Machine Learning en la predicción de supervivencia en el cáncer de pulmón y metástasis cerebrales

Júlia Giner

Resumen — El cáncer de pulmón no célula pequeña (CPCNP) es una de las neoplasias malignas más común en nuestro ámbito y su pronóstico empeora significativamente en presencia de metástasis cerebrales, afectando entre un 30-60% de los pacientes. Aunque se han realizado avances en el tratamiento del CPCNP en estadios avanzados, los pacientes con metástasis cerebrales han sido excluidos de muchos ensayos clínicos, lo que limita el conocimiento sobre la efectividad de los nuevos tratamientos. El objetivo de este trabajo es desarrollar un modelo de machine learning para predecir la supervivencia en pacientes con CPCNP y metástasis cerebrales, evaluando también el impacto del tratamiento local. Para ello, en el presente estudio se analizaron diferentes modelos de machine learning, como el Árbol de Decisión, Random Forest y Red Neuronal, entre otros. Los resultados obtenidos fueron muy prometedores y sugieren que el uso de estas herramientas puede ayudar a la toma de decisiones clínicas en pacientes con metástasis cerebrales.

Palabras clave —Cáncer de pulmón no célula pequeña, Metástasis cerebrales, Supervivencia, Machine learning, modelos predictivos, Random Forest, Red Neuronal

Abstract— Non-small cell lung cancer (NSCLC) is one of the most common malignant neoplasms in our setting, and its prognosis worsens significantly in the presence of brain metastases, affecting 30-60% of patients. Although there have been advances in the treatment of advanced NSCLC, patients with brain metastases have often been excluded from clinical trials, limiting knowledge of treatment effectiveness. The objective of this study was to develop a machine learning model to predict survival in NSCLC patients with brain metastases, while also evaluating the impact of local treatment. Three models were trained: Decision Tree, Random Forest, and Deepnet. The Random Forest model with 100 trees showed the best performance, with an AUC-ROC of 0.7601. These results suggest that the use of machine learning can improve clinical decision-making in patients with brain metastases.

Index Terms— Non-small cell lung cancer, Brain Metastases, Survival, Machine Learning, Predictive models, Random Forest, Deepnet

Máster en Inteligencia Artificial y Big Data en Salud

1. Introducción

El Cáncer de Pulmón es, hoy en día, una de las neoplasias malignas más frecuentes y representa la primera causa de muerte por cáncer a nivel mundial (1). El cáncer de pulmón célula no pequeña (CPCNP) es el subtipo más frecuente y su incidencia sigue en aumento, especialmente en mujeres. (1,2) Cada año, se diagnostican más de 2,2 millones de nuevos casos de cáncer de pulmón, lo que representa alrededor del 11,4% de todos los diagnósticos de cáncer. El pronóstico de esta enfermedad varía significativamente dependiendo del subtipo, el estadio en el que se diagnostica o de múltiples factores relacionados con el paciente.

En términos de pronóstico, la tasa de supervivencia a cinco años para el CPCNP es del 25% (3), reflejando su naturaleza altamente invasiva y la tendencia a diagnosticarse en etapas avanzadas.

Entre un 30-60% de pacientes CPNCP presentan metástasis cerebrales durante la evolución de la enfermedad, y hasta un 10% presentan enfermedad cerebral en el momento del diagnóstico. Esto se asocia a un peor pronóstico, alta mortalidad y deterioro de la calidad de vida (4,5).

Aunque hay importantes avances en el tratamiento del CPCNP en estadios avanzados, los pacientes con metástasis cerebrales han sido, en general, excluidos de los ensayos clínicos, por lo que se dispone de escasa información sobre cómo los nuevos agentes terapéuticos son o no efectivos. El tratamiento local de las metástasis cerebrales sigue siendo controvertido (6).

