, logrando un AUC=0.85 en una cohorte de validación externa.

Estos son algunos ejemplos de publicaciones relevantes cuyo objetivo era predecir algún tipo de toxicidad inducida por la RT en pacientes oncológicos que pueden ayudar a ilustrar el estado del arte actual. En la Tabla 1 se muestra un resumen de estas publicaciones en la que se destacan los criterios de valoración clínicos, la población, los métodos de selección de características, los métodos de aprendizaje automático, la división de la muestra entre conjunto de entrenamiento y conjunto de validación, las variables empleadas por el modelo y las métricas de rendimiento utilizadas en cada una de estas publicaciones.

Un problema común al que se enfrentan los científicos de datos cuando tratan con grandes cohortes de conjuntos de datos sanitarios obtenidos en el entorno asistencial (*Real-World Health Data* - RWHD) es el relacionado con la alta dimensionalidad que pueden alcanzar estos conjuntos de datos. En este escenario, los científicos de datos necesitan averiguar cuáles de esas variables, o características, son predictores fiables del criterio de valoración clínico seleccionado. Por ejemplo, si una característica tiene información irrelevante o redundante, o los datos recogidos son ruidosos o poco fiables, la introducción de estas variables en el proceso de entrenamiento del modelo de predicción penalizará su precisión (77).

Es relevante resaltar la situación regulatoria que actualmente enfrentan las aplicaciones que incluyen algún elemento basado en IA en el ámbito europeo. En este sentido, la denominada EU AI Act (78), actualmente en estado de borrador, busca regular el uso de estas aplicaciones con un enfoque basado en el riesgo del potencial daño que pueda causar dicha aplicación sobre la salud, seguridad o los derechos fundamentales de las personas. Por tanto, la propuesta regulatoria divide las aplicaciones que incluyen algún elemento basado en IA en tres grandes bloques: prácticas prohibidas; aplicaciones de alto riesgo; y resto de aplicaciones.

Las aplicaciones de IA en salud, generalmente clasificadas como aplicaciones de alto riesgo, requerirán una evaluación de conformidad por parte del proveedor y, en el caso de que dicha aplicación sea considerado un producto sanitario, también deberá estar a lo dispuesto en la Regulación sobre Productos Sanitarios (79).

Previo al desarrollo normativo en curso, la Comisión Europea ya había desarrollado un catálogo de requisitos a cumplir para que los sistemas de IA pudieran considerarse confiables a través del grupo de expertos de alto nivel en IA. Este grupo elaboró en 2020 el informe ALTAI (“*Assessment List for Trustworthy AI for self-assessment*”) (80) en el que se recogían los siguientes requisitos sobre los sistemas de IA: 1) deben trabajar bajo supervisión humana; 2) deben cumplir criterios de seguridad y robustez técnica; 3) debe existir una gobernanza de datos que preserve la privacidad; 4) transparencia; 5) deben evitar sesgos que provoquen inequidades; 6) deben preservar el bienestar social y medioambiental; y 7) deben tener mecanismos de rendición de cuentas (*accountability*).

1.4. Hipótesis y objetivos

Según la bibliografía revisada, se ha podido constatar que existe un alto grado de variabilidad metodológica respecto a las diferentes iniciativas aplicadas al desarrollo y validación de modelos de predicción de toxicidad inducida por la RT. Esta variabilidad dificulta la comparación entre diferentes experiencias y lastra la generación de evidencia suficiente para su aplicación a gran escala en la rutina asistencial. Por tanto, en base a esta falta de homogeneidad metodológica, construimos nuestra hipótesis principal de investigación:

H1. Se puede aplicar una metodología común que permita la comparación de resultados obtenidos por los diferentes enfoques basados en modelos de aprendizaje automático a la hora de desarrollar y validar modelos de predicción de toxicidad inducida por RT en pacientes con cáncer de pulmón.

Por otro lado, se tienen evidencias de que el riesgo de sufrir algún tipo de toxicidad inducida por la RT aumenta de forma directamente proporcional con las dosis de radiación que recibe el tejido sano circundante al tumor. Además, parece existir una correlación con características clínicas como la edad, el estado funcional del paciente, el hábito tabáquico, los tratamientos complementarios o concomitantes (quimioterapia, hormonoterapia, etc.), otras comorbilidades y el genotipo de cada paciente, entre otros factores. Esta evidencia procede de los resultados obtenidos en investigaciones realizadas en entornos controlados a partir de información recogida y estructurada con esta finalidad específica. Sin embargo, estas situaciones controladas distan mucho de las que se encuentran en la rutina asistencial, por lo que su aplicación en el mundo real aún presenta numerosos desafíos (81). En base a este conocimiento, y derivado de la hipótesis principal de esta tesis, planteamos la siguiente hipótesis secundaria:

H2. El uso de datos del mundo real para el desarrollo de modelos basados en algoritmos de aprendizaje automático de predicción de toxicidad asociada al tratamiento de RT en pacientes con cáncer de pulmón permite la validación del conocimiento obtenido a partir de investigaciones realizadas en entornos controlados.

Para dar una respuesta satisfactoria a las hipótesis planteadas, el objetivo principal de esta investigación es el de validar una metodología clara y generalizable para la implementación y validación externa de los modelos predictivos de toxicidad asociada a la RT basados en algoritmos de aprendizaje automático y entrenados con datos obtenidos en el entorno clínico-asistencial (datos del mundo real) para facilitar el avance y adopción de esta tecnología de apoyo a la toma de decisiones clínicas en la rutina asistencial durante el manejo de los pacientes con cáncer de pulmón y, por extensión, al resto de patologías que puedan beneficiarse de esta tecnología. Además, como resultado de la validación de la metodología propuesta, podremos obtener evidencias acerca del nivel de aplicabilidad del conocimiento generado a través de iniciativas de investigación en entornos controlados, lo que contribuirá a facilitar la adopción de la tecnología en este contexto. Relacionado con este último objetivo, y con carácter eminentemente exploratorio, este trabajo también permitiría evidenciar la

capacidad de esta metodología para la generación de hipótesis de investigación basadas en el análisis avanzado de datos procedentes del entorno asistencial. En este sentido, podemos formular los objetivos de la tesis de la siguiente forma:

O1. Validar una metodología clara y generalizable para el desarrollo y validación de modelos de predicción basados en técnicas de aprendizaje automático para predecir la toxicidad asociada al tratamiento de RT en pacientes con cáncer de pulmón con un conjunto de modelos de predicción representativos y generados a partir de datos obtenidos desde la práctica asistencial (datos del mundo real).

O2. Obtener evidencias acerca del nivel de aplicabilidad del conocimiento obtenido mediante investigaciones en entornos controlados a partir de la información recogida en el entorno asistencial.

O3 (exploratorio). Evidenciar la capacidad de esta metodología para la generación de nuevas hipótesis de investigación basadas en el análisis avanzado de datos procedentes del mundo real acerca de factores predictores de toxicidad inducida por RT.

1.5. Trabajos realizados en el marco de esta investigación

Esta tesis doctoral se enmarca en el proyecto de investigación “Sistema de soporte al tratamiento sistémico y de dosificación radioterápica en pacientes con cáncer de pulmón (S32)” (PI16/02104) financiado por el Instituto de Salud Carlos III (ISCIII) a través de la convocatoria de Proyectos de Investigación del año 2016, formando parte de la Acción Estratégica en Salud que promueve el ISCIII. El proyecto comenzó a ejecutarse en el año 2017 y finalizó en el año 2020.

El proyecto S32 tenía como objetivo desarrollar una infraestructura informática orientada a la normalización, reutilización y análisis de la información de los pacientes de

cáncer de pulmón para aplicarla en la definición y validación de nuevos modelos terapéuticos asociados a las nuevas terapias diana y a las nuevas técnicas de administración de RT.

El proyecto contemplaba la aplicación de técnicas de minería de datos basadas en algoritmos de redes neuronales artificiales, árboles de decisión, clasificadores Bayesianos y reglas de asociación sobre la información generada durante la aplicación de tratamientos de quimioterapia, moléculas diana y RT entre los años 2005 y 2015 en el Hospital Universitario Virgen del Rocío (HUVR). Estos análisis permitirían identificar patrones terapéuticos individualizados que incorporaran los resultados en salud del tratamiento del histórico de pacientes, estableciendo una nueva base para la definición de un protocolo de atención a pacientes de cáncer de pulmón en base a modelos predictivos de toxicidad y supervivencia.

El proyecto se ha basado en la infraestructura desarrollada en el proyecto S31 (82), financiado por ISCIII (PI13/01155), para implementar un sistema de ayuda a la decisión que estableciera recomendaciones personalizadas para la planificación del tratamiento en pacientes con patología tumoral pulmonar. Inicialmente, el proyecto preveía evaluar la sensibilidad y especificidad de los modelos terapéuticos propuestos mediante un estudio cuasiexperimental de 10 meses de duración sobre una población estimada de en torno a 100 pacientes.

Por tanto, el objetivo del proyecto era el desarrollo de una infraestructura informática orientada a la normalización, reutilización y análisis de la información de los pacientes de cáncer de pulmón para aplicarla en la definición y validación de nuevos modelos terapéuticos asociados a los tratamientos sistémicos, incluyendo las nuevas terapias diana, y a las nuevas técnicas de administración de RT.

Teniendo en cuenta este marco de trabajo, esta tesis se ha centrado en el desarrollo y validación de la metodología que respaldaría a un sistema de ayuda a la toma de decisiones terapéuticas basado en la implementación de varios modelos predictivos de toxicidades asociadas al tratamiento de RT. Como consecuencia de las actividades desarrolladas en torno a esta investigación, se han realizado las siguientes acciones de difusión a la comunidad científica en los ámbitos de oncología radioterápica e informática médica:

1. Año 2018: Presentación de la metodología.

- a. Póster en el congreso internacional ESTRO 37.
 - i. Autores: JL López-Guerra, Beatriz Pontes, A Moreno, Cristina Rubio, **F Núñez**, Isabel Nepomuceno, J Moreno, J Cacicedo, JM Praena-Fernandez, GA Escobar Rodriguez, C Parra, Jose Riquelme, MJ Ortiz-Gordillo.
 - ii. Título de la comunicación: “Project S32: decision support system for lung cancer patients”.
 - iii. Organizado por la Sociedad Europea para la Radioterapia y la Oncología, celebrado en Barcelona (España).
 - b. Póster en el congreso nacional INFORSALUD 2018.
 - i. Autores: J. Moreno Conde, A. Moreno Conde, **F. Núñez Benjumea**, M. J. Ortiz Gordillo, C. L. Parra Calderón, J. L. López Guerra.
 - ii. Título de la comunicación: “S32: Sistema de Soporte a la Decisión para el Tratamiento Sistémico y la Dosificación de Radioterapia”.
 - iii. Organizado por la Sociedad Española de Informática de la Salud, celebrado en Madrid (España)
2. Año 2019: Presentación de resultados preliminares.
- a. Comunicación oral en el congreso nacional GRX Health Data 2019.
 - i. Autores: **Benjumea, F. N.**, Conde, J. M., Conde, A. M., García, S. G., Gordillo, M. O., Riquelme, J., & Guerra, J. L.
 - ii. Título de la comunicación: “Predicción de toxicidad a través de la plataforma ITC-Bio”.
 - iii. Organizado por la Escuela Andaluza de Salud Pública, celebrado en Granada (España).
 - b. Póster en el congreso internacional ESTRO 2019.
 - i. Autores: **Nunez Benjumea F**; Moreno Conde J; Moreno Conde A; Gonzalez Garcia S; Ortiz Gordillo MJ; Riquelme J; Fernández Fernández MDC; Parra Calderón CL; López Guerra JL.
 - ii. Título de la comunicación: “S32: A decision Support System to predict radiation toxicity in lung cancer patients”.
 - iii. Organizado por la Sociedad Europea para la Radioterapia y la Oncología, celebrado en Milán (Italia).
-

- c. Comunicación oral en el congreso nacional INFORSALUD 2019.
 - i. Autores: Sara González García, **Francisco José Núñez Benjumea**, Jesús Moreno Conde, Alberto Moreno Conde, José Luis López Guerra, Blas David Delgado León, Carlos Luis Parra Calderón, María José Ortiz Gordillo.
 - ii. Título de la comunicación: “Métodos de selección de características para la predicción de toxicidad inducida por radioterapia”.
 - iii. Organizado por la Sociedad Española de Informática de la Salud, celebrado en Madrid (España). Cabe destacar que esta comunicación recibió el 1^{er} premio a la mejor comunicación del congreso.
 - d. Comunicación oral en el congreso internacional EFMI STC 2019.
 - i. Autores: **Núñez-Benjumea, F. J.**, Moreno-Conde, J., González-García, S., Moreno-Conde, A., López-Guerra, J. L., Ortiz-Gordillo, M. J., & Calderón, C. L. P.
 - ii. Título de la comunicación: “Comparison of Feature Selection Methods for Predicting RT-Induced Toxicity”.
 - iii. Organizado por la Federación Europea de Informática Médica (EFMI), celebrado en Hannover (Alemania).
3. Año 2020: Presentación de resultados parciales.
- a. Póster en el congreso internacional ESTRO 2020. Autores:
 - i. Manuel Borrego, Jon Cacicedo, Blas David Delgado, Sara González-García, Jose Luis Lopez Guerra, Alberto Moreno-Conde, Jesús Moreno-Conde, **Francisco Núñez-Benjumea**, Carlos Luis Parra-Calderón, Sonia Perez Luque.
 - ii. Título de la comunicación: “Feature selection methods improve accuracy in radiation toxicity prediction for lung cancer”.
 - iii. Organizado por la Sociedad Europea para la Radioterapia y la Oncología, celebrado online debido a la pandemia de COVID-19.
 - b. Póster en congreso nacional XVI Foro de Investigadores.
 - i. Autores: S. González-García, **F.J. Núñez Benjumea**, J. Moreno Conde, A. Moreno Conde, M. Borrego Reina, Sonia Pérez Luque, B.D. Delgado León, M.J. Ortiz Gordillo, C. L. Parra Calderón, J.L. López Guerra.
-

- ii. Título de la comunicación: “S32: Métodos de selección de características para predicción de toxicidad inducida por Radioterapia”.
- iii. Organizado por el Instituto de Biomedicina de Sevilla, celebrado en Sevilla (España).

Además de estas ocho comunicaciones, recientemente ha sido publicado el artículo “Benchmarking machine learning approaches to predict radiation-induced toxicities in lung cancer patients” (83) en la revista “Clinical and Translational Radiation Oncology”, donde se presentan los resultados obtenidos a partir de esta investigación.

2. Metodología para el desarrollo y validación de modelos de predicción de toxicidad asociada al tratamiento de RT

La metodología propuesta para el desarrollo y validación de modelos de predicción de toxicidad asociada al tratamiento de RT basados en técnicas de aprendizaje automático desarrolladas a partir de datos obtenidos en entorno asistencial gira en torno a dos elementos principales: el conjunto de datos usado y los modelos predictivos basados en técnicas de aprendizaje automático desarrollados.

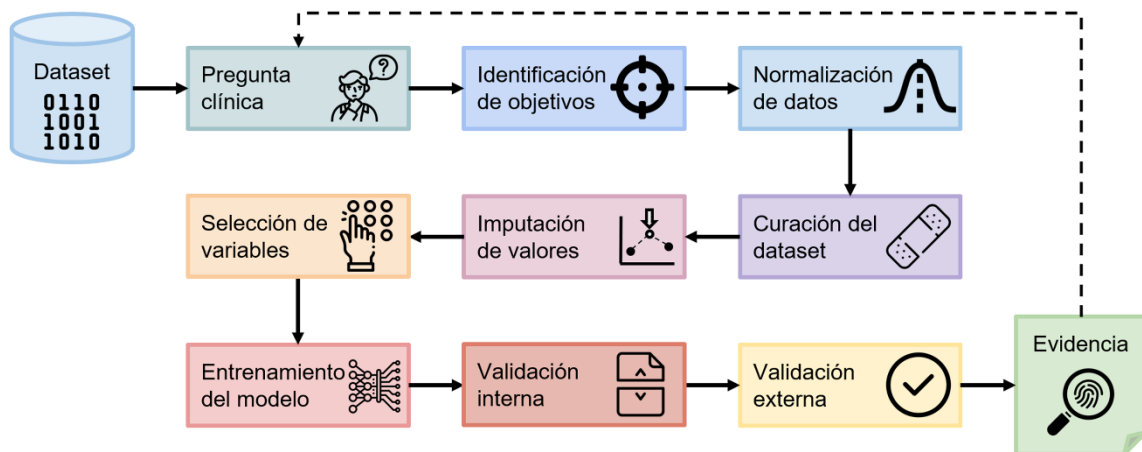


FIGURA 5. METODOLOGÍA PROPUESTA PARA LA IMPLEMENTACIÓN DE MODELOS DE PREDICCIÓN DE TOXICIDAD ASOCIADA AL TRATAMIENTO DE RT BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO DESARROLLADOS A PARTIR DE DATOS OBTENIDOS EN EL ENTORNO ASISTENCIAL.

Para facilitar la implementación técnica de esta metodología en la práctica clínica, se han definido un conjunto de pasos que pueden resumirse de la siguiente manera: 1) Definir la pregunta clínica; 2) Identificar el objetivo del modelo de aprendizaje automático; 3)

Normalización de los datos; 4) Curación del conjunto de datos; 5) Imputación de valores perdidos; 6) Validación interna; 7) Entrenamiento de los modelos de aprendizaje automático; 8) Selección de las variables del modelo; 9) Validación externa; y 10) Generación de nueva evidencia, como se muestra en la Figura 5.

La implementación de esta estrategia puede facilitar el descubrimiento de nuevas evidencias basadas en datos que, a su vez, puede desencadenar la generación de nuevas hipótesis y preguntas clínicas, cerrando así el ciclo y contribuyendo al establecimiento de un sistema sanitario con capacidad de aprendizaje (84).

2.1. El conjunto de datos

El conjunto de datos utilizado para esta tesis tiene su origen en el registro de cáncer de pulmón S31 (82). Este registro contiene información obtenida de forma automática a partir de diferentes sistemas de información hospitalaria y complementado con información registrada manualmente durante las consultas de oncología radioterápica del HUVR. Este registro, que se diseñó partiendo de un consenso de todos los profesionales sanitarios implicados en la asistencia al paciente con CP, incluyendo cirujanos, oncólogos radioterápicos, oncólogos médicos, anatomopatólogos, neumólogos y radiólogos, aborda un conjunto completo de conceptos clínicos asociados al tratamiento y seguimiento de los pacientes con CP sometidos a cualquier tratamiento oncológico.