El objetivo principal de este trabajo es desarrollar un modelo de machine learning para predecir la supervivencia en pacientes con CPCNP y metástasis cerebrales, explorando

E-mail de contacto jginer@tauli.cat
Trabajo tutorizado por: Remo Suppi

también el impacto del tratamiento local y la cirugía sobre la supervivencia. Este estudio busca proporcionar un punto de partida para identificar patrones complejos y mejorar la toma de decisiones clínicas en pacientes con metástasis cerebrales.

2. MATERIAL Y MÉTODOS

2.1 Población a estudio

Para este trabajo se han utilizado los datos obtenidos de forma retrospectiva de pacientes diagnosticados de cáncer de pulmón con metástasis cerebrales, tratados en el servicio de Oncologia Mèdica del Parc Tauí Hospital Universitari (CSPT) entre enero de 2019 hasta mayo de 2024.
Del total de 285 sujetos introducidos en la base de datos, se han utilizado sólo los datos de pacientes con CPCNP y sospecha de metástasis cerebrales en el momento de su diagnóstico, siendo un total de 97 pacientes.
Dado el tamaño relativamente pequeño de la muestra original, se consideró necesario incrementar el número de casos para mejorar la robustez estadística y la generalización de los modelos de aprendizaje automático utilizados en este estudio.
Los datos originales se encuentran almacenados en la plataforma *onQos*, la cual ha sido aprobada por el Comitẽ de ética del Hospital Parc Taulí *(28 de septiembre del 2020).* Debido a que los datos han sido anonimizados no se requiere el consentimiento informado de los pacientes.

2.2 Generación de Datos Sintéticos

Dada la limitación del tamaño de la muestra original de sólo 97 pacientes, se generaron datos sintéticos para aumentar el tamaño de la muestra y mejorar la capacidad predictiva de los modelos de aprendizaje. Para la generación de datos sintéticos, se utilizó la biblioteca *Synthetic Data Vault (SDV)* (7), implementada en el entorno de Google Colab.

El procedimiento para la generación de datos sintéticos incluyó los siguientes pasos:

- Conversión del archivo de datos originales en formato Excel a formato CSV para su procesamiento con SDV
- Importación de los datos y preparación
- Ajuste y corrección de datos: Una vez cargados los datos se realizó una revisión manual para asegurar que se ajustaban correctamente a las características de los datos originales.
- Validación y guardado de la metainformación.
- Exportación a BigML

A través de la generación de datos sintéticos, se obtuvo una muestra de 470 sujetos.

2.3 Variables Utilizadas

Para el desarrollo del modelo se eligieron 38 variables, teniéndose en cuenta los factores pronósticos conocidos, según la literatura disponible en la actualidad. (8)

Además, de cada paciente se recogió la siguiente información;

- Datos demográficos y clínicos: fecha de nacimiento, sexo biológico, hábito tabáquico
- Escala ECOG (*Eastern Cooperative Oncology Group*) para medir la calidad de vida en el momento del diagnóstico a través de la escala: (9)
    - 0: paciente asintomático
    - 1: paciente con síntomas, pero sin afectación en sus actividades cotidianas
    - 2: paciente sin capacidad de desempeñar ningún trabajo, con síntomas que obligan a permanecer en cama durante varias horas al día
    - 3: necesidad de estar encamado más del 50% del día
    - 4: encamado el 100% del día necesitando ayuda para todas las actividades
    - 5: fallecido

En cuanto a los antecedentes patológicos de los pacientes, se tuvo en cuenta si habían tenido:

- Antecedente de enfermedad oncológica previa
- Factores de riesgo cardiovascular (hipertensión, diabetes mellitus, dislipemia y obesidad)
- Antecedentes de cardiopatía, neumopatía o enfermedad autoinmune
- Causa de la muerte en caso de defunción (enfermedad, infección, toxicidad al tratamiento u otras)

Con relación a las variables relacionadas con la enfermedad neoplásica, se registró:

- Tipo histológico del tumor (Adenocarcinoma, Carcinoma escamoso, No especificado)
- Estadificación *TNM* de la 8ª edición (donde la letra *T* describe el tamaño del tumor, la letra *N* describe la afectación por cáncer de los ganglios linfáticos cercanos y la letra *M* describe las metástasis) (10)

Las alteraciones moleculares que se consideraron relevantes para el análisis fueron:

- Mutación de EGFR *(Epidermal Growth Factor Receptor*)
- Reordenamiento de ALK *(Anaplastic Lymphoma Kinase)*
- Reordenamiento de ROS1
- Fusión de NTRK (*Neurotrophic Tyrosine Receptor Kinase*)
- Amplificación de HER/neu (*erbB-Tyrosine Receptor Kinase)*
- Mutación de RET
- Mutación de BRAF
- Mutación de KRAS
- Otras

Para la determinación de PDL1 (*Programmed Death-ligand 1*) se clasificaron en tres grupos según su expresión;

- <1%
- 1-49
- >50%

La presencia de metástasis extracerebrales fueron registradas y se tuvo en cuenta si el paciente presentaba enfermedad

metastática en:

- Pulmón
- Hueso
- Hígado
- Glándulas adrenales
- Adenopatías extratorácicas
- Pleura
- Pericardio
- Otras

Para conocer el impacto en la supervivencia de los pacientes del tratamiento local sobre las metástasis cerebrales, se registró el número de lesiones cerebrales (una/pocas/muchas) así como el tipo de tratamiento;

- Cirugía
- Radiocirugía
- Radioterapia focal
- Radioterapia holocraneal

Por último, se tuvo en cuenta, como variable objetivo el estado de supervivencia:

- Muerto con enfermedad
- Vivo con enfermedad
- Vivo sin evidencia de enfermedad

2.4 Preprocesamiento y Revisión de Datos

Para garantizar la calidad del conjunto de datos, se llevó a cabo un preprocesamiento y revisión de las variables recopiladas, dentro del entorno de BigML. Este proceso incluyó;

Revisión y Validación de las Variables:
Se comprobó que todas las variables estuvieran correctamente catalogadas en función de su tipo. Se identificaron y clasificaron como categóricas, numéricas o fechas, asegurando que cada una correspondiera al tipo de datos adecuado para su análisis posterior. Se eliminó la variable identificativa "Patient Number" por no aportar información relevante en el análisis.

Cálculo de la Edad de los Pacientes:
Dado que la información demográfica de los pacientes incluía la fecha de nacimiento y la fecha de diagnóstico, se procedió a calcular la edad de cada paciente en el momento del diagnóstico. Este cálculo se realizó utilizando la fórmula Lisp en la función "Add Fields" de BigML, generando una nueva variable de edad que fue incorporada al conjunto de datos.

Análisis de Valores Faltantes:
Se analizaron las variables que contenían valores faltantes. Se encontró que la variable "cantidad de paquetes/año de tabaco consumido", tenía un elevado número de *missing values*, por lo que se decidió eliminarla para el análisis definitivo.
En las demás variables, donde el número de valores faltantes era mínimo (inferior al 3%) se decidió no aplicar técnicas de imputación de datos, considerando que no afectarían significativamente los resultados del análisis.

Detección de Outliers:
Se llevó a cabo un análisis de detección de *outliers* para identificar posibles valores atípicos que pudieran conllevar sesgos en los resultados. Este análisis no reveló la presencia de *outliers* en las variables evaluadas, por lo que no fue necesario aplicar técnicas de manejo de valores atípicos.

2.5 Análisis Estadístico

Para las variables categóricas se utilizaron las frecuencias absolutas y porcentajes. Para las variables numéricas, la media.

2.5.1. Análisis de Clústers

Como parte del análisis exploratorio de datos, se realizó un análisis exploratorio no supervisado de clúster utilizando el algoritmo K-means en BigML. El objetivo de este análisis era identificar posibles agrupaciones de los pacientes basadas en sus características clínicas y demográficas, lo cual podría revelar patrones subyacentes o grupos de pacientes con comportamientos de supervivencia similares.