El registro S31 incluye información de 875 pacientes consecutivos con CP atendidos por la Unidad de Oncología Radioterápica del HUVR entre los años 2013 y 2020. Los conceptos clínicos fundamentales incluidos en el registro son:

- Datos demográficos y situación socioeconómica.
- Datos de asistencia clínica: registro del histórico de asistencia clínica recibida incluyendo fecha, especialidad, motivo de consulta, profesional sanitario consultado y centro de referencia.
- Antecedentes familiares: parentesco y localización del tumor.

- Antecedentes personales: alergias, hábitos tóxicos (tabaquismo, alcoholismo), antecedentes respiratorios (EPOC, oxigenoterapia), antecedentes neoplásicos, antecedentes generales, antecedentes quirúrgicos, calidad de vida.
- Medicación: medicación actual, especificando nombre, dosis, unidades, vía, frecuencia, fechas de inicio y fin.
- Diagnóstico: diagnóstico clínico, diagnóstico radiológico, anatomía patológica, marcadores tumorales séricos, pruebas genéticas, estadificación del tumor.
- Examen físico: Índice de Karnofsky, hallazgos de auscultación pulmonar y cardíaca.
- Tratamiento: quimioterapia, cirugía, radioterapia y dosimetría.
- Evolución: toxicidad y calidad de vida.
- Respuesta al tratamiento: fecha, tipo de respuesta, prueba clínica y decisión tomada.
- Último contacto: fecha, estado del paciente y hallazgos.

Además, este registro se diseñó con la intención de poder reconstruir el recorrido asistencial del paciente con CP y, por ello, la información de cada paciente se recoge según las cuatro etapas principales de este recorrido en los siguientes documentos:

- Informe de consulta: este documento incluye los datos clínicos del paciente antes de iniciar el tratamiento. Incluye datos demográficos, antecedentes personales, estado actual de la enfermedad, exploración física y tratamiento propuesto, entre otros.
- Dosimetría: Este documento contiene todos los datos relacionados con los informes dosimétricos de los sistemas de información de radioterapia y quimioterapia.
- Evolución y curso clínico: este documento recoge la información asociada a la evolución del paciente justo después del tratamiento, incluyendo información sobre toxicidad, eventos adversos y recaídas.
- Calidad de vida: este documento final incorpora las respuestas al cuestionario EORTC QLQ-C30 con el complemento específico de cáncer de pulmón QLQ-LC13 (85). Este cuestionario constituye un conjunto de ítems de valoración de diferentes aspectos relacionados con la calidad de vida del paciente de cáncer de pulmón. Esta escala se recoge durante las consultas de seguimiento al paciente.

El registro S31 se desarrolló haciendo uso de la plataforma ITC-Bio (86), que integra un conjunto de aplicaciones desarrolladas a partir de diferentes iniciativas de software de

código abierto para dar soporte a los procesos de investigación clínica y traslacional. En concreto, el registro S31 hace uso de las capacidades de gestión de *data warehouse* de la herramienta i2b2 (87) y, como interfaz para el registro manual de la información, se usó el software de gestión de ensayos clínicos OpenClinica (88). La integración de la plataforma ITC-Bio con los sistemas de información hospitalarios (HIS) se llevó a cabo mediante el bus de integración Mirth Connect (89). Esta aplicación es un motor de software de código abierto desarrollado específicamente para implementar integraciones de información y documentos relacionados con la salud, y que implementa de forma nativa el estándar de mensajería HL7. Haciendo uso de este motor, se habilitó la reutilización bajo demanda de la información clínica de los pacientes almacenada en el HIS para completar de forma automática algunas de las variables del registro S31.

2.2. Definición de la pregunta clínica

El primer paso para diseñar e implementar un modelo predictivo basado en técnicas de aprendizaje automático es definir la pregunta clínica que planeamos responder y luego asegurarnos de que nuestro conjunto de datos tiene toda la información que necesitamos para construir el modelo predictivo. La pregunta debe ser formulada por un experto clínico en la materia con un profundo conocimiento del tema, idealmente a partir de la experiencia adquirida en el ámbito asistencial, dado que aproximadamente la mitad de las preguntas planteadas e hipótesis que surgen en este contexto quedan sin respuesta (90).

Como ya se ha comentado en la introducción, uno de los factores más limitantes a la hora de planificar el tratamiento de RT es la gestión de la probabilidad de la aparición de toxicidades inducidas por el tratamiento (14,15). Este hecho limita la administración de dosis elevadas de radiación que permitirían una mejor respuesta tumoral, lo que se traduciría en una potencial mejora de la supervivencia global del paciente. Esto plantea la cuestión de con qué precisión y nivel de personalización podemos predecir la aparición y gravedad de toxicidades inducidas por el tratamiento de RT, de modo que la planificación del tratamiento pueda optimizarse para mejorar las posibilidades de supervivencia global de los pacientes.

En este trabajo, esta pregunta fue planteada por un oncólogo radioterápico experto en planificación de tratamientos de RT y, tras confirmar con la literatura existente que la pregunta aún no ha sido respondida satisfactoriamente, y que se disponía de los recursos necesarios para tratar de responderla, el equipo de investigación consideró que era una pregunta relevante a la que se podía dar respuesta con un enfoque basado en el desarrollo de modelos predictivos de toxicidad asociada al tratamiento de RT mediante técnicas de aprendizaje automático.

2.3. Definición de las variables clínicas objetivo

Como se ha comentado extensamente en la sección 1.1, en los pacientes con CP sometidos a tratamiento de RT, las toxicidades comúnmente reportadas están principalmente relacionadas con las vías respiratorias y el tórax (neumonitis, tos, disnea, dolor de la pared torácica, fibrosis, hipoxia, derrame pleural, fractura de costillas), el tracto digestivo (disfagia, esofagitis, anorexia, náuseas), la piel (dermatitis, eritema), el sistema cardiovascular (hemoptisis, leucopenia, trombocitopenia, derrame pericárdico) y la fatiga (91–94). Estas toxicidades inducidas por la RT se clasifican cuantitativamente de 0 a 5 puntos según su nivel de gravedad, que va desde la no presencia (0 puntos) hasta la muerte (5 puntos), en base a los Criterios Terminológicos Comunes para Eventos Adversos (CTCAE) (95), desarrollados y mantenidos por el Instituto Nacional del Cáncer de los NIH, siendo una de las escalas de clasificación más utilizadas para este fin.

La definición de las toxicidades específicas que serían abordadas en esta investigación se realizó en función de la relevancia para el manejo clínico del paciente y su impacto en la planificación del tratamiento de RT y en base a la frecuencia con la que fueron notificadas en el registro S31. Estas toxicidades se clasificaron en aguda o crónica, considerando toxicidad aguda aquella que aparece en los primeros 3 meses tras el inicio de la terapia, y toxicidad crónica la que permanece presentes más allá de ese tiempo. Las definiciones clínicas para cada una de las toxicidades abordadas se realizaron de acuerdo con la guía de CTCAE v5.0.

Teniendo en cuenta estos condicionantes, se definieron las siguientes toxicidades como variables objetivo para implementar los modelos de predicción basados en aprendizaje

automático: esofagitis aguda, tos aguda, disnea aguda, neumonitis aguda, disnea crónica y neumonitis crónica.

2.4. Normalización del conjunto de datos

La normalización de los datos es un paso importante para optimizar el rendimiento de los modelos predictivos basados en técnicas de aprendizaje automático, ya que puede aliviar la carga computacional durante la fase de entrenamiento y, también, porque permite minimizar el ruido no deseado que puede enmascarar la señal de interés (96). Por ejemplo, en las redes neuronales artificiales, mantener las entradas entre 0 y 1 evita que los pesos de la red sean demasiado grandes y, por tanto, optimiza la carga computacional y el tiempo de ejecución (97).

El registro S31 se inició en el año 2013, incluyendo a finales del año 2022 información de 875 pacientes consecutivos con CP. De cada paciente, se recogió información de alrededor de 400 conceptos clínicos pertenecientes a los dominios definidos en la sección 2.1. Para facilitar el tratamiento de esta información por parte de los algoritmos de aprendizaje automático, se procedió a la binarización de todas las variables de la siguiente forma: para las variables cuantitativas, se asignó un valor de '1' cuando el valor de la variable era superior o igual al valor medio de esa variable teniendo en cuenta todos los pacientes, asignando el valor '0' en caso contrario. Para las variables cualitativas o categóricas, se aplicó la técnica '*one hot encoding*'. Para las variables objetivo, es decir, para las toxicidades inducidas por RT, se aplicó la binarización de la siguiente manera: en el caso de que el paciente no hubiera reportado dicha toxicidad, o la hubiera reportado con carácter leve (valores 0 o 1 según CTCAE v5.0), se les asignó el valor '0', mientras que si la toxicidad había sido reportada con valores iguales o superiores a 2 según CTCAE v5.0, entonces se les asignó el valor de '1'. Esta decisión está basada en que únicamente las toxicidades reportadas con valor 2 o más según CTCAE v5.0 requieren de asistencia clínica.

Tras la binarización de todas las variables, cada paciente quedó representado en el conjunto de datos por un total de 4499 variables binarias.

2.5. Curación del conjunto de datos

Otro paso importante del proceso es la identificación de las variables que tienen potencial para contribuir a la predicción de la variable objetivo (y descartar las que no lo harán) haciendo uso del conocimiento que tengamos *a priori* sobre las variables objetivo en cuestión o del resto de variables del conjunto de datos.

Una variable se puede clasificar como: 1) fuertemente relevante, 2) débilmente relevante, pero no redundante, 3) irrelevante y 4) redundante (98). En este punto del proceso, se recomienda identificar las variables para las que ya existe alguna evidencia sobre su contribución a la predicción de la variable objetivo. Por ejemplo, existe una fuerte evidencia sobre la relación entre las variables dosimétricas y el inicio de la toxicidad, por lo que en nuestro caso de uso se recomienda mantener estas variables dentro del conjunto de datos que se usará posteriormente para entrenar el modelo de predicción.

Una estrategia complementaria sería eliminar aquellas variables con información redundante. Esto se puede hacer, por ejemplo, analizando la matriz de correlación entre todas las variables independientes, y eliminando las que obtengan una puntuación de ± 1 (es decir, están totalmente correlacionadas con alguna otra variable independiente del conjunto de datos). Por último, también es útil realizar un análisis estadístico básico para identificar y eliminar posibles variables que aporten ruido al modelo. De este modo, podríamos identificar variables "estáticas" (es decir, que tienen el mismo valor para todos los casos) que sólo añadirán ruido y coste computacional al proceso, por lo que también podríamos querer eliminarlas.

Por otra parte, dado que en este trabajo únicamente abordamos la predicción de seis toxicidades inducidas por la RT (esofagitis aguda, tos aguda, disnea aguda, neumonitis aguda, disnea crónica y neumonitis crónica), no se incluyeron todas las toxicidades reportadas por los pacientes y recogidas en el registro S31, sino únicamente las relacionadas con las seis toxicidades anteriormente mencionadas. Estas variables se extrajeron a partir de los valores reportados en el informe de evolución y curso clínico.

Además, dado que este trabajo estaba orientado a proporcionar a los oncólogos radioterápicos una herramienta de ayuda a la decisión clínica basada en la predicción de la toxicidad inducida por la RT para ayudarles a decidir entre las diferentes opciones de tratamiento disponibles en su entorno clínico, únicamente se utilizaron como variables independientes aquéllas que estaban a disposición del oncólogo radioterápico en el momento de la decisión terapéutica, es decir, aquéllas variables recogidas en los informes de consulta, dosimetría y los informes de calidad de vida basales.

Tras la eliminación de estas variables, la dimensionalidad del conjunto de datos se redujo de 4499 a 1104 variables.

2.6. Imputación de valores perdidos

Los valores perdidos son un reto recurrente cuando se trabaja con conjuntos de datos que incluyen series temporales multivariantes obtenidos a partir de un entorno asistencial, entre otros ámbitos. Estos valores que faltan pueden proporcionar información importante cuando se aplican metodologías de aprendizaje supervisado, como en el caso de los problemas de clasificación (99). Los valores perdidos se suelen clasificar en tres tipos diferentes en función del mecanismo que origina la pérdida del dato: valor perdido completamente aleatorio, valor perdido aleatorio y valor perdido no aleatorio (100). En una evaluación *a priori*, una suposición razonable sería considerar que los valores perdidos son del tipo completamente aleatorio. Esta suposición podría limitar el número de métodos de imputación de valores perdidos que podrían aplicarse posteriormente. Sin embargo, permite abordar el problema desde un punto de vista más generalizable (101).

Existen varios enfoques para la imputación de valores perdidos que pueden agruparse en dos categorías principales: la eliminación de la lista (es decir, eliminar los casos y/o las variables con valores perdidos) y la imputación de valores perdidos. Esta última incluye una amplia variedad de métodos que van desde la imputación con valores medios, la sustitución mediante técnicas de regresión y la imputación múltiple, entre otros (102). Little *et al.* (103) han realizado un análisis exhaustivo y profundo sobre los métodos de imputación de valores perdidos que puede consultarse para más información sobre este aspecto en particular.

En el caso que nos ocupa, se asumió la hipótesis de que la distribución de los valores perdidos era completamente al azar. Los valores perdidos en este conjunto de datos se gestionaron de la siguiente manera: por un lado, se eliminaron los casos (pacientes) que presentaban valores perdidos en más del 20% de las variables contempladas. Del mismo modo, se eliminaron aquellas variables independientes con valores perdidos en más del 20% de los casos. Por otra parte, la imputación de los valores perdidos en el conjunto de datos resultante se realizó haciendo uso de una técnica no paramétrica para datos de tipo mixto basada en el algoritmo Random Forest desarrollado por Stekhoven *et al.* (104), ya que se informó de que este método superaba a otros métodos en condiciones de trabajo similares (105).

Siguiendo este enfoque, conseguimos reducir la dimensionalidad de nuestro conjunto de datos inicial a 573 pacientes representados por 464 variables independientes, tras eliminar aquellos casos y variables con una contribución *a priori* escasa para los modelos de predicción, mientras que nos apoyamos en una técnica sólida para imputar el resto de los valores perdidos, aprovechando así la mayor cantidad de información posible para ser procesada en etapas posteriores.

Tras la aplicación de estos métodos de preprocesamiento sobre el conjunto de datos inicial, se obtuvieron los seis subconjuntos de datos representados en la Tabla 2.

TABLA 2. SUBCONJUNTOS DE DATOS RESULTANTES TRAS EL PREPROCESAMIENTO. ENTRE PARÉNTESIS, NÚMERO DE CASOS CON TOXICIDAD POSITIVA EN CADA SUBCONJUNTO.

Variable dependiente	Casos totales	Entrenamiento	Validación externa
Esofagitis aguda	454 (220)	408 (204)	46 (16)
Tos aguda	291 (134)	246 (123)	45 (11)
Disnea aguda	276 (127)	230 (115)	46 (12)
Neumonitis aguda	378 (184)	332 (166)	46 (18)
Disnea crónica	267 (119)	236 (118)	31 (1)
Neumonitis crónica	293 (138)	262 (131)	31 (7)

2.7. Métodos de selección de características

Las técnicas de selección de características suelen emplearse para reducir la dimensionalidad de los conjuntos de datos y optimizar el rendimiento de los modelos predictivos basados en aprendizaje automático (129). Según la estrategia utilizada para la selección de características, estas técnicas pueden dividirse en tres tipos principales según la estrategia de búsqueda (130,131):

- Métodos de filtro: las características se evalúan de acuerdo con las propiedades inherentes de los datos. Normalmente, se obtiene un buen rendimiento en términos de velocidad y escalabilidad.
- Métodos *wrapper* o de envoltura: generan subconjuntos de características utilizando un algoritmo de agrupación específico con una función de optimización dirigida por la precisión a la salida del clasificador. Computacionalmente, los métodos *wrapper* son más costosos que los de filtro.
- Métodos embebidos o híbridos: intentan aprovechar las ventajas de ambos enfoques, filtro y *wrapper*, tratando de equilibrar el esfuerzo computacional y el rendimiento.

Además, dependiendo del resultado de la selección realizada, estas técnicas pueden ser de evaluación individual (también conocida como clasificación de características) o de evaluación de subconjuntos de características. En la primera, las características se evalúan individualmente y se les asigna un peso que refleja su relevancia. En el segundo, los subconjuntos de características candidatas se evalúan utilizando una métrica determinada, con el fin de seleccionar el mejor subconjunto de características (132).

En este trabajo, se consideraron varios métodos de selección de características para ilustrar diferentes formas de reducir la dimensionalidad de los conjuntos de datos y mejorar el rendimiento de los modelos de aprendizaje automático. Dado que nuestros conjuntos de datos incluyen un gran número de variables independientes (464), decidimos implementar algunas técnicas de selección de características comúnmente utilizados para probar su idoneidad, viabilidad y para evaluar su contribución a la precisión de la predicción de los puntos finales clínicos seleccionados. Las técnicas utilizadas fueron las siguientes:

- Métodos de subconjunto:

- Selección de características basada en la correlación (CFS) (133). CFS utiliza una heurística basada en la correlación para evaluar la contribución de los subconjuntos de características al resultado de la clasificación. CFS primero evalúa y luego clasifica subconjuntos de características en lugar de características individuales. Se basa en la siguiente hipótesis: los subconjuntos de características más adecuados contienen características altamente correlacionadas con la clase, pero no correlacionadas entre sí. CFS calcula primero una matriz de correlaciones característica-clase y característica-característica a partir de los datos de entrenamiento y luego busca el mejor subconjunto en el espacio de subconjuntos de características utilizando un algoritmo de optimización.
 - Chi-cuadrado (χ^2) (134). Esta función selecciona el subconjunto de características sobre la base del algoritmo de discretización χ^2 . La función calcula los pesos de las características utilizando el estadístico χ^2 y realiza la clasificación de las características. De acuerdo con el valor del estadístico χ^2 para cada par de características consecutivas ordenadas según el valor χ^2 , la adición de nuevas características al subconjunto continúa hasta que se supere un nivel de inconsistencia. El algoritmo χ^2 determina automáticamente un umbral adecuado que mantiene la fidelidad de los datos originales.
 - Boruta (135). Boruta es una extensión del método de bosques aleatorios (*random forest*) que utiliza la medida de importancia generada por este método. En los bosques aleatorios, la puntuación de importancia por sí sola no es suficiente para identificar correlaciones significativas entre las variables y el atributo de decisión. Boruta supera este problema haciendo una copia aleatorizada del sistema, fusionando la copia con el original y construyendo el clasificador para este sistema ampliado. Para evaluar la importancia de la variable en el sistema original la comparamos con la de las variables aleatorias. Sólo se consideran importantes las variables cuya importancia es mayor que la de las variables aleatorias.
- Métodos de clasificación:

- Mínima redundancia-máxima relevancia (mRMR) (136). El método mRMR fue desarrollado originalmente para ser aplicado en el campo de la bioinformática para extraer los genes relevantes que mejor predicen los fenotipos a partir de los datos de expresión génica de *microarrays*, evitando la inclusión de información redundante. La mínima redundancia entre pares de características y la máxima relevancia entre una característica y la clase objetivo se proporciona maximizando la información mutua en el caso de las variables categóricas o la prueba F (F -test) para las variables continuas.
- Relief (137). Relief es un algoritmo basado en la ponderación de las características que detecta aquellas que son estadísticamente relevantes para la clase objetivo. Los pesos se calculan en función de la distancia euclídea entre las características y la clase objetivo, y se establece un umbral de antemano para decidir si la característica se considera relevante o no.
- Ganancia de información (IG) (134). Esta función calcula las ponderaciones de las características utilizando la medida del criterio de ganancia de información y realiza la clasificación de las características en orden decreciente en base a este criterio. Puede manejar tanto valores numéricos como categóricos. IG mide el valor de una característica calculando la medida del criterio de ganancia de información con respecto a la clase.
- Bosque aleatorio (RF) (138). Un bosque aleatorio es un clasificador que consiste en una colección de clasificadores estructurados en forma de árbol, donde cada árbol emite un voto unitario para la clase más popular en cada entrada. Las características utilizadas en cada árbol se asignan al azar, pero manteniendo una distribución idéntica en cada uno de ellos.

TABLA 3. RESUMEN DE LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS IMPLEMENTADOS A PARTIR DE LIBRERÍAS DE R.

Técnica	Librería R	Función	Estrategia	Método
mRMR	<i>Praznik</i>	<i>MRMR</i>	Wrapper	Clasificación
Relief	<i>FSelector</i>	<i>relief</i>	Wrapper	Clasificación
IG	<i>Biocomb</i>	<i>select.inf.gain</i>	Wrapper	Clasificación
CFS	<i>Biocomb</i>	<i>select.cfs</i>	Filtro	Subconjunto
χ^2	<i>Biocomb</i>	<i>select.inf.chi2</i>	Filtro	Subconjunto
RF	<i>randomForest</i>	<i>randomForest</i>	Wrapper	Clasificación
Boruta	<i>Boruta</i>	<i>Boruta</i>	Filtro	Subconjunto

Además de estas siete técnicas de selección de características, se implementaron otros dos métodos basados en votación, uno para los métodos de subconjunto (CFS, χ^2 y Boruta) y otro para los métodos de clasificación (mRMR, Relief, IG y RF). Estos métodos consideraron sólo aquellas variables que fueron seleccionadas por, al menos, 2 de las técnicas de selección de características implementados.

Por último, un oncólogo experto también proporcionó, para cada una de las toxicidades que se deseaba predecir, subconjuntos de características seleccionadas en base a la evidencia clínica, conformando así este último método el *gold standard* para ser comparado su rendimiento con el resto de las técnicas implementadas.

En total, este trabajo implementa 10 técnicas diferentes de selección de características que, cuando se utilizaron en combinación con los 5 modelos predictivos basados en aprendizaje automático descritos en la sección 2.8, produjeron un total de 50 modelos de predicción diferentes basados en técnicas de aprendizaje automático para cada toxicidad bajo estudio, haciendo un total de 300 modelos de predicción diferentes, permitiendo de esta forma realizar una evaluación comparativa de su rendimiento en base al conjunto de datos obtenidos en entorno asistencial.

2.8. Entrenamiento de los modelos de predicción basados en técnicas de aprendizaje automático

Entre los diferentes retos que debemos afrontar durante el desarrollo de un modelo de predicción de toxicidad inducida por RT basado en técnicas de aprendizaje automático, debemos decidir qué algoritmo de aprendizaje automático se utilizará y cómo abordar los posibles problemas relacionados con el ajuste del modelo durante la fase de entrenamiento (109). La decisión sobre el qué modelo de aprendizaje automático utilizar depende en gran medida del tipo de problema que pretendemos resolver. Tanto si se trata de una tarea de clasificación, regresión o agrupación (*clustering*), y dependiendo del método de aprendizaje que se vaya a implementar (supervisado, no supervisado, semi-supervisado o por refuerzo),

se deben considerar diferentes algoritmos de aprendizaje automático que se adapten mejor a cada situación (110). Además, dependiendo del algoritmo de aprendizaje automático elegido, es posible que queramos optimizar su rendimiento en función de las características específicas de nuestro conjunto de datos. Esto puede hacerse mediante el ajuste de los hiperparámetros del algoritmo (111).

En nuestro caso, las variables objetivo se definieron como una función binaria (toxicidad clínicamente relevante presente/ausente) y pueden considerarse como dos clases diferentes, lo que hace que el uso de algoritmos clasificadores sea el más adecuado para el propósito de este trabajo.

Para evaluar la idoneidad de la metodología propuesta, se han implementado cinco clasificadores comúnmente utilizados que representan diferentes niveles de complejidad (e interpretabilidad) y rendimiento. Además, en esta investigación se profundizará acerca de la idoneidad de estos clasificadores en función de las métricas de precisión obtenidas para cada uno de ellos.

Los cinco clasificadores seleccionados son: 1) Máquina de vector soporte (SVM) con kernel radial (112); 2) k -ésimo vecino más cercano (kNN) con kernel radial (113), Red neuronal artificial (ANN) con retroalimentación hacia delante (114), Modelo lineal generalizado (GLM) siguiendo una distribución binomial (115) y Naïve Bayes (NB) con suavizado Laplaciano (116).

En cuanto al ajuste de los hiperparámetros de estos clasificadores, se siguió un enfoque común a todos los clasificadores para demostrar su reproducibilidad. En este sentido, se realizó una optimización automática de los hiperparámetros siguiendo el método de búsqueda en rejilla (117) dentro de los rangos recomendados en la literatura. Esta búsqueda en rejilla se realizó siguiendo una validación cruzada de 5 particiones tomadas del subconjunto de datos de entrenamiento, es decir, se ejecuta sobre 5 subconjuntos de entrenamiento diferentes, y la precisión media resultante se utiliza para seleccionar los hiperparámetros optimizados. Una vez determinados los hiperparámetros óptimos, se siguió el proceso de validación interna comentado en la sección 2.9. La Tabla 4 resume las características de los clasificadores basados en técnicas de aprendizaje automático utilizados en este trabajo para producir los modelos de predicción.

TABLA 4. CARACTERÍSTICAS DE LOS ALGORITMOS CLASIFICADORES EMPLEADOS.

Clasificador	Librería R	Hiperparámetros	Valores de búsqueda
SVM	<i>e1071</i>	C, γ	$0.1 < C < 1000; 0.001 < \gamma < 1000$
kNN	<i>class</i>	k	$5 < k < 23$
ANN	<i>nnet</i>	Size, decay	$2 < \text{size} < 232; 0.0001 < \text{decay} < 0.1$
GLM	<i>stats</i>	N/A	N/A
NB	<i>e1071</i>	N/A	N/A

2.8.1. Máquina de vector soporte (SVM)

SVM es un método popular de aprendizaje automático que puede emplearse para la clasificación, regresión y otras tareas de aprendizaje supervisado. Desarrollado originalmente por Cortes y Vapnik en el año 1995 (112), SVM implementa conceptualmente la siguiente idea: los vectores de entrada se asignan de forma no lineal a un espacio de características de muy alta dimensión. En este espacio de características se construye una superficie de decisión lineal. Las propiedades especiales de esta superficie de decisión garantizan una gran capacidad de generalización de esta máquina de aprendizaje.

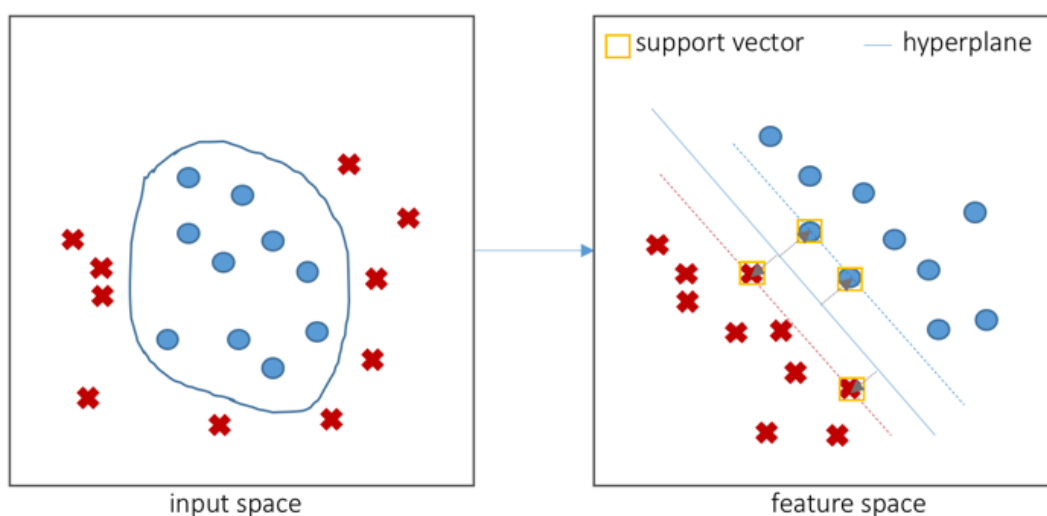


FIGURA 6. EJEMPLO DE SVM CON CUATRO VECTORES SOPORTE EN EL ESPACIO DE CARACTERÍSTICAS. FUENTE: PARK ET AL. (118).

En este trabajo, se ha implementado una SVM con un kernel de base radial basado en la función *svm* proporcionada en la librería *e1071* de R (119). Los hiperparámetros C (para controlar el error de clasificación) y γ (para controlar la curvatura de la superficie de decisión) se optimizaron siguiendo un enfoque de búsqueda en rejilla con valores que van de 0.1 a 1000 y de 0.001 a 1000, respectivamente.

2.8.2. k-ésimo vecino más cercano (kNN)

El método kNN fue propuesto originalmente por Fix y Hodges en 1951 (113) como un método de clasificación supervisado no paramétrico, y su funcionamiento ha sido adaptado para resolver problemas de regresión también (120). En el aprendizaje supervisado, el algoritmo calcula una medida de distancia entre la muestra que se quiere clasificar y las k muestras del conjunto de entrenamiento más cercanas, y clasifica la muestra nueva con la clase de las muestras más prevalentes en ese vecindario.

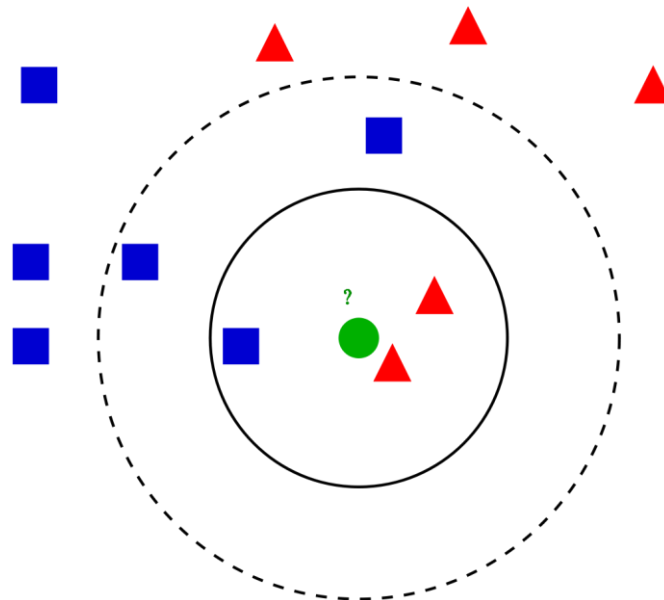


FIGURA 7. ILUSTRACIÓN DEL ALGORITMO kNN. LA NUEVA MUESTRA ES EL PUNTO VERDE. PARA $k=3$ VECINOS, SE CONSIDERA QUE LA NUEVA MUESTRA PERTENECE A LA CLASE TRIANGULAR, MIENTRAS QUE PARA $k=5$ VECINOS, EL ALGORITMO LA CLASIFICA EN LA CLASE CUADRADA. FUENTE: WIKIPEDIA (121).

En este trabajo, se utilizó una implementación del algoritmo kNN con un kernel radial disponible en la biblioteca *caret* de R (122). El hiperparámetro k , que representa el número de vecinos considerados para clasificar las nuevas muestras, se optimizó siguiendo un enfoque de búsqueda en rejilla con valores que van desde 5 a 23 vecinos.

2.8.3. Red neuronal artificial (ANN)

Las ANN se concibieron por primera vez en 1949 (114) como un modelo computacional inspirado en un modelo biológico capaz de resolver tareas de regresión, clasificación, reconocimiento de patrones y reconocimiento del habla, entre otras, usando para ello enfoques de aprendizaje tanto supervisado, como no supervisado o reforzado. Se han diseñado y desarrollado muchos tipos de redes neuronales artificiales, pero todas ellas pueden describirse mediante las funciones de transferencia entre sus neuronas, su regla de aprendizaje y la fórmula de conexión (123). Una ANN está formada por varias unidades (neuronas) conectadas a una o varias entradas (pesos) y organizadas en capas. El funcionamiento básico consiste en que cada neurona procesa las entradas recibidas y, dependiendo de su función de transferencia, produce una salida que se envía a la siguiente capa. El número de neuronas, la función de transferencia y los pesos pueden adaptarse a una salida deseada en los problemas de clasificación durante la etapa de entrenamiento del algoritmo.

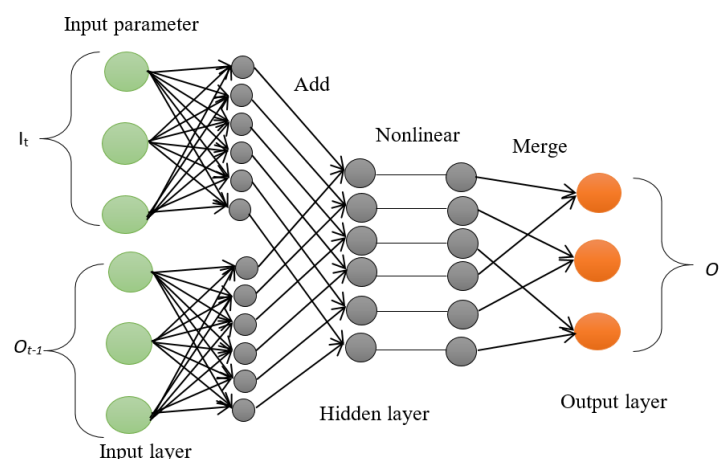


FIGURA 8. REPRESENTACIÓN DE UNA RED NEURONAL ARTIFICIAL RECURRENTE CON RETROALIMENTACIÓN HACIA ADELANTE. FUENTE: TOBORE ET AL. (124).

En este trabajo, hemos utilizado la biblioteca *nnet* de R (125) para implementar una ANN con retroalimentación hacia adelante con una sola capa oculta. Los hiperparámetros considerados fueron *size*, que representa el número de neuronas que constituye la capa oculta, y *decay*, para controlar el factor de decaimiento de los pesos en las iteraciones durante la fase de aprendizaje. Los rangos considerados para la búsqueda en rejilla de los valores óptimos de estos hiperparámetros fueron, hemos seguido un enfoque de búsqueda en cuadrícula con valores que van de 2 a 232 (la mitad de las 464 variables independientes presentes en nuestro conjunto de datos) y de 0.0001 a 0.1, respectivamente.

2.8.4. Modelo Lineal Generalizado (GLM)

El GLM es una técnica avanzada de modelización estadística propuesta por primera vez por Nelder y Wedderburn en 1972 (115) que puede entenderse como una generalización flexible del método de regresión lineal, ya que permite describir la variable de salida con una función de distribución de error distinta de la Gaussiana. Puede utilizarse tanto para tareas de clasificación como de regresión. El GLM consta de tres elementos principales: 1) una familia exponencial de distribuciones de probabilidad para describir las posibles distribuciones de error de la variable de salida; 2) un predictor lineal para incorporar la información de las variables independientes al modelo; y 3) una función de enlace para relacionar los predictores lineales con la media de la familia de las distribuciones de probabilidad. En las tareas de clasificación, el método GLM calcula la probabilidad de que una nueva muestra pertenezca a una u otra clase basándose en las estimaciones de máxima verosimilitud.

El GLM implementado en este trabajo fue instanciado desde la librería *stats* de R (126), haciendo uso de una función de distribución binomial y la función de enlace *logit*, ya que se trataba de una tarea de clasificación binaria.

2.8.5. Nàive Bayes (NB)

NB es una familia de clasificadores probabilísticos supervisados que se basan en el teorema de Bayes y en la suposición de que las variables de entrada son independientes entre sí (127). Los clasificadores NB generan una función de probabilidad asociada a cada clase basada en los descriptores de la muestra (características o variables independientes), y para cada nueva muestra, evalúa la probabilidad de que dicha muestra pertenezca a una clase determinada. A continuación, se aplica una regla de decisión, que suele basarse en el método de máxima verosimilitud, para clasificar las nuevas muestras.