Se exploraron diferentes configuraciones del algoritmo, determinando el cálculo de 6, 8 y 11 clústeres. Para cada configuración, se analizaron las características y la distribución de los pacientes dentro de los clústeres.

2.5.2 Análisis de Predicción de supervivencia

Se evaluaron tres modelos de predicción para determinar su efectividad en la clasificación de los estados de supervivencia de los pacientes:

1. Árbol de decisión individual:

Este modelo utiliza un enfoque de árbol binario para dividir el espacio de características de los datos en subconjuntos más pequeños, basándose en la variable objetivo. Es fácil de interpretar y proporciona una visión clara de cómo se toman las decisiones de predicción. Sin embargo, es susceptible al *overfitting*, por lo que, en este estudio, se utilizó principalmente como un modelo de referencia para comparar su rendimiento con los modelos más complejos.

2. Random forest:

Compuesto por múltiples árboles de decisión que se entrenan de forma independiente con diferentes datos y características, y cuyas predicciones se combinan para mejorar la precisión y reducir el riesgo de sobreajuste.
En este estudio, se implementaron dos configuraciones del modelo en BigML: una con 50 árboles y otra con 100 árboles. Cada configuración se entrenó utilizando una proporción de división de datos del 80% para entrenamiento y 20% para prueba.
Las métricas de evaluación incluyeron la exactitud (accuracy), el F1-Score, la sensibilidad (recall), y el AUC-ROC.

3. Red Neuronal (Deepnet):

Implementada para capturar relaciones no lineales complejas en los datos. En este caso, la red fue configurada con dos capas ocultas. Se seleccionó el algoritmo 'Adam' como optimizador, con una tasa de aprendizaje de 0.01. Se permitió el manejo de

valores faltantes en datos numéricos, lo cual resulta crucial para mantener la integridad del análisis.
La red fue entrenada con un total de 376 instancias, utilizando una división del 80% para entrenamiento y 20% para validación.

3. RESULTADOS

3.1. Estadística Descriptiva

En la *Tabla 1* se resumen las características demográficas y clínicas de los pacientes, incluyendo el sexo, edad, hábito tabáquico, y el estado basal según el ECOG. En general, el 64.68% de los pacientes eran hombres, con una edad media de 65 años (rango 33-84 años). Casi la mitad de los pacientes eran ex-fumadores, mientras que un 36.38% continuaban fumando al momento del diagnóstico. La mayoría de los pacientes presentaban un estado basal ECOG de 1 o 2.

Como se puede observar en la *Tabla 1*, la mayoría de la población nunca había tenido una enfermedad neoplásica (94.89%) siendo el cáncer de pulmón el primer diagnóstico oncológico.

En cuanto a las comorbilidades, el 44.89% de los pacientes presentaban factores de riesgo cardiovascular, el 19.15% tenían cardiopatía asociada, y la mayoría (77.23%) no presentaba neumopatías relevantes. Únicamente el 6.38% padecían de una enfermedad autoinmune, mientras que un 32.34% (152 pacientes) presentaban otros antecedentes médicos relevantes.

Los resultados del estudio histológico, así como las alteraciones moleculares detectadas en la población, se resumen en la *Tabla 1*. El tipo histológico más común fue el adenocarcinoma, que representó el 82.98% de los casos. En cuanto a las alteraciones moleculares, el 20.64% de los pacientes presentaban mutaciones en EGFR, siendo las más comunes la mutación L858R en el exón 21 y la deleción del exón 19. Además, se detectaron reordenamientos en ALK y ROS1, así como alteraciones en cMET y KRAS.
Las localizaciones de metástasis extracerebrales más frecuentemente observadas fueron en pulmón, glándulas suprarrenales y hueso, seguidas de adenopatías extrapulmonares y otras localizaciones, como se muestra en la *Tabla 1*.