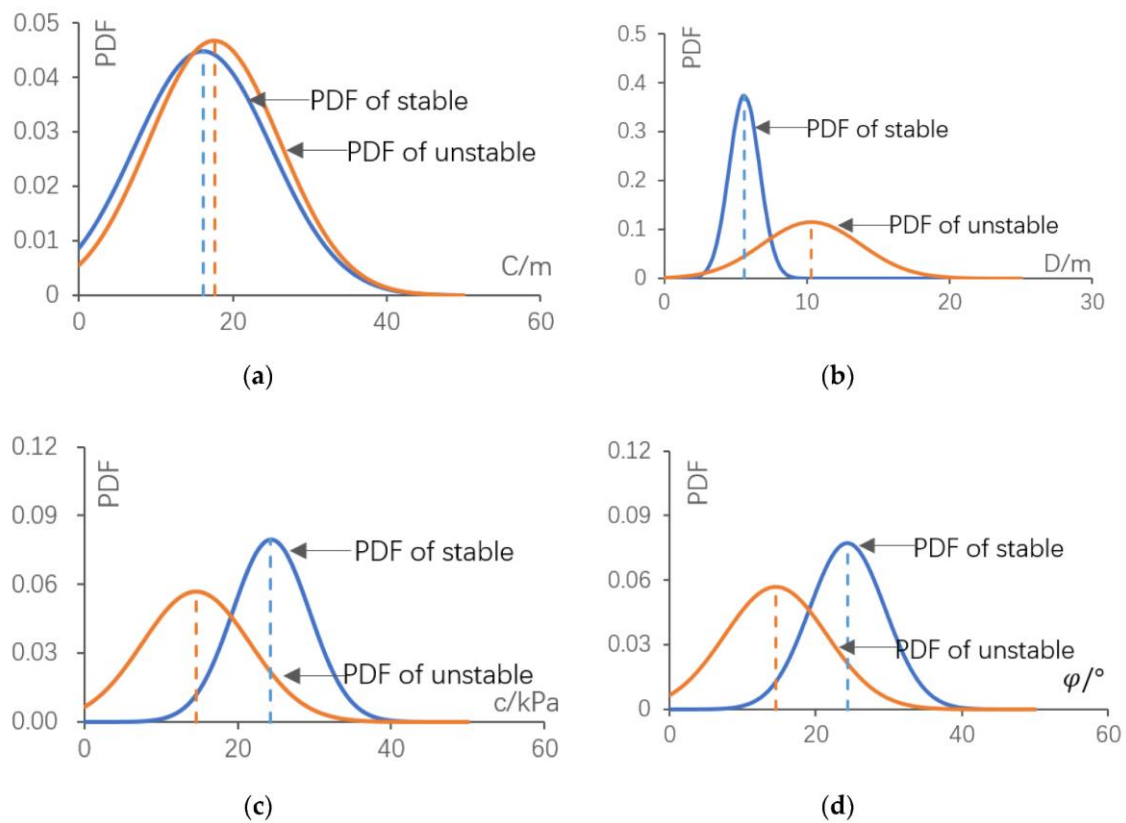


FIGURA 9. FUNCIONES DE DENSIDAD DE PROBABILIDAD DE CUATRO CASOS DE USO DE UN CLASIFICADOR NB UTILIZADO PARA DETERMINAR LA ESTABILIDAD DE LOS TÚNELES. FUENTE: LI ET AL. (128).

En nuestro caso, utilizamos el clasificador NB incluido en la biblioteca *e1071* de R (119) con la configuración por defecto, es decir, sin el suavizado laplaciano.

2.9. Validación interna

Para evitar los problemas de sobreajuste en el entrenamiento de un modelo predictivo, es necesario dejar fuera una parte del conjunto de datos con fines de validación interna, de modo que no utilicemos los mismos datos durante las fases de entrenamiento y validación posterior del modelo.

Esto se suele conseguir siguiendo una estrategia de validación cruzada *leave-one-out* (LOO), que puede hacerse más robusta repitiendo el proceso de LOO varias veces con diferentes particiones del conjunto de datos. En este caso, el proceso se denomina validación cruzada de k particiones, siendo k el número de veces que se repite el proceso y, también, el número de particiones generadas a partir del conjunto de datos original (106).

Un valor de k comúnmente aceptado, siempre que el tamaño del conjunto de datos lo permita, es 10, ya que proporciona una estimación aproximada del error medio de predicción del modelo entrenado (107). Sin embargo, se han realizado algunos estudios sobre métodos de optimización aplicados al cálculo del valor óptimo de k en problemas de clasificación (108).

En este trabajo, para el proceso de validación interna se ha realizado una validación cruzada de 10 particiones para cada uno de los modelos de predicción desarrollados haciendo uso de los subconjuntos de entrenamiento representados en la Tabla 2.

2.10. Validación externa y medidas de desempeño

Según la guía de Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) (62), es muy recomendable evaluar el desempeño de los modelos predictivos con un conjunto de datos que no se haya utilizado para el desarrollo del modelo. Según como se genere este conjunto de datos, se puede realizar una validación temporal (cuando los casos se recogen siguiendo los mismos procedimientos que los originales, pero en una fase posterior), una validación geográfica (cuando los datos

proceden de diferentes entornos clínicos y/o utilizando diferentes procedimientos y mediciones para recoger la información), e incluso incluyendo otros tipos de pacientes.

En nuestro caso, dado que nuestro conjunto de datos incluye información procedente de un único entorno clínico, decidimos realizar una validación externa siguiendo un enfoque temporal. Como se muestra en la Tabla 2, los conjuntos de datos de entrenamiento se generaron con aquellos casos registrados antes de una fecha determinada (31 de mayo de 2018) de una forma balanceada, es decir, se consideraron el mismo número de casos con toxicidades positivas y negativas en cada conjunto de datos de entrenamiento. Los conjuntos de datos para la validación externa se generaron con aquellos casos registrados después de esa fecha, y no se aplicó ningún tipo de balanceo en este caso.

El desempeño de un modelo de predicción se suele informar en términos de calibración y discriminación. La calibración es la concordancia entre la predicción y la observación real, y la discriminación es la capacidad del modelo para distinguir entre los casos que tuvieron un resultado positivo o negativo (139). Por lo general, la discriminación se evalúa mediante el área bajo la curva característica operativa del receptor (AUROC o AUC), también conocida como concordancia o estadístico C, y la calibración suele evaluarse mediante una curva de calibración (140).

En este trabajo, la sección de resultados incluye el AUC y las curvas de calibración obtenidos por las mejores combinaciones de métodos de selección de características y clasificadores basados en algoritmos aprendizaje automático por cada toxicidad analizada. También se incluye el AUC promedio obtenido por cada uno de los métodos de selección de características y por cada uno de los clasificadores empleados, con la idea de comparar su desempeño. En el anexo se incluye el AUC obtenido por los 300 modelos analizados tanto en la etapa de validación cruzada como en la de validación externa.

Todos los análisis se ejecutaron en un MacBook Pro de 15", de finales de 2011 (Apple, Inc. Cupertino, California, USA) usando el software R versión 4.1.3 y RStudio versión 1.3.1093.

3. Resultados

3.1. AUC y curvas de calibración obtenidas para cada toxicidad

En la Tabla 5 se presentan los resultados obtenidos para cada una de las toxicidades bajo estudio por las mejores combinaciones de métodos de selección de características y clasificadores basados en algoritmos de aprendizaje automático en términos de AUC. Además, se incluyen las dimensiones clínicas utilizadas por cada modelo de predicción desarrollado, que posteriormente analizaremos en la discusión de los resultados.

Además, en la Figura 10 se muestran las curvas de calibración correspondientes a estos modelos de predicción.

TABLA 5. MEJORES MODELOS SEGÚN AUC DURANTE VALIDACIÓN INTERNA PARA CADA TOXICIDAD BAJO ESTUDIO.

Toxicidad	Mejor modelo	AUC*	AUC**	Características	Dimensiones clínicas
Esofagitis aguda	mRMR + GLM	0.85	0.81	69	Edad, nivel socio-económico, VIH, etnia, tabaquismo, síntoma primario, anorexia, pérdida de peso, KPS, altura, localización del tumor, histología, EGFR, ALK, TNM, creatinina, hematocritos, historia de cáncer familiar, QoL, QT concurrente, dosimetría (pulmón, esófago, corazón, GTV, CTV)
Tos aguda	IG + ANN	0.90	0.77	13	Nivel socio-económico, QoL, QT concurrente, dosimetría (pulmón, esófago, corazón, GTV, CTV)
Disnea aguda	mRMR + GLM	0.81	0.57	32	Nivel socio-económico, EPOC, oxigenoterapia, síntoma primario, anorexia, KPS, altura, histología, ALK, TNM, PFT, historia de cáncer familiar, QoL, dosimetría (pulmón, esófago, GTV)
Neumonitis aguda	χ^2 + NB	0.81	0.85	24	Nivel socio-económico, disnea, tos, histología, TNM, QoL, dosimetría (CTV, GTV)
Disnea crónica	mRMR + GLM	0.87	0.97	19	Nivel socio-económico, síntoma primario, disfagia, PET, TNM, ALK, historia de cáncer familiar, QoL, GTV
Neumonitis crónica	mRMR + ANN	0.90	0.73	12	Nivel socio-económico, síntoma primario, disnea, dolor pleurítico, PET, localización del tumor, ALK, TNM, PFT, historia de cáncer familiar, QoL, dosimetría (CTV, GTV, corazón)

* AUC obtenida durante la validación interna

** AUC obtenida durante la validación externa

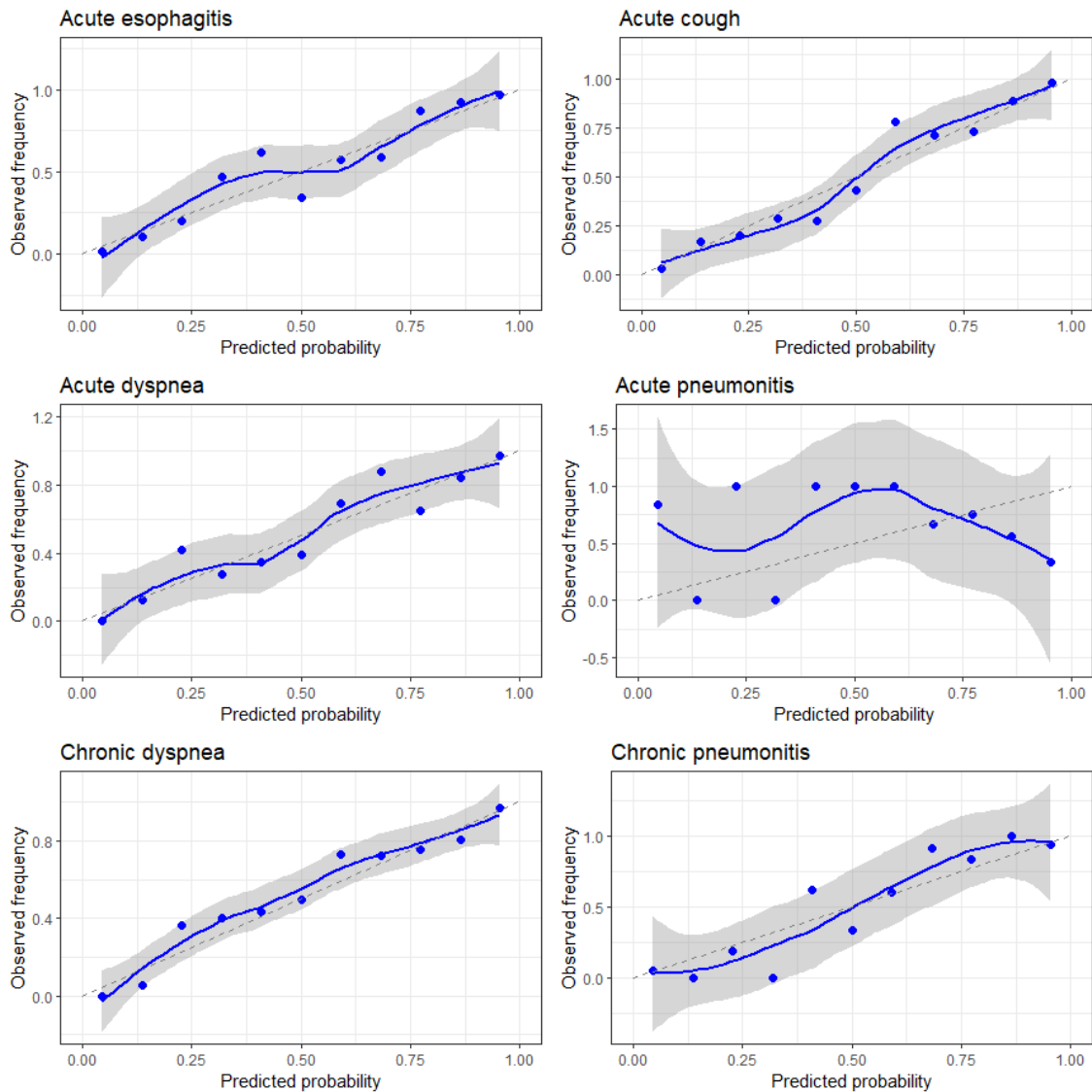


FIGURA 10. CURVAS DE CALIBRACIÓN CORRESPONDIENTES A LOS MEJORES MODELOS DE PREDICCIÓN OBTENIDOS PARA CADA TOXICIDAD.

3.2. AUC según método de selección de características

En esta sección se resumen los resultados obtenidos, en términos de AUC, por cada uno de los métodos de selección de características empleados, sin tener en cuenta ni la toxicidad objetivo ni el algoritmo de clasificación basado en técnicas de aprendizaje automático con el que se ha usado. La idea es poder comparar el rendimiento de cada uno de los métodos de selección de características estudiados con el *gold-standard*, que es la AUC obtenida tras utilizar las características seleccionadas por el experto oncólogo. Esta comparación se realiza tanto a nivel de validación interna como de validación externa.

TABLA 6. AUC PROMEDIO Y DESVIACIÓN ESTÁNDAR (ENTRE PARÉNTESIS) DURANTE LAS FASES DE VALIDACIÓN INTERNA Y VALIDACIÓN EXTERNA PARA CADA UNO DE LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS ESTUDIADOS. SE INCLUYE COMPARACIÓN CON EL GOLD STANDARD.

Método	AUC en validación interna	AUC en validación externa
CFS	0.71 (0.11) ***	0.66 (0.13)
χ^2	0.73 (0.10) ***	0.69 (0.10) **
Boruta	0.74 (0.11) ***	0.69 (0.10) *
Votación (subconjunto)	0.73 (0.10) ***	0.70 (0.12) **
mRMR	0.74 (0.11) ***	0.67 (0.15)
Relief	0.65 (0.05) ***	0.66 (0.14)
RF	0.74 (0.09) ***	0.66 (0.13)
IG	0.72 (0.10) ***	0.67 (0.11)
Votación (clasificación)	0.74 (0.10) ***	0.64 (0.13)
Gold standard	0.59 (0.04)	0.62 (0.10)

* p -value < 0.05 (95% IC)

** p -value < 0.01 (95% IC)

*** p -value < 0.001 (95% IC)

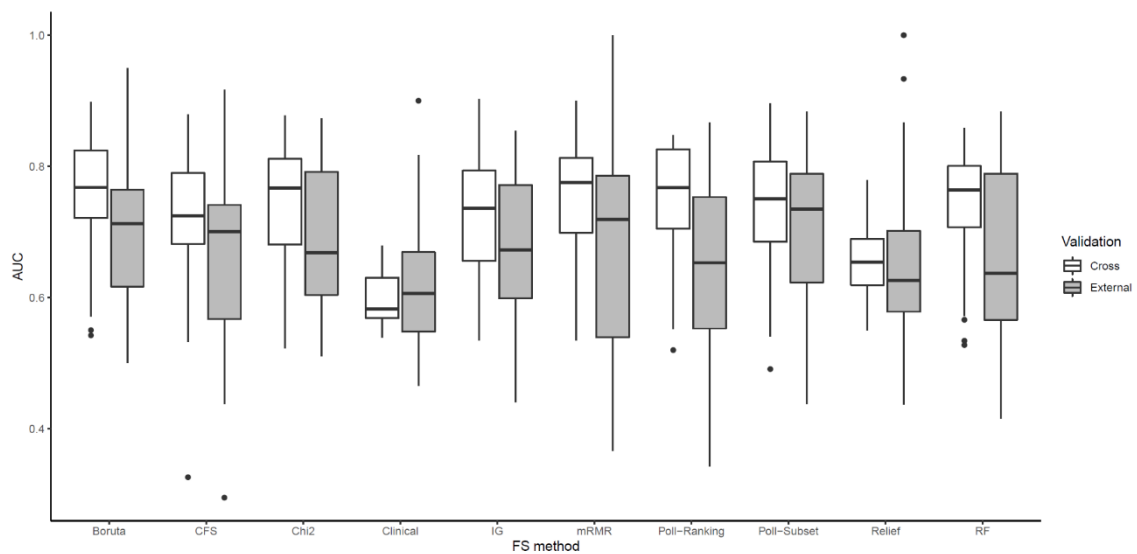


FIGURA 11. DIAGRAMAS DE CAJA (BOXPLOT) DE LAS AUCs OBTENIDAS POR CADA MÉTODO DE SELECCIÓN DE CARACTERÍSTICAS EN LAS ETAPAS DE VALIDACIÓN INTERNA (FONDO BLANCO) Y VALIDACIÓN EXTERNA (FONDO GRIS). DE IZQUIERDA A DERECHA, LOS MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS SON: BORUTA, CFS, CHI CUADRADO, GOLD STANDARD, IG, mRMR, VOTACIÓN (CLASIFICACIÓN), VOTACIÓN (SUBCONJUNTOS), RELIEF Y RF.

3.3. AUC según técnica de aprendizaje automático

En esta sección se resumen los resultados obtenidos, en términos de AUC, por cada uno de los algoritmos de clasificación basados en técnicas de aprendizaje automático, sin tener en cuenta ni los métodos de selección de características empleados ni la toxicidad objetivo. La idea es poder comparar el rendimiento de cada uno de los clasificadores entre sí. Esta comparación se realiza tanto a nivel de validación interna como de validación externa.

TABLA 7. AUC PROMEDIO Y DESVIACIÓN ESTÁNDAR (ENTRE PARÉNTESIS) OBTENIDOS DURANTE LAS FASES DE VALIDACIÓN INTERNA Y VALIDACIÓN EXTERNA PARA CADA UNO DE LOS ALGORITMOS DE CLASIFICACIÓN EMPLEADOS.

Método	AUC en validación interna	AUC en validación externa
kNN	0.57 (0.03)	0.58 (0.13)
GLM	0.77 (0.08)	0.72 (0.10)
NB	0.74 (0.09)	0.67 (0.14)
SVM	0.71 (0.07)	0.68 (0.10)
ANN	0.76 (0.08)	0.68 (0.10)

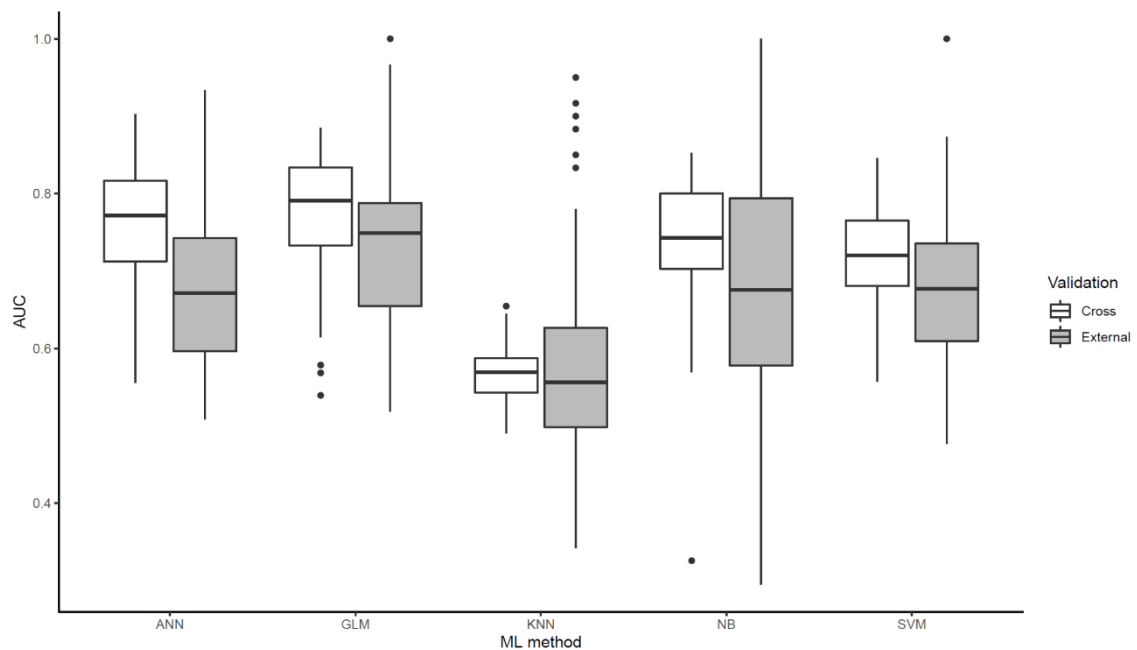


FIGURA 12. DIAGRAMAS DE CAJA (BOXPLOT) DE LAS AUCS OBTENIDAS POR CADA ALGORITMO DE CLASIFICACIÓN EN LAS ETAPAS DE VALIDACIÓN INTERNA (FONDO BLANCO) Y VALIDACIÓN EXTERNA (FONDO GRIS). DE IZQUIERDA A DERECHA: ANN, GLM, kNN, NB Y SVM.

4. Discusión

4.1. AUC y curvas de calibración obtenidas para cada toxicidad

A continuación, analizamos los resultados obtenidos en términos de AUC por los modelos de predicción basados en técnicas de aprendizaje automático para cada toxicidad bajo estudio, y finalizamos la sección con el análisis de las curvas de calibración.

En términos generales, un hallazgo interesante de este trabajo es que, según las variables seleccionadas por los mejores modelos de predicción en términos de AUC, es que hay una serie de variables clínicas que son comunes a todos los modelos de predicción inducida por RT, que son: el nivel socioeconómico, la calidad de vida basal, el GTV y el TNM. Además, las variables relacionadas con la expresión de ALK, el CTV, el síntoma primario y los antecedentes familiares de cáncer estuvieron presentes en 4 de los 6 modelos predictivos.

En cuanto al rendimiento y precisión alcanzados por los modelos seleccionados siguiendo la metodología propuesta, también es destacable que todos ellos están en línea con los modelos predictivos actuales del estado del arte, y algunos de ellos incluso los superan.

4.1.1. Esofagitis aguda

En el caso de la esofagitis aguda, el modelo de mejor rendimiento alcanzó un AUC=0.85 durante la fase de validación interna y un AUC=0.81 tras probarlo con la muestra de validación externa.

Este resultado aparentemente mejora otros métodos del estado del arte (141,142) y potencialmente resuelve el problema señalado por las conclusiones de investigaciones anteriores sobre métodos de aprendizaje automático aplicados a la predicción de esta toxicidad que sugieren que estas técnicas aún no están preparados para lograr una buena precisión de predicción para la esofagitis aguda (143).

El modelo se construyó utilizando el método de selección de características mRMR seguido por el clasificador GLM, considerando un total de 69 variables que describen el perfil sociodemográfico del paciente, las afecciones previas, los síntomas primarios, el KPS (índice de estado funcional de Karnofsky), la caracterización del tumor (localización, histología, TNM), el perfil molecular (expresión genética de EGFR y ALK), los antecedentes familiares de cáncer, la calidad de vida, el régimen de tratamiento y la dosimetría en pulmón, esófago, y corazón, además de valores de planificación de tratamiento como GTV (volumen tumoral bruto) y CTV (volumen objetivo clínico).

Algunas de estas variables ya se habían discutido en la literatura como posibles factores predictivos de la esofagitis aguda inducida por la RT (144–146), lo que demuestra que la metodología seguida consiguió reflejar este conocimiento en el modelo desarrollado. Además, el modelo también destacó otras variables relacionadas con los niveles de creatinina y hematocrito que, hasta donde sabemos, aún no se han relacionado con la aparición de esofagitis aguda inducida por la RT, abriendo la puerta por tanto al establecimiento de nuevas hipótesis de investigación basadas exclusivamente en métodos de aprendizaje automático a partir de los datos obtenidos en la rutina asistencial.

4.1.2. Tos aguda

Respecto a la predicción de la tos aguda inducida por la RT, el modelo de mejor rendimiento alcanzó un AUC=0.90 en la etapa de validación interna, y un AUC=0.77 en la muestra de validación externa.

El modelo, basado en el método de selección de características IG y el clasificador ANN, utilizó 13 variables que describen el nivel socioeconómico del paciente, la calidad de vida y la dosis de RT en los pulmones, el esófago y el corazón.

Esta toxicidad se considera generalmente un efecto secundario de la neumonitis por radiación y se tienen en cuenta como criterio de clasificación para las escalas de gravedad de la toxicidad inducida por la RT, como la RTOG y la EORTC (147) y, hasta donde sabemos, no hay trabajos previos sobre validaciones de modelos predictivos para la tos aguda inducida por la RT, a pesar de su reconocida importancia y contribución a la CVRS de los pacientes (148,149).

4.1.3. Disnea aguda

En cuanto a la disnea aguda, el modelo predictivo con mejor rendimiento en la fase de validación interna alcanzó un AUC=0.81 y, con las muestras de validación externa, un AUC=0.57. En comparación con otros modelos predictivos de esta toxicidad reportados en la literatura (28,150,151), el modelo desarrollado parece superar a la mayoría de ellos en la fase de validación interna. Sin embargo, su bajo rendimiento con la muestra de validación externa debe ser objeto de una revisión crítica para evaluar si se debe a una potencial falta de generalización del modelo o a un factor de sesgo diferente, tal vez relacionado con el escaso número de muestras utilizadas durante la validación externa.

Este modelo se construyó utilizando el método de selección de características mRMR junto con el clasificador GLM, e incluyó 32 variables relacionadas con el nivel socioeconómico, enfermedades basales, el síntoma primario, el estado funcional, la caracterización del tumor, la expresión de ALK, los resultados de las pruebas de función pulmonar, los antecedentes familiares de cáncer, la calidad de vida y la dosis de RT en los pulmones y el esófago.

Estas variables clínicas ya han sido reportadas en la literatura como posibles predictores de la disnea aguda inducida por la RT (152–155) por lo que, a pesar del bajo rendimiento alcanzado durante la validación externa, el modelo fue capaz de representar con fidelidad la evidencia conocida sobre los factores que contribuyen a esta toxicidad.

4.1.4. Neumonitis aguda

En cuanto a la neumonitis aguda inducida por RT, el modelo de predicción de mejor rendimiento alcanzó un AUC=0.81 en la validación interna, y un AUC=0.85 con la muestra de validación externa. En términos de rendimiento, el modelo desarrollado logra resultados comparables a los modelos más recientes del estado del arte (75,156,157).

Este modelo se construyó a partir del método de selección de características Chi-cuadrado seguido de un clasificador Naïve Bayes, teniendo en cuenta 24 variables que describían el nivel socioeconómico del paciente, las afecciones previas (disnea, tos), la caracterización del tumor (histología y TNM), la QoL y la dosis de RT (CTV y GTV).

Todos estos potenciales predictores de la neumonitis aguda ya han sido discutidos en la literatura con diferentes niveles de acuerdo (158–160), por lo que el modelo parece ajustarse de forma fiable a la evidencia actual sobre los factores que contribuyen a la neumonitis aguda inducida por la RT.

4.1.5. Disnea crónica

Respecto a la predicción de la disnea crónica inducida por la RT, el modelo de mejor rendimiento alcanzó un AUC=0.87 durante la etapa de validación interna, y un AUC=0.97 en la muestra de validación externa. Este resultado supera a todos los modelos predictivos existentes en el estado del arte para la disnea inducida por la RT (28,153,161).

Este modelo se construyó a partir de un método un método de selección de características basado en la técnica mRMR seguido de un clasificador GLM entrenado con 19 variables que describían el nivel socioeconómico del paciente, los síntomas primarios, las afecciones previas (disfagia), registro previo de PET, el TNM del tumor, la expresión de ALK, los antecedentes de cáncer familiar, la calidad de vida basal y el GTV.

Algunas de estas variables clínicas ya han sido mencionadas en la literatura como potenciales predictores de la disnea crónica inducida por la RT (153,162,163). Sin embargo, no hemos podido encontrar ninguna mención sobre la expresión de ALK y su relación con esta toxicidad específica en la literatura revisada.

4.1.6. Neumonitis crónica

Por último, en lo que respecta a la predicción de la neumonitis crónica inducida por RT, el modelo con mejor rendimiento alcanzó un AUC=0.90 en la etapa de validación interna, mientras que su rendimiento durante la validación externa fue de un AUC=0.73, resultado que parece estar en consonancia con los métodos de predicción de esta toxicidad más recientes (75,156,157).

Este modelo se basó en el método de selección de características mRMR seguido de un clasificador basado en una ANN, entrenada con 32 variables que describían los siguientes conceptos clínicos: nivel socioeconómico, síntoma primario, disnea previa, dolor pleurítico, registro previo de PET y de pruebas de función pulmonar, características del tumor (localización y TNM), expresión de ALK, antecedentes familiares de cáncer, QoL y dosis de RT en el corazón, CTV y GTV.

La mayoría de estos posibles factores que contribuyen a la aparición de la neumonitis crónica inducida por la RT ya se han discutido en la literatura (158,164–167), lo que puede considerarse como una señal de que el modelo fue capaz de capturar con fidelidad estas relaciones ya conocidas.

4.1.7. Curvas de calibración

Según los gráficos de calibración mostrados en la Figura 10, 5 de los 6 modelos desarrollados representan fielmente una relación casi lineal entre las probabilidades predichas y las frecuencias observadas en los casos positivos, lo que representa que el modelo estaría correctamente calibrado.

En cuanto al modelo de neumonitis aguda, el gráfico de calibración no se ajusta correctamente a esta relación lineal necesaria para representar una correcta calibración, lo que ilustra un problema bien conocido sobre la incapacidad de los clasificadores Nãive Bayes para producir estimaciones de probabilidad correctamente calibradas (168). Este inconveniente podría mitigarse utilizando modelos de calibración en una fase posterior al desarrollo del modelo (169).

4.2. AUC según método de selección de características

Como se muestra en la Tabla 6, se ha llevado a cabo un análisis del rendimiento de los 10 métodos de selección de características implementados en términos de AUC, tanto en la etapa de validación interna como en la de validación externa, y sin tener en cuenta las toxicidades objetivo de la predicción ni el tipo de clasificadores basados en técnicas de aprendizaje automático empleados.

El método de selección de características considerado como *gold standard* fue la selección por parte de un oncólogo experto en RT de aquellas variables que la evidencia actual y las guías de práctica clínica recomiendan tener en cuenta a la hora de predecir las posibles toxicidades derivadas de la exposición al tratamiento de RT. En este sentido, para la toxicidad relacionada con la esofagitis, el clínico seleccionó las variables que representaban las siguientes características clínicas: dosis de RT en esófago (dosis media, V35, V50 y V70), afecciones digestivas previas, edad, estado funcional basal del paciente, hábito tabáquico, TNM y plan de tratamiento. Para el resto de las toxicidades, el clínico seleccionó las variables que representaban las siguientes características clínicas: dosis de RT en pulmón (dosis media y V20), edad, EPOC previa, oxigenoterapia previa, hábito tabáquico, estado funcional basal del paciente, TNM y plan de tratamiento.

Los otros nueve métodos de selección de características analizados se compararon con el clínico para encontrar posibles diferencias estadísticamente significativas entre ellos. En la fase de validación interna, el método de selección de características *gold standard* alcanzó una media de $AUC=0.59$ (desviación estándar 0.04), mientras que los otros 9 métodos selección de características oscilaron entre una media de $AUC=0.65$ (0.05) obtenida por el

método Relief y una media de AUC=0.74 (0.09) obtenida por el método RF. Esta media de AUC también fue alcanzada por los métodos de votación (clasificación), Boruta y mRMR, con una desviación estándar ligeramente superior (0.10, 0.11 y 0.11, respectivamente). Cabe destacar que, en esta fase, todos los métodos de selección de características analizados lograron mejoras estadísticamente significativas en su rendimiento en comparación con el *gold standard* ($p < 0,001$).

En la fase de validación externa, el *gold standard* obtuvo un AUC medio de 0.62 (0.10), mientras que los demás métodos de selección de características oscilaron entre un AUC medio de 0.64 (0.13), alcanzado por el método de selección de características de votación (clasificación), y un AUC medio de 0.70 (0.12), alcanzado por el método de selección de características basado en votación (subconjunto). En este caso, sólo tres métodos de selección de características (Chi-cuadrado, Boruta y votación entre los métodos de subconjunto) lograron una mejora estadísticamente significativa de su rendimiento en comparación con el *gold standard*.

En general, el rendimiento medio alcanzado por todos los métodos de selección de características en la fase de validación externa fue considerablemente inferior al obtenido durante la validación interna, lo que coincide con la evidencia actual (170).

4.3. AUC según método de predicción basado en técnicas de aprendizaje automático

Como se indica en la Tabla 7, se ha llevado a cabo el análisis del rendimiento en términos de AUC promedio y desviación estándar de los clasificadores basados en técnicas de aprendizaje automático implementados sin tener en cuenta las toxicidades objetivo ni el método de selección de características implementado.

En la etapa de validación interna, el rendimiento de los clasificadores osciló entre un AUC medio de 0.57 (0.03) logrado por el clasificador kNN y un AUC medio de 0.77 (0.08) logrado por el clasificador GLM. Cabe destacar que los otros tres clasificadores analizados (NB, SVM y ANN) obtuvieron un rendimiento medio superior a 0.70.

En la fase de validación externa, el rendimiento de los clasificadores osciló entre un AUC medio de 0.58 (0.13) obtenido por el clasificador kNN y un AUC medio de 0.72 (0.10) obtenido por el clasificador GLM. En esta etapa, los clasificadores NB, SVM y ANN obtuvieron un rendimiento medio similar (0.67, 0.68 y 0.68, respectivamente).

Estos resultados informan de manera consistente sobre el pobre rendimiento del clasificador kNN para predecir las toxicidades inducidas por la RT en general cuando se utiliza para su entrenamiento un conjunto de datos obtenidos en un entorno asistencial que podría incorporar un alto nivel de variables ruidosas (171). El clasificador GLM logró el AUC promedio más alto en ambas etapas (validación interna y validación externa), pero seguido de cerca por los clasificadores NB, SVM y ANN en la etapa de validación externa, lo que parece indicar que estos cuatro clasificadores podrían adaptarse mejor que el kNN a un entorno de aprendizaje automático basado en datos obtenidos en la rutina asistencial.

4.4. Cumplimiento de los objetivos planteados

El objetivo principal de esta investigación era la de validar una metodología clara y generalizable para la implementación y validación externa de los modelos predictivos de toxicidad asociada a la RT basados en métodos de selección de características y algoritmos de aprendizaje automático que fueran entrenados con conjuntos de datos obtenidos en el entorno clínico-asistencial para facilitar el avance y adopción de esta tecnología de apoyo a la toma de decisiones clínicas en la rutina asistencial durante el manejo de los pacientes con cáncer de pulmón.

La metodología utilizada, ilustrada en la Figura 5 y desarrollada en consonancia con las recomendaciones de la declaración TRIPOD (62), incluía los elementos básicos y los pasos a seguir para conseguir este objetivo, desde la recogida del dato en el entorno asistencial, la formulación de la pregunta o hipótesis clínica, la identificación de las variables objetivos, el tratamiento del conjunto de datos (normalización, curación e imputación de valores), el proceso de validación interna, entrenamiento del modelo de predicción, selección de características relevantes, la validación externa y, por último, el descubrimiento de nuevas evidencias o indicios que permitan la elaboración de nuevas hipótesis de investigación.

Para validar la aplicabilidad de esta metodología, se desarrollaron un total de 300 casos de uso diferentes, basados en todas las combinaciones posibles entre 6 toxicidades objetivo, 10 métodos de selección de características y 5 clasificadores basados en técnicas de aprendizaje automático haciendo uso de un conjunto de datos de 875 pacientes con cáncer de pulmón cuya información se recogió en un contexto clínico-asistencial.

Los resultados obtenidos tras la aplicación de esta metodología resaltan su capacidad para lograr, en la mayoría de los casos, resultados comparables a otros enfoques similares desarrollados en el estado del arte para predecir la aparición de toxicidades inducidas por la RT en esta población. Además, este trabajo demuestra el potencial para escalar esta metodología a un mayor número de métodos de selección de características, clasificadores basados en técnicas de aprendizaje automático y, potencialmente, otros conjuntos de datos recogidos en entorno asistencial.

Respecto al segundo objetivo planteado, los mejores modelos de predicción en términos de AUC han empleado una serie de variables clínicas comunes a todas las variables objetivo que se han estudiado, que son: el nivel socioeconómico, la calidad de vida basal, el GTV y el TNM. Además, las variables relacionadas con el CTV, el síntoma primario y los antecedentes familiares de cáncer estuvieron presentes en 4 de los 6 modelos predictivos. En la literatura se ha descrito el potencial predictor de estas variables, entre otras, a través de estudios realizados en entornos controlados, tal y como se ha destacado en la introducción de este trabajo. Este resultado nos permite, por tanto, validar parte de este conocimiento a través del análisis de los datos procedentes de la rutina asistencial.

Por otra parte, y como resultado de la experimentación llevada a cabo, los modelos desarrollados con mejor rendimiento han puesto de manifiesto algunas cuestiones interesantes acerca del papel de ciertas variables clínicas que no suelen ser tenidas en consideración durante la gestión de las posibles toxicidades inducidas por RT que un paciente de CP puede sufrir tras el tratamiento. En este sentido, esta investigación plantea la existencia de alguna relación entre los niveles basales de creatinina y hematocritos del paciente y la aparición de esofagitis aguda, o la relación entre la expresión de ALK y la aparición de disnea crónica tras la RT.

Más allá de las implicaciones que el descubrimiento de estas potenciales relaciones entre variables clínicas y toxicidades pudiera tener en futuras investigaciones clínicas, el mero hecho de que la metodología presentada haya demostrado tener capacidad de generar nuevas hipótesis de investigación basadas en datos recogidos en el entorno asistencial confirma su validez y aplicabilidad en este ámbito.

4.5. Limitaciones

Esta investigación tiene varias limitaciones que deben ser consideradas cuidadosamente antes de generalizar las conclusiones presentadas. En cuanto al conjunto de datos obtenidos en el entorno asistencial utilizado, este trabajo no ha considerado en profundidad la información sobre el estilo de vida en el sentido amplio, es decir, incluyendo valoraciones sobre el ejercicio físico, higiene del sueño, alimentación y bienestar emocional del paciente, elementos que han demostrado tener el potencial de contribuir a la predicción de las toxicidades inducidas por la RT (172). Tampoco se han tenido en cuenta otras cuestiones habitualmente poco reconocidas que afectan a la calidad de vida relacionada con la salud en los pacientes con CP durante y después de la RT, como la angustia psicológica, el estigma, las disparidades sanitarias, el deterioro cognitivo y la disfunción sexual (173). Esta información no se ha incluido debido a que su registro no forma parte de la rutina asistencial a estos pacientes y, por tanto, no se encuentra disponible para su análisis.

Los avances tecnológicos están promoviendo un uso creciente de dispositivos portátiles para monitorizar diferentes aspectos de la evolución de los pacientes con cáncer entre las visitas de seguimiento (174). La información recopilada por estos dispositivos añadiría un conjunto adicional de variables que podrían ser útiles para predecir eventos de salud como son las toxicidades inducidas por la RT. Lo mismo ocurre con el uso de escalas de evaluación de los resultados en salud comunicados por los pacientes (PROM) en los entornos oncológicos, que ya han demostrado su capacidad para predecir eventos de salud en pacientes con CP (175). Además, la información derivada de características cuantificables obtenidas a partir de imágenes médicas (radiómica) tampoco se incluyó entre las variables de análisis a pesar de su reconocido potencial para predecir los resultados de salud tras la RT en pacientes con CP (176). La decisión de no incluir estos conjuntos de datos derivó de la falta de

dicha información recogida de forma consistente en los sistemas de información hospitalarios durante la rutina asistencial.

La metodología propuesta en esta tesis facilitaría la explotación de estos nuevos conjuntos de datos gracias al proceso de tratamiento de información planteado (normalización, curación e imputación de valores perdidos) y a la automatización del proceso de selección de características. Una línea de trabajo futuro podría ser la exploración de la adición de variables procedentes de estos dominios al proceso de desarrollo del modelo predictivo para así comprobar su contribución a la precisión global del modelo.

A pesar de haber implementado un proceso de validación externa que permitiera verificar la capacidad de generalizar de la metodología propuesta, debe tenerse en cuenta un potencial sesgo de selección dado que todos los pacientes con CP incluidos en el análisis fueron reclutados en un único hospital y, por tanto, no son representativos de los diferentes grupos étnicos que podrían encontrarse en otras geografías. Otra limitación relacionada con el tema de la localización tiene que ver con las opciones terapéuticas disponibles en este hospital, que fueron proporcionadas de acuerdo con las directrices de la NCCN para los pacientes con CP. Por lo tanto, las conclusiones obtenidas en este trabajo no deben extenderse directamente a otros entornos clínicos que proporcionen atención a poblaciones diferentes y/o que sigan un proceso asistencial diferente al recomendado por la NCCN.

La validación experimental de la metodología presentada incluía un conjunto limitado de diez métodos de selección de características y cinco clasificadores basados en algoritmos de aprendizaje automático, y el ajuste de los hiperparámetros de estos clasificadores, siempre que fuera posible, se ha limitado a una estrategia básica de búsqueda en rejilla. Esta estrategia podría haber producido resultados subóptimos en términos de precisión de los clasificadores empleados, ya que ninguno de los 300 modelos predictivos probados fue optimizado manualmente. Esta decisión se tomó para no menoscabar la capacidad de generalización de la metodología propuesta. Trabajos futuros podrían incluir otros métodos de selección de características, así como otras familias de clasificadores basados en algoritmos de aprendizaje automático para comparar su rendimiento con la experimentación llevada a cabo en este trabajo y demostrar la reproducibilidad de la metodología propuesta.

Por último, se debe resaltar que se ha observado una gran variabilidad entre los rendimientos en términos de AUC obtenidos durante la validación interna y la validación externa. Este es un problema bien documentado en la literatura (177,178). Con el fin de proporcionar una estimación más precisa de la precisión de los modelos, otras líneas futuras de trabajo podrían incorporar la inclusión de un mayor tamaño muestral en el conjunto de datos de validación externa.

Para poder implementar los resultados obtenidos en la rutina asistencial, debe tenerse en cuenta una serie de consideraciones desde el punto de vista regulatorio y operativo que den respuesta a las necesidades del proceso asistencial en su conjunto, más allá de la puesta a disposición de profesionales y pacientes de una herramienta que facilite la toma de decisiones compartida sobre las opciones terapéuticas disponibles que tenga en consideración la posibilidad de aparición de efectos secundarios que pueden llegar a reducir de manera significativa la calidad de vida de los pacientes.

La metodología validada en esta tesis permite la generación y validación de modelos de predicción basados en técnicas de aprendizaje automático con la finalidad de dar soporte a la hora de tomar decisiones terapéuticas en el ámbito del cáncer de pulmón. Según la regulación Europea sobre productos sanitarios (Medical Devices Regulation, MDR) (179), en su Anexo VIII, sección 6.3. Regla 11, se especifica que “los programas informáticos destinados a proporcionar información que se utiliza para tomar decisiones con fines terapéuticos o de diagnóstico se clasifican en la clase IIa”. Por tanto, para poder aplicar los modelos de predicción desarrollados en la rutina asistencial, se deberá obtener previamente el marcado CE del producto sanitario resultante conforme a los procedimientos establecidos por la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) como organismo notificado.

También es relevante el actual desarrollo del borrador de la Ley de IA Europea (EU AI Act) (78), en el que se propone una clasificación de los sistemas que hagan uso de elementos de inteligencia artificial en base al riesgo potencial para la salud, seguridad o los derechos fundamentales de los ciudadanos. En este contexto, tal y como se expone en el borrador de la legislación, aquellos sistemas que caigan dentro del ámbito de aplicación de la MDR, serían considerados como sistemas de alto riesgo por la EU AI Act. Esto implica que estos sistemas o

aplicaciones deberán cumplir con unos requisitos específicos relativos a los datos de alta calidad, la documentación y la trazabilidad, la transparencia, la vigilancia humana, la precisión y la solidez para poder garantizar una correcta mitigación de los riesgos derivados por la IA. En la práctica, esto se traduciría en la implantación de un sistema de gestión de riesgos auditable y certificado por parte del proveedor del sistema que abarque aspectos de datos y gobernanza de datos, documentación técnica, registros, transparencia y comunicación de información a los usuarios, vigilancia humana, precisión, solidez y ciberseguridad, tal y como se detalla en el Capítulo 2, art. 10 – 15 del citado borrador.

5. Conclusiones

El uso de modelos predictivos validados e integrados en la atención clínica tiene el potencial de fomentar una toma de decisiones informada y compartida entre el paciente y el profesional sanitario, siempre que a ambos se les pueda proporcionar información fiable y contrastada sobre la posible progresión de la enfermedad y la calidad de vida esperada tras el tratamiento. Esto facilitaría un cambio de paradigma hacia una elección de tratamiento más personalizada y centrada en el paciente (180).

Este trabajo espera contribuir al estado del arte sobre la idoneidad y viabilidad de la implementación de los modelos predictivos basados en algoritmos de aprendizaje automático desarrollados a través de la metodología presentada con la finalidad de ofrecer una atención de mayor calidad al paciente con CP.

Agradecimientos

Esta tesis ha sido posible gracias a la financiación recibida por parte del Instituto de Salud Carlos III a través del proyecto de investigación “Sistema de Soporte al Tratamiento Sistémico y de Dosificación Radioterápica en Pacientes con Cáncer de Pulmón (S32)” (PI16/02104) y por parte de la Consejería de Salud de la Junta de Andalucía con cargo al proyecto de innovación “Servicio sostenible y escalable de apoyo a la investigación clínica y traslacional mediante la plataforma ITC-Bio” (PIN-0441-2017).

A título personal, agradecer el apoyo y los ánimos recibidos por parte de la familia, amigos, compañeros y todas las personas con las que he tenido la oportunidad de compartir algunas de las etapas de este intenso viaje.

Bibliografía

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209–49.
2. Faivre-Finn C, Vicente D, Kurata T, Planchard D, Paz-Ares L, Vansteenkiste JF, et al. LBA49 Durvalumab after chemoradiotherapy in stage III NSCLC: 4-year survival update from the phase III PACIFIC trial. *Annals of Oncology*. 2020 Sep 1;31:S1178–9.
3. Smith-Palmer J, Takizawa C, Valentine W. Literature review of the burden of prostate cancer in Germany, France, the United Kingdom and Canada. *BMC Urology*. 2019 Mar 18;19(1):19.
4. National Institutes of Health. National Cancer Institute. Definition of radiotherapy - NCI Dictionary of Cancer Terms - NCI [Internet]. 2011 [cited 2022 Aug 12]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/radiotherapy>
5. De Ruyscher D, Niedermann G, Burnet NG, Siva S, Lee AWM, Hegi-Johnson F. Radiotherapy toxicity. *Nat Rev Dis Primers*. 2019 Feb 21;5(1):1–20.
6. Ettinger DS, Wood DE, Aisner DL, Akerley W, Bauman JR, Bharat A, et al. NCCN Guidelines Insights: Non–Small Cell Lung Cancer, Version 2.2021: Featured Updates to the NCCN Guidelines. *Journal of the National Comprehensive Cancer Network*. 2021 Mar;19(3):254–66.
7. Kalemkerian GP, Loo BW, Akerley W, Attia A, Bassetti M, Bumber Y, et al. NCCN Guidelines Insights: Small Cell Lung Cancer, Version 2.2018. *J Natl Compr Canc Netw*. 2018 Oct;16(10):1171–82.
8. Maconachie R, Mercer T, Navani N, McVeigh G. Lung cancer: diagnosis and management: summary of updated NICE guidance. *BMJ*. 2019 Mar 28;l1049.
9. Mohan G, T P AH, A J J, K M SD, Narayanasamy A, Vellingiri B. Recent advances in radiotherapy and its associated side effects in cancer—a review. *The Journal of Basic and Applied Zoology*. 2019 Feb 27;80(1):14.
10. Palmer JD, Zaorsky NG, Witek M, Lu B. Molecular markers to predict clinical outcome and radiation induced toxicity in lung cancer. *J Thorac Dis*. 2014 Apr;6(4):387–98.
11. Käsmann L, Dietrich A, Staab-Weijnitz CA, Manapov F, Behr J, Rimner A, et al. Radiation-induced lung toxicity – cellular and molecular mechanisms of pathogenesis, management, and literature review. *Radiation Oncology*. 2020 Sep 10;15(1):214.
12. Kleckner IR, Dunne RF, Asare M, Cole C, Fleming F, Fung C, et al. Exercise for Toxicity Management in Cancer-A Narrative Review. *Oncol Hematol Rev*. 2018;14(1):28–37.

13. Wedlake L, Shaw C, McNair H, Lalji A, Mohammed K, Klopper T, et al. Randomized controlled trial of dietary fiber for the prevention of radiation-induced gastrointestinal toxicity during pelvic radiotherapy. *Am J Clin Nutr*. 2017 Sep;106(3):849–57.
14. Spetz J, Moslehi J, Sarosiek K. Radiation-Induced Cardiovascular Toxicity: Mechanisms, Prevention, and Treatment. *Curr Treat Options Cardio Med*. 2018 Apr;20(4):31.
15. Wilson JD, Hammond EM, Higgins GS, Petersson K. Ultra-High Dose Rate (FLASH) Radiotherapy: Silver Bullet or Fool's Gold? *Frontiers in Oncology* [Internet]. 2020 [cited 2022 Aug 12];9. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2019.01563>
16. Shrestha A, Martin C, Burton M, Walters S, Collins K, Wyld L. Quality of life versus length of life considerations in cancer patients: A systematic literature review. *Psycho-Oncology*. 2019;28(7):1367–80.
17. Liang W, Zhang L, Jiang G, Wang Q, Liu L, Liu D, et al. Development and Validation of a Nomogram for Predicting Survival in Patients With Resected Non–Small-Cell Lung Cancer. *JCO*. 2015 Mar 10;33(8):861–9.
18. Agrawal A, Choudhary A. Health Services Data: Big Data Analytics for Deriving Predictive Healthcare Insights. In: Sobolev B, Levy A, Goring S, editors. *Data and Measures in Health Services Research* [Internet]. Boston, MA: Springer US; 2016 [cited 2022 Aug 12]. p. 1–17. Available from: http://link.springer.com/10.1007/978-1-4899-7673-4_2-1
19. Zheng XQ, Huang JF, Lin JL, Chen L, Zhou TT, Chen D, et al. Incidence, prognostic factors, and a nomogram of lung cancer with bone metastasis at initial diagnosis: a population-based study. *Transl Lung Cancer Res*. 2019 Aug;8(4):367–79.
20. Pontes B, Núñez F, Rubio C, Moreno A, Nepomuceno I, Moreno J, et al. A data mining based clinical decision support system for survival in lung cancer. *Reports of Practical Oncology and Radiotherapy*. 2021;26(6):839–48.
21. Huang S, Yang J, Shen N, Xu Q, Zhao Q. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. *Seminars in Cancer Biology*. 2023 Feb 1;89:30–7.
22. Rajaraman P, Hauptmann M, Bouffler S, Wojcik A. Human individual radiation sensitivity and prospects for prediction. *Ann ICRP*. 2018 Oct 1;47(3–4):126–41.
23. Giuranno L, Ient J, De Ruyscher D, Vooijs MA. Radiation-Induced Lung Injury (RILI). *Frontiers in Oncology* [Internet]. 2019 [cited 2022 Aug 12];9. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2019.00877>
24. Torre-Bouscoulet L, Muñoz-Montaña WR, Martínez-Briseño D, Lozano-Ruiz FJ, Fernández-Plata R, Beck-Magaña JA, et al. Abnormal pulmonary function tests predict the development of radiation-induced pneumonitis in advanced non-small cell lung Cancer. *Respiratory Research*. 2018 Apr 24;19(1):72.

25. Shi A, Zhu G, Wu H, Yu R, Li F, Xu B. Analysis of clinical and dosimetric factors associated with severe acute radiation pneumonitis in patients with locally advanced non-small cell lung cancer treated with concurrent chemotherapy and intensity-modulated radiotherapy. *Radiation Oncology*. 2010 May 12;5(1):35.
26. Tang W, Li X, Yu H, Yin X, Zou B, Zhang T, et al. A novel nomogram containing acute radiation esophagitis predicting radiation pneumonitis in thoracic cancer receiving radiotherapy. *BMC Cancer*. 2021 May 22;21(1):585.
27. Chen H, Senan S, Nossent EJ, Boldt RG, Warner A, Palma DA, et al. Treatment-Related Toxicity in Patients With Early-Stage Non-Small Cell Lung Cancer and Coexisting Interstitial Lung Disease: A Systematic Review. *International Journal of Radiation Oncology*Biography*Physics*. 2017 Jul 1;98(3):622–31.
28. Shi Z, Foley KG, Pablo de Mey J, Spezi E, Whybra P, Crosby T, et al. External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients. *Frontiers in Oncology* [Internet]. 2019 [cited 2022 Aug 13];9. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2019.01411>
29. Ha D, Ries AL. Characterization of Dyspnea in Veteran Lung Cancer Survivors Following Curative-Intent Therapy. *J Cardiopulm Rehabil Prev*. 2020 Mar;40(2):120–7.
30. Sardaro A, McDonald F, Bardoscia L, Lavrenkov K, Singh S, Ashley S, et al. Dyspnea in Patients Receiving Radical Radiotherapy for Non-Small Cell Lung Cancer: A Prospective Study. *Front Oncol*. 2020 Dec 23;10:594590.
31. Albers J, Parker W, Kildea J, Pembroke C, Faria S. Chest wall pain following lung stereotactic body radiation therapy using 48Gy in three fractions: A search for predictors. *Cancer/Radiothérapie*. 2019 Apr 1;23(2):98–103.
32. Voruganti IS, Donovan E, Walker-Dilks C, Swaminath A. Chest Wall Toxicity after Stereotactic Radiation in Early Lung Cancer: A Systematic Review. *Current Oncology*. 2020 Aug;27(4):179–89.
33. Ahn SJ, Kahn D, Zhou S, Yu X, Hollis D, Shafman TD, et al. Dosimetric and clinical predictors for radiation-induced esophageal injury. *International Journal of Radiation Oncology*Biography*Physics*. 2005 Feb 1;61(2):335–47.
34. Nieder C, Imingen KS, Mannsåker B, Yobuta R, Haukland E. Risk factors for esophagitis after hypofractionated palliative (chemo) radiotherapy for non-small cell lung cancer. *Radiation Oncology*. 2020 May 1;15(1):91.
35. Huang J, He T, Yang R, Ji T, Li G. Clinical, dosimetric, and position factors for radiation-induced acute esophagitis in intensity-modulated (chemo)radiotherapy for locally advanced non-small-cell lung cancer. *Onco Targets Ther*. 2018 Sep 21;11:6167–75.
36. Dzul S, Ninia J, Jang H, Kim S, Dominello M. Predictors of Acute Radiation Dermatitis and Esophagitis in African American Patients Receiving Whole-Breast Radiation Therapy. *Practical Radiation Oncology*. 2022 Jan 1;12(1):52–9.

37. Siavashpour Z, Rabiee N, Bakhshandeh M, Anvari A, Zayeri F. Evaluating the relationship between dosimetric parameters and esophagitis in patients who experienced radiotherapy in the cervicothoracic region. *J Can Res Ther.* 2021;0(0):0.
38. Gallegos C, Brito-de la Fuente E, Clavé P, Costa A, Assegehegn G. Chapter Eight - Nutritional Aspects of Dysphagia Management. In: Toldrá F, editor. *Advances in Food and Nutrition Research* [Internet]. Academic Press; 2017 [cited 2022 Aug 13]. p. 271–318. Available from: <https://www.sciencedirect.com/science/article/pii/S1043452616300687>
39. Li H, Li L, Huang X, Li Y, Zou T, Zhuo X, et al. Radiotherapy-induced dysphagia and its impact on quality of life in patients with nasopharyngeal carcinoma. *Strahlenther Onkol.* 2019 Jun 1;195(6):457–67.
40. Petersson K, Finizia C, Tuomi L. Predictors of severe dysphagia following radiotherapy for head and neck cancer. *Laryngoscope Investigative Otolaryngology.* 2021;6(6):1395–405.
41. Paia F, Cristaudo A, Gonnelli A, Giannini N, Cocuzza P, Montrone S, et al. Radiation-induced nausea and vomiting in head and neck cancer: Is it something worth considering in the intensity modulated radiotherapy era? “A narrative review.” *Head & Neck.* 2020;42(1):131–7.
42. Taunk NK, Haffty BG, Kostis JB, Goyal S. Radiation-Induced Heart Disease: Pathologic Abnormalities and Putative Mechanisms. *Frontiers in Oncology* [Internet]. 2015 [cited 2022 Aug 13];5. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2015.00039>
43. Borkenhagen JF, Bergom C, Rapp CT, Klawikowski SJ, Rein LE, Gore EM. Dosimetric Predictors of Cardiotoxicity in Thoracic Radiotherapy for Lung Cancer. *Clinical Lung Cancer.* 2019 Nov 1;20(6):435–41.
44. Ratosa I, Ivanetic Pantar M. Cardiotoxicity of mediastinal radiotherapy. *Reports of Practical Oncology & Radiotherapy.* 2019 Nov;24(6):629–43.
45. da Silva RMFL. Effects of Radiotherapy in Coronary Artery Disease. *Curr Atheroscler Rep.* 2019 Nov 19;21(12):50.
46. Garant A, Spears G, Routman D, Whitaker T, Liao Z, Harmsen W, et al. A Multi-Institutional Analysis of Radiation Dosimetric Predictors of Toxicity After Trimodality Therapy for Esophageal Cancer. *Practical Radiation Oncology.* 2021 Jul 1;11(4):e415–25.
47. Tan LL, Lyon AR. Role of Biomarkers in Prediction of Cardiotoxicity During Cancer Treatment. *Curr Treat Options Cardio Med.* 2018 Jun 19;20(7):55.
48. Zhang TW, Snir J, Boldt RG, Rodrigues GB, Louie AV, Gaede S, et al. Is the Importance of Heart Dose Overstated in the Treatment of Non-Small Cell Lung Cancer? A Systematic Review of the Literature. *International Journal of Radiation Oncology*Biophysics*Physics.* 2019 Jul 1;104(3):582–9.

49. Liang X, Bradley JA, Zheng D, Rutenberg M, Yeung D, Mendenhall N, et al. Prognostic factors of radiation dermatitis following passive-scattering proton therapy for breast cancer. *Radiation Oncology*. 2018 Apr 19;13(1):72.
50. Rattay T, Seibold P, Aguado-Barrera ME, Altabas M, Azria D, Barnett GC, et al. External Validation of a Predictive Model for Acute Skin Radiation Toxicity in the REQUITE Breast Cohort. *Frontiers in Oncology* [Internet]. 2020 [cited 2022 Aug 13];10. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2020.575909>
51. Behroozian T, Milton L, Li N, Zhang L, Lou J, Karam I, et al. Predictive factors associated with radiation dermatitis in breast cancer. *Cancer Treatment and Research Communications*. 2021 Jan 1;28:100403.
52. Curt GA, Breitbart W, Cella D, Groopman JE, Horning SJ, Itri LM, et al. Impact of Cancer-Related Fatigue on the Lives of Patients: New Findings From the Fatigue Coalition. *The Oncologist*. 2000 Oct 1;5(5):353–60.
53. Bandara V, Capp A, Ahmed G, Arm J, Martin J. Assessment and predictors of fatigue in men with prostate cancer receiving radiotherapy and androgen deprivation therapy. *Journal of Medical Imaging and Radiation Oncology*. 2019;63(5):683–90.
54. Bower JE, Wiley J, Petersen L, Irwin MR, Cole SW, Ganz PA. Fatigue after breast cancer treatment: Biobehavioral predictors of fatigue trajectories. *Health Psychology*. 2018;37(11):1025–34.
55. Ruiz-Casado A, Álvarez-Bustos A, de Pedro CG, Méndez-Otero M, Romero-Elías M. Cancer-related Fatigue in Breast Cancer Survivors: A Review. *Clinical Breast Cancer*. 2021 Feb 1;21(1):10–25.
56. Feng LR, Fuss T, Dickinson K, Ross A, Saligan LN. Co-Occurring Symptoms Contribute to Persistent Fatigue in Prostate Cancer. *Oncology*. 2019;96(4):183–91.
57. Feng LR, Fernández-Martínez JL, Zaal KJM, deAndrés-Galiana EJ, Wolff BS, Saligan LN. mGluR5 mediates post-radiotherapy fatigue development in cancer patients. *Transl Psychiatry*. 2018 May 30;8(1):1–11.
58. Feng LR, Juneau P, Regan JM, Liwang J, Alshawi S, Wang A, et al. Brain-derived neurotrophic factor polymorphism Val66Met protects against cancer-related fatigue. *Transl Psychiatry*. 2020 Aug 26;10(1):1–9.
59. Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013 Jan;10(1):27–40.
60. Toesca DAS, Ibragimov B, Koong AJ, Xing L, Koong AC, Chang DT. Strategies for prediction and mitigation of radiation-induced liver toxicity. *Journal of Radiation Research*. 2018 Mar 1;59(suppl_1):i40–9.

61. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective. *Int J Radiat Oncol Biol Phys*. 2015 Dec 1;93(5):1127–35.
62. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015 Jan 6;162(1):55–63.
63. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Medical Research Methodology*. 2022 Apr 8;22(1):101.
64. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*. 2022 Jan 1;77:29–52.
65. Luo Y, Tseng HH, Cui S, Wei L, Ten Haken RK, El Naqa I. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR|Open*. 2019 Nov;1(1):20190021.
66. Isaksson LJ, Pepa M, Zaffaroni M, Marvaso G, Alterio D, Volpe S, et al. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. *Frontiers in Oncology* [Internet]. 2020 [cited 2022 Aug 14];10. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2020.00790>
67. Luo Y, Chen S, Valdes G. Machine learning for radiation outcome modeling and prediction. *Medical Physics*. 2020;47(5):e178–84.
68. Boldrini L, Bibault JE, Masciocchi C, Shen Y, Bittner MI. Deep Learning: A Review for the Radiation Oncologist. *Frontiers in Oncology* [Internet]. 2019 [cited 2022 Aug 14];9. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2019.00977>
69. Dehing-Oberije C, Fung G, Ruyscher DD, Weide H van der, Krishnan S, Rao RB, et al. The Limitations of Dosimetric Parameters for the Prediction of Radiation-Induced Lung Toxicity: An Approach Based on Machine Learning Techniques. *International Journal of Radiation Oncology, Biology, Physics*. 2007 Nov 1;69(3):S488–9.
70. Oh JH, Apte A, Al-Lozi R, Bradley J, Naqa IME. Towards Prediction of Radiation Pneumonitis Arising from Lung Cancer Patients Using Machine Learning Approaches. *Journal of Radiation Oncology Informatics*. 2009;1(1):30–43.
71. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*. 1975 Oct 20;405(2):442–51.
72. Pella A, Cambria R, Riboldi M, Jereczek-Fossa BA, Fodor C, Zerini D, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. *Medical Physics*. 2011;38(6Part1):2859–67.

73. Lee S, Kerns S, Ostrer H, Rosenstein B, Deasy JO, Oh JH. Machine Learning on a Genome-wide Association Study to Predict Late Genitourinary Toxicity After Prostate Radiation Therapy. *International Journal of Radiation Oncology*Biography*Physics*. 2018 May 1;101(1):128–35.
74. Cui S, Luo Y, Tseng HH, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Medical Physics*. 2019;46(5):2497–511.
75. Luna JM, Chao HH, Diffenderfer ES, Valdes G, Chinniah C, Ma G, et al. Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning. *Radiotherapy and Oncology*. 2019 Apr 1;133:106–12.
76. Yu H, Lam KO, Wu H, Green M, Wang W, Jin JY, et al. Weighted-Support Vector Machine Learning Classifier of Circulating Cytokine Biomarkers to Predict Radiation-Induced Lung Fibrosis in Non-Small-Cell Lung Cancer Patients. *Frontiers in Oncology* [Internet]. 2021 [cited 2022 Aug 14];10. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2020.601979>
77. Kotsiantis SB, et al. *Data Preprocessing for Supervised Learning*. 2006.
78. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS [Internet]. 2021. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
79. Reglamento (UE) 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios, por el que se modifican la Directiva 2001/83/CE, el Reglamento (CE) n.º 178/2002 y el Reglamento (CE) n.º 1223/2009 y por el que se derogan las Directivas 90/385/CEE y 93/42/CEE del Consejo (Texto pertinente a efectos del EEE.) [Internet]. OJ L Apr 5, 2017. Available from: <http://data.europa.eu/eli/reg/2017/745/oj/spa>
80. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe’s digital future [Internet]. 2020 [cited 2023 Jul 9]. Available from: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
81. Abravan A, Price G, Banfill K, Marchant T, Craddock M, Wood J, et al. Role of Real-World Data in Assessing Cardiac Toxicity After Lung Cancer Radiotherapy. *Frontiers in Oncology* [Internet]. 2022 [cited 2023 Jul 9];12. Available from: <https://www.frontiersin.org/articles/10.3389/fonc.2022.934369>
82. Conde JM, Guerra JL, Conde AM, Calderon CP, Garcia AM, Benjumea FN, et al. EP-1164 Infrastructure to integrate translational research into clinical decision making for patients with lung cancer. *Radiotherapy and Oncology*. 2015;115:S633.
83. Núñez-Benjumea FJ, González-García S, Moreno-Conde A, Riquelme-Santos JC, López-Guerra JL. Benchmarking machine learning approaches to predict radiation-induced

- toxicities in lung cancer patients. ctRO [Internet]. 2023 Jul 1 [cited 2023 Jul 9];41. Available from: [https://www.ctro.science/article/S2405-6308\(23\)00065-4/fulltext](https://www.ctro.science/article/S2405-6308(23)00065-4/fulltext)
84. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. *Science Translational Medicine*. 2010 Nov 10;2(57):57cm29-57cm29.
 85. Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. *European journal of cancer*. 1994;30(5):635–42.
 86. Conde JM, Moreno-Conde A, Salas-Fernández S, Parra-Calderón CL. ITCBio, a Clinical and Translational Research Platform. *AMIA Annu Symp Proc*. 2020 Mar 4;2019:673–80.
 87. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010 Apr;17(2):124–30.
 88. Cavelaars M, Rousseau J, Parlayan C, de Ridder S, Verburg A, Ross R, et al. OpenClinica. *J Clin Bioinform*. 2015 May 22;5(1):S2.
 89. Brauer J. Mirth: Standards-Based Open Source Healthcare Interface Engine. *Open Source Business Resource*. 2008;(November 2008).
 90. Del Fiol G, Workman TE, Gorman PN. Clinical Questions Raised by Clinicians at the Point of Care: A Systematic Review. *JAMA Internal Medicine*. 2014 May 1;174(5):710–8.
 91. Fairchild A, Harris K, Barnes E, Wong R, Lutz S, Bezjak A, et al. Palliative Thoracic Radiotherapy for Lung Cancer: A Systematic Review. *JCO*. 2008 Aug 20;26(24):4001–11.
 92. De Ruyscher D, Pijls-Johannesma M, Vansteenkiste J, Kester A, Rutten I, Lambin P. Systematic review and meta-analysis of randomised, controlled trials of the timing of chest radiotherapy in patients with limited-stage, small-cell lung cancer. *Annals of Oncology*. 2006 Apr 1;17(4):543–52.
 93. Prezzano KM, Ma SJ, Hermann GM, Rivers CI, Gomez-Suescun JA, Singh AK. Stereotactic body radiation therapy for non-small cell lung cancer: A review. *World J Clin Oncol*. 2019 Jan 10;10(1):14–27.
 94. Murray P, Franks K, Hanna GG. A systematic review of outcomes following stereotactic ablative radiotherapy in the treatment of early-stage primary lung cancer. *BJR*. 2017 Mar;90(1071):20160732.
 95. US Department of Health and Human Services. Common Terminology Criteria for Adverse Events (CTCAE). 2017;155.
 96. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*. 2019 Oct 1;50:71–91.
-

97. Marsland S. *Machine Learning: An Algorithmic Perspective*. New York: Chapman and Hall/CRC; 2011. 406 p.
98. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2015. p. 1200–5.
99. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci Rep*. 2018 Apr 17;8(1):6085.
100. Allison PD. *Missing data*. Thousand Oaks, Calif: Sage Publications; 2002. 93 p. (Sage university papers. Quantitative applications in the social sciences).
101. Marsden PV, Wright JD, editors. *Handbook of survey research*. Second edition. Bingley, UK: Emerald; 2010. 886 p.
102. Tripathi AK, Rathee G, Saini H. Taxonomy of Missing Data along with their handling Methods. In: 2019 Fifth International Conference on Image Information Processing (ICIIP) [Internet]. Shimla, India: IEEE; 2019 [cited 2022 Oct 21]. p. 463–8. Available from: <https://ieeexplore.ieee.org/document/8985715/>
103. Little RJA, Rubin DB. *Statistical analysis with missing data*. Third edition. Hoboken, NJ: Wiley; 2020. 449 p. (Wiley series in probability and statistics).
104. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012 Jan 1;28(1):112–8.
105. Muharemi F, Logofătu D, Leon F. Review on General Techniques and Packages for Data Imputation in R on a Real World Dataset. In: Nguyen NT, Pimenidis E, Khan Z, Trawiński B, editors. *Computational Collective Intelligence* [Internet]. Cham: Springer International Publishing; 2018 [cited 2022 Aug 14]. p. 386–95. (Lecture Notes in Computer Science; vol. 11056). Available from: http://link.springer.com/10.1007/978-3-319-98446-9_36
106. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: Liu L, Özsu MT, editors. *Encyclopedia of Database Systems* [Internet]. New York, NY: Springer; 2016 [cited 2022 Oct 21]. p. 1–7. Available from: https://doi.org/10.1007/978-1-4899-7993-3_565-2
107. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011 Apr 1;21(2):137–46.
108. Gunasegaran T, Cheah YN. Evolutionary cross validation. In: 2017 8th International Conference on Information Technology (ICIT). 2017. p. 89–95.
109. Hawkins DM. The Problem of Overfitting. *J Chem Inf Comput Sci*. 2004 Jan 1;44(1):1–12.
110. Ray S. A Quick Review of Machine Learning Algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). 2019. p. 35–9.

111. Probst P, Boulesteix AL, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*. 2019;20(53):1–32.
112. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995 Sep;20(3):273–97.
113. Evelyn Fix, J.L. Hodges. Nonparametric discrimination: Consistency Properties [Internet]. Berkeley: University of California; 1951 Feb [cited 2022 Aug 14]. (Discriminatory Analysis). Report No.: Report number 4. Available from: <https://apps.dtic.mil/sti/pdfs/ADA800276.pdf>
114. McCulloch WS. The brain computing machine. *Electrical Engineering*. 1949 Jun;68(6):492–7.
115. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*. 1972;135(3):370–84.
116. Lewis DD. Naive (Bayes) at forty: The independence assumption in information retrieval. In: Nédellec C, Rouveirol C, editors. *Machine Learning: ECML-98*. Berlin, Heidelberg: Springer; 1998. p. 4–15. (Lecture Notes in Computer Science).
117. Hutter F, Lücke J, Schmidt-Thieme L. Beyond Manual Tuning of Hyperparameters. *Künstl Intell*. 2015 Nov 1;29(4):329–37.
118. Park E, Chang HJ, Nam HS. Use of Machine Learning Classifiers and Sensor Data to Detect Neurological Deficit in Stroke Patients. *Journal of Medical Internet Research*. 2017 Apr 18;19(4):e7092.
119. Meyer D, editor. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2022 [cited 2022 Oct 21]. Available from: <https://cran.r-project.org/web/packages/e1071/e1071.pdf>
120. Devroye L, Györfi L, Krzyżak A, Lugosi G. On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. *The Annals of Statistics*. 1994 Sep;22(3):1371–85.
121. k-nearest neighbors algorithm. In: Wikipedia [Internet]. 2022 [cited 2022 Oct 21]. Available from: https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1116232528
122. Kuhn M. Classification and Regression Training [Internet]. 2022 [cited 2022 Oct 21]. Available from: <https://github.com/topepo/caret>
123. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*. 2000 Jun 1;22(5):717–27.
124. Tobore I, Li J, Yuhang L, Al-Handarish Y, Kandwal A, Nie Z, et al. Deep Learning Intervention for Health Care Challenges: Some Biomedical Domain Considerations. *JMIR Mhealth Uhealth*. 2019 Aug 2;7(8):e11966.

125. Venables WN, Ripley BD, Venables WN. *Modern applied statistics with S*. 4th ed. New York: Springer; 2002. 495 p. (Statistics and computing).
126. R Core Team. *The R Stats Package*. 2022.
127. Rish I. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001. p. 41–6.
128. Li B, Li H. Prediction of Tunnel Face Stability Using a Naive Bayes Classifier. *Applied Sciences*. 2019 Jan;9(19):4139.
129. Venkatesh B, Anuradha J. A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*. 2019 Mar 1;19(1):3–26.
130. Alelyani S, Tang J, Liu H. Feature Selection for Clustering: A Review. In: *Data Clustering*. Chapman and Hall/CRC; 2014.
131. Dong G, Liu H, editors. *Feature engineering for machine learning and data analytics*. First edition. Boca Raton: CRC Press/Taylor & Francis Group; 2018. 400 p. (Chapman & Hall/CRC data mining & knowledge discovery series).
132. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*. 2019 Sep 1;112:103375.
133. Hall MA. Correlation-based feature selection of discrete and numeric class machine learning. 2000;
134. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, et al. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry*. 2005 Feb 1;29(1):37–46.
135. Kurasa MB, Jankowski A, Rudnicki WR. Boruta – A System for Feature Selection. *Fundamenta Informaticae*. 2010;101(4):271–85.
136. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. 2005 Apr;03(02):185–205.
137. Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In: *Aaai*. 1992. p. 129–34.
138. Breiman L. Random Forests. *Machine Learning*. 2001 Oct 1;45(1):5–32.
139. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*. 2014 Mar 19;14(1):40.
140. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019 Dec 16;17(1):230.

141. Chen M, Wang Z, Jiang S, Sun J, Wang L, Sahoo N, et al. Predictive performance of different NTCP techniques for radiation-induced esophagitis in NSCLC patients receiving proton radiotherapy. *Sci Rep.* 2022 Jun 2;12(1):9178.
 142. Chang E, Du J, Decker R, Yu J, Park H, Aneja S. Development and Validation of a Deep Learning Algorithm to Predict Radiation Pneumonitis and Esophagitis Following Lung Stereotactic Body Radiation Therapy. *International Journal of Radiation Oncology, Biology, Physics.* 2020 Oct 1;108(2):E14.
 143. Luna JM, Chao HH, Shinohara RT, Ungar LH, Cengel KA, Pryma DA, et al. Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation. *Clinical and Translational Radiation Oncology.* 2020 May 1;22:69–75.
 144. El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *International Journal of Radiation Oncology*Biolog*Physics.* 2006 Mar 15;64(4):1275–86.
 145. Wrona A, Dziadziuszko R, Jassem J. Combining radiotherapy with targeted therapies in non-small cell lung cancer: focus on anti-EGFR, anti-ALK and anti-angiogenic agents. *Transl Lung Cancer Res.* 2021 Apr;10(4):2032–47.
 146. De Ruyck K, Sabbe N, Oberije C, Vandecasteele K, Thas O, De Ruyscher D, et al. Development of a Multicomponent Prediction Model for Acute Esophagitis in Lung Cancer Patients Receiving Chemoradiotherapy. *International Journal of Radiation Oncology*Biolog*Physics.* 2011 Oct 1;81(2):537–44.
 147. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). *Int J Radiat Oncol Biol Phys.* 1995 Mar 30;31(5):1341–6.
 148. French CT, Irwin RS, Fletcher KE, Adams TM. Evaluation of a Cough-Specific Quality-of-Life Questionnaire. *Chest.* 2002 Apr 1;121(4):1123–31.
 149. Yousaf N, Lee KK, Jayaraman B, Pavord ID, Birring SS. The assessment of quality of life in acute cough with the Leicester Cough Questionnaire (LCQ-acute). *Cough.* 2011 Jul 18;7(1):4.
 150. Defraene G, Schuit E, De Ruyscher D. Development and internal validation of a multinomial NTCP model for the severity of acute dyspnea after radiotherapy for lung cancer. *Radiotherapy and Oncology.* 2019 Jul 1;136:176–84.
 151. Seow H, Tanuseputro P, Barbera L, Earle CC, Guthrie DM, Isenberg SR, et al. Development and validation of a prediction model of poor performance status and severe symptoms over time in cancer patients (PROVIEW+). *Palliat Med.* 2021 Oct 1;35(9):1713–23.
 152. Dehing-Oberije C, Ruyscher DD, Baardwijk A van, Yu S, Rao B, Lambin P. The importance of patient characteristics for the prediction of radiation-induced lung toxicity. *Radiotherapy and Oncology.* 2009 Jun 1;91(3):421–6.
-

153. Cella L, Monti S, Thor M, Rimner A, Deasy JO, Palma G. Radiation-Induced Dyspnea in Lung Cancer Patients Treated with Stereotactic Body Radiation Therapy. *Cancers*. 2021 Jan;13(15):3734.
154. Tanaka K, Akechi T, Okuyama T, Nishiwaki Y, Uchitomi Y. Factors Correlated with Dyspnea in Advanced Lung Cancer Patients: Organic Causes and What Else? *Journal of Pain and Symptom Management*. 2002 Jun 1;23(6):490–500.
155. Myall NJ, Lei AQ, Wakelee HA. Safety of lorlatinib following alectinib-induced pneumonitis in two patients with ALK-rearranged non-small cell lung cancer: a case series. *Transl Lung Cancer Res*. 2021 Jan;10(1):487–95.
156. Yu H, Wu H, Wang W, Jolly S, Jin JY, Hu C, et al. Machine Learning to Build and Validate a Model for Radiation Pneumonitis Prediction in Patients with Non–Small Cell Lung Cancer. *Clinical Cancer Research*. 2019 Jul 15;25(14):4343–50.
157. Yakar M, Etiz D, Metintas M, Ak G, Celik O. Prediction of Radiation Pneumonitis With Machine Learning in Stage III Lung Cancer: A Pilot Study. *Technol Cancer Res Treat*. 2021 Jan 1;20:15330338211016372.
158. Kirkland RS, Kole AJ, Batra H, Boggs DH, Spencer SA, Dobelbower MC, et al. Predictors of In-Hospital Death in Patients with Lung Cancer Admitted for Acute Radiation Pneumonitis: A Healthcare Cost and Utilization Project (HCUP) Analysis. *Clinical Lung Cancer*. 2021 Sep 1;22(5):e716–22.
159. Tonison JJ, Fischer SG, Viehrig M, Welz S, Boeke S, Zwirner K, et al. Radiation Pneumonitis after Intensity-Modulated Radiotherapy for Esophageal Cancer: Institutional Data and a Systematic Review. *Sci Rep*. 2019 Feb 19;9(1):2255.
160. Mehta V. Radiation pneumonitis and pulmonary fibrosis in non–small-cell lung cancer: Pulmonary function, prediction, and prevention. *International Journal of Radiation Oncology*Biography*Physics*. 2005 Sep 1;63(1):5–24.
161. Nalbantov G, Dekker A, De Ruyscher D, Lambin P, Smirnov EN. The Combination of Clinical, Dose-Related and Imaging Features Helps Predict Radiation-Induced Normal-Tissue Toxicity in Lung-cancer Patients – An in-silico Trial Using Machine Learning Techniques. In: 2011 10th International Conference on Machine Learning and Applications and Workshops. 2011. p. 220–4.
162. Tribius S, Meyer MS, Pflug C, Hanken H, Busch CJ, Krüll A, et al. Socioeconomic status and quality of life in patients with locally advanced head and neck cancer. *Strahlenther Onkol*. 2018 Aug 1;194(8):737–49.
163. Fang FM, Tsai WL, Chiu HC, Kuo WR, Hsiung CY. Quality of life as a survival predictor for esophageal squamous cell carcinoma treated with radiotherapy. *International Journal of Radiation Oncology*Biography*Physics*. 2004 Apr 1;58(5):1394–404.
164. Zhang XJ, Sun JG, Sun J, Ming H, Wang XX, Wu L, et al. Prediction of radiation pneumonitis in lung cancer patients: a systematic review. *J Cancer Res Clin Oncol*. 2012 Dec 1;138(12):2103–16.

165. Vogelius IR, Bentzen SM. A literature-based meta-analysis of clinical risk factors for development of radiation induced pneumonitis. *Acta Oncologica*. 2012 Nov 1;51(8):975–83.
166. Lu C, Lei Z, Wu H, Lu H. Evaluating risk factors of radiation pneumonitis after stereotactic body radiation therapy in lung tumor: Meta-analysis of 9 observational studies. *PLOS ONE*. 2018 Dec 6;13(12):e0208637.
167. Hwang HJ, Kim MY, Choi CM, Lee JC. Anaplastic lymphoma kinase inhibitor related pneumonitis in patients with non-small cell lung cancer. *Medicine (Baltimore)*. 2019 Nov 27;98(48):e18131.
168. Bennett PN. Assessing the calibration of naive bayes posterior estimates. Carnegie-Mellon Univ Pittsburgh PA School of Computer Science; 2000.
169. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*. 2020;27(4):621–33.
170. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015 Jan;68(1):25–34.
171. Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*. 2019 Dec;7(4):221–48.
172. Kikuchi R, Shah NP, Dent SF. Strategies to Prevent Cardiovascular Toxicity in Breast Cancer: Is It Ready for Primetime? *Journal of Clinical Medicine*. 2020 Apr;9(4):896.
173. Rajapakse P. An Update on Survivorship Issues in Lung Cancer Patients. *World J Oncol*. 2021 Jun;12(2–3):45–9.
174. Albino de Queiroz D, André da Costa C, Aparecida Isquierdo Fonseca de Queiroz E, Folchini da Silveira E, da Rosa Righi R. Internet of Things in active cancer Treatment: A systematic review. *Journal of Biomedical Informatics*. 2021 Jun 1;118:103814.
175. Bouazza YB, Chiari I, El Kharbouchi O, De Backer L, Vanhoutte G, Janssens A, et al. Patient-reported outcome measures (PROMs) in the management of lung cancer: A systematic review. *Lung Cancer*. 2017 Nov 1;113:140–51.
176. Walls GM, Osman SOS, Brown KH, Butterworth KT, Hanna GG, Hounsell AR, et al. Radiomics for Predicting Lung Cancer Outcomes Following Radiotherapy: A Systematic Review. *Clinical Oncology*. 2022 Mar 1;34(3):e107–22.
177. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*. 2003 May 1;56(5):441–7.

178. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*. 2021 Jan 1;14(1):49–58.
179. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.) [Internet]. OJ L Apr 5, 2017. Available from: <http://data.europa.eu/eli/reg/2017/745/oj/eng>
180. Craddock M, Crockett C, McWilliam A, Price G, Sperrin M, van der Veer SN, et al. Evaluation of Prognostic and Predictive Models in the Oncology Clinic. *Clinical Oncology*. 2022 Feb 1;34(2):102–13.

Anexo A. AUCs obtenidos tras la validación interna

TABLA A.1. ESOFAGITIS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.55	0.72	0.70	0.73	0.75
χ^2	0.55	0.82	0.72	0.68	0.78
Boruta	0.58	0.84	0.77	0.72	0.83
Votación (subconjunto)	0.54	0.77	0.73	0.69	0.80
mRMR	0.55	0.85	0.72	0.71	0.76
Relief	0.55	0.66	0.65	0.67	0.62
RF	0.57	0.79	0.73	0.73	0.80
IG	0.53	0.80	0.72	0.62	0.78
Votación (clasificación)	0.55	0.83	0.74	0.74	0.70
Gold standard	0.54	0.68	0.61	0.60	0.67

TABLA A.2. TOS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.60	0.87	0.84	0.79	0.88
χ^2	0.57	0.88	0.84	0.76	0.83
Boruta	0.58	0.88	0.85	0.75	0.90
Votación (subconjunto)	0.61	0.87	0.84	0.74	0.90
mRMR	0.62	0.88	0.81	0.81	0.88
Relief	0.64	0.72	0.71	0.78	0.71
RF	0.61	0.86	0.82	0.80	0.83
IG	0.59	0.89	0.83	0.66	0.90
Votación (clasificación)	0.58	0.85	0.84	0.83	0.84
Gold standard	0.57	0.64	0.60	0.60	0.64

TABLA A.3. DISNEA AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.57	0.75	0.72	0.68	0.75
χ^2	0.61	0.78	0.77	0.72	0.77
Boruta	0.57	0.74	0.73	0.72	0.73
Votación (subconjunto)	0.57	0.75	0.72	0.68	0.75
mRMR	0.54	0.81	0.74	0.77	0.79
Relief	0.59	0.68	0.67	0.71	0.64
RF	0.62	0.76	0.70	0.80	0.76
IG	0.64	0.72	0.63	0.71	0.76
Votación (clasificación)	0.52	0.75	0.70	0.78	0.78
Gold standard	0.58	0.64	0.64	0.68	0.67

TABLA A.4. NEUMONITIS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.53	0.71	0.33	0.68	0.69
χ^2	0.56	0.77	0.81	0.69	0.76
Boruta	0.54	0.80	0.80	0.60	0.77
Votación (subconjunto)	0.54	0.78	0.78	0.68	0.77
mRMR	0.57	0.78	0.79	0.73	0.82
Relief	0.56	0.64	0.66	0.66	0.64
RF	0.53	0.76	0.78	0.69	0.73
IG	0.54	0.79	0.77	0.68	0.79
Votación (clasificación)	0.55	0.79	0.80	0.74	0.71
Gold standard	0.54	0.58	0.59	0.56	0.57

TABLA A.5. DISNEA CRÓNICA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.63	0.81	0.79	0.77	0.81
χ^2	0.59	0.83	0.82	0.66	0.68
Boruta	0.60	0.86	0.85	0.79	0.83
Votación (subconjunto)	0.59	0.83	0.81	0.74	0.82
mRMR	0.54	0.87	0.65	0.79	0.81
Relief	0.63	0.62	0.69	0.73	0.67
RF	0.57	0.83	0.78	0.76	0.80
IG	0.65	0.80	0.79	0.71	0.76
Votación (clasificación)	0.58	0.83	0.78	0.76	0.82
Gold standard	0.54	0.54	0.59	0.57	0.56

TABLA A.6. NEUMONITIS CRÓNICA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.59	0.79	0.72	0.76	0.79
χ^2	0.52	0.84	0.81	0.70	0.82
Boruta	0.55	0.82	0.76	0.79	0.77
Votación (subconjunto)	0.49	0.81	0.72	0.79	0.81
mRMR	0.53	0.86	0.73	0.70	0.90
Relief	0.58	0.64	0.69	0.59	0.61
RF	0.53	0.83	0.80	0.83	0.74
IG	0.57	0.79	0.75	0.72	0.83
Votación (clasificación)	0.55	0.84	0.83	0.85	0.73
Gold standard	0.57	0.57	0.57	0.56	0.59

Anexo B. AUCs obtenidos tras la validación externa

TABLA B.1. ESOFAGITIS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.78	0.66	0.76	0.67	0.70
χ^2	0.65	0.80	0.79	0.87	0.83
Boruta	0.62	0.79	0.76	0.72	0.73
Votación (subconjunto)	0.61	0.79	0.79	0.73	0.83
mRMR	0.47	0.81	0.72	0.67	0.84
Relief	0.44	0.81	0.69	0.68	0.60
RF	0.42	0.73	0.81	0.78	0.81
IG	0.61	0.77	0.80	0.68	0.85
Votación (clasificación)	0.53	0.85	0.82	0.77	0.77
Gold standard	0.47	0.75	0.67	0.63	0.82

TABLA B.2. TOS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.57	0.74	0.76	0.70	0.79
χ^2	0.51	0.78	0.81	0.58	0.66
Boruta	0.61	0.77	0.82	0.72	0.75
Votación (subconjunto)	0.53	0.79	0.82	0.67	0.74
mRMR	0.37	0.75	0.46	0.53	0.66
Relief	0.53	0.58	0.68	0.70	0.51
RF	0.57	0.79	0.61	0.61	0.52
IG	0.61	0.80	0.82	0.61	0.77
Votación (clasificación)	0.52	0.76	0.60	0.64	0.55
Gold standard	0.54	0.52	0.68	0.60	0.58

TABLA B.3. DISNEA AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.44	0.56	0.68	0.48	0.56
χ^2	0.59	0.62	0.58	0.69	0.62
Boruta	0.53	0.53	0.56	0.50	0.57
Votación (subconjunto)	0.44	0.56	0.68	0.48	0.56
mRMR	0.49	0.57	0.52	0.64	0.57
Relief	0.65	0.70	0.76	0.73	0.60
RF	0.58	0.69	0.67	0.66	0.51
IG	0.63	0.55	0.60	0.60	0.54
Votación (clasificación)	0.63	0.68	0.75	0.61	0.60
Gold standard	0.47	0.61	0.57	0.56	0.65

TABLA B.4. NEUMONITIS AGUDA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.54	0.77	0.29	0.80	0.71
χ^2	0.56	0.83	0.85	0.66	0.71
Boruta	0.60	0.75	0.69	0.76	0.68
Votación (subconjunto)	0.67	0.84	0.83	0.76	0.83
mRMR	0.48	0.80	0.82	0.74	0.73
Relief	0.59	0.64	0.59	0.62	0.66
RF	0.54	0.81	0.49	0.79	0.68
IG	0.56	0.80	0.83	0.72	0.70
Votación (clasificación)	0.55	0.74	0.69	0.56	0.59
Gold standard	0.47	0.59	0.54	0.54	0.61

TABLA B.5. DISNEA CRÓNICA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.92	0.73	0.57	0.70	0.73
χ^2	0.83	0.57	0.63	0.63	0.60
Boruta	0.95	0.77	0.67	0.77	0.77
Votación (subconjunto)	0.88	0.67	0.60	0.80	0.67
mRMR	0.67	0.97	1.00	0.82	0.72
Relief	0.52	0.97	0.87	1.00	0.88
RF	0.85	0.57	0.88	0.87	0.53
IG	0.45	0.67	0.60	0.73	0.67
Votación (clasificación)	0.45	0.77	0.87	0.67	0.67
Gold standard	0.90	0.77	0.67	0.70	0.67

TABLA B.6. NEUMONITIS CRÓNICA

Método	kNN	GLM	NB	SVM	ANN
CFS	0.62	0.74	0.47	0.73	0.74
χ^2	0.70	0.79	0.68	0.59	0.71
Boruta	0.63	0.70	0.50	0.80	0.71
Votación (subconjunto)	0.77	0.78	0.47	0.67	0.79
mRMR	0.52	0.75	0.74	0.80	0.73
Relief	0.58	0.58	0.53	0.53	0.55
RF	0.55	0.63	0.57	0.57	0.64
IG	0.50	0.76	0.44	0.68	0.79
Votación (clasificación)	0.34	0.76	0.46	0.68	0.51
Gold standard	0.48	0.67	0.65	0.58	0.63