Un 51.06% de los pacientes recibió tratamiento local para las metástasis cerebrales. El tratamiento más común fue la radioterapia holocraneal, seguida de la radioterapia focal, radiocirugía y cirugía. La *Figura 1* ilustra la distribución de los distintos tipos de tratamiento local administrados.

De los 470 pacientes, 322 (68.51%) habían fallecido en el momento del análisis, 138 (29.36%) estaban vivos con enfermedad activa, y solo un 2.13% del total de la población estaban vivos sin evidencia de enfermedad oncológica activa. La causa de la defunción fue predominantemente debida a la enfermedad (84.87%).

TABLA 1
CARACTERÍSTICAS DE LA POBLACIÓN

Características	Total (n=470)
Edad- (años)	Media 65 (rango: 33-84)
Sexo- (%)	
Hombres	64.68%
Mujeres	35.32%
Hábito tabáquico- (%)	
Exfumadores	47.87%
Fumadores activos	36.38%
Nunca fumadores	15.74%
ECOG *Performance Status*- (%)	
0	18.65%
1	52.81%
2	24.04%
3	4.49%
Antecedente neoplasia previa- (%)	
Si	94.89%
No	5.11%
Histología del Tumor- (%)	
Adenocarcinoma	82.98%
Carcinoma escamoso	4.26%
Carcinoma de células grandes	0.43%
No especificado (NOS)	12.34%
Alteraciones Moleculares- (%)	
EGFR	20.64%
ALK	3.19%
ROS1	5.96%
cMET	5.75%
KRAS	26.38%
Metástasis Extracerebrales- (%)	
Pulmón	37.23%
Glándulas Suprarrenales	32.77%
Hueso	36.38%
Adenopatías extrapulmonares	22.13%
Estado de Supervivencia- (%)	
Muertos	68.51%
Vivos con enfermedad activa	29.36%
Vivos sin evidencia de enfermedad	2.13%

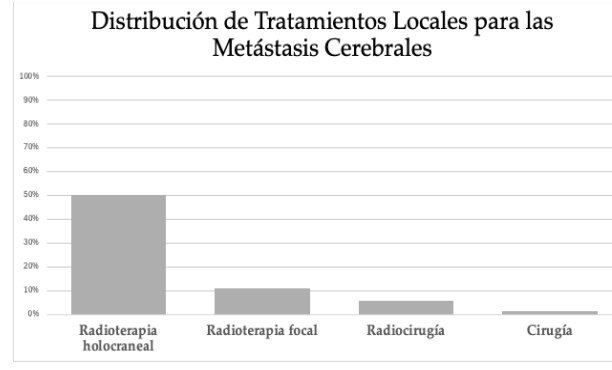

Fig 1. Distribución de los tipos de tratamientos locales que los pacientes recibieron en las metástasis del sistema nervioso central

3.2. Clustering

Los resultados de este análisis no revelaron una segmentación concluyente para la predicción de la supervivencia en los pacientes. Aunque se observaron algunas agrupaciones

consistentes en términos de variables específicas, la variabilidad interna dentro de los clústers fue considerable (with_ss: 2.623) y la separación entre ellos fue limitada (between_ss: 0.511, ratio_ss: 0.163).
Los centroides de los clústers, que representan las características promedio de cada grupo, mostraron diferencias sutiles, pero no se observaron patrones claramente distintos en cuanto a la supervivencia de los pacientes ni en relación con las alteraciones moleculares.
Además, las distancias entre los centroides fueron pequeñas, lo que sugiere que los clústers están cerca unos de otros y no proporcionan una segmentación útil. Estos hallazgos subrayan la complejidad de los datos y podrían indicar la necesidad de explorar métodos alternativos de clúster o incorporar más variables que puedan influir en los patrones de agrupación

En la *Figura 2,* se visualiza la distribución de los pacientes en los distintos clústers.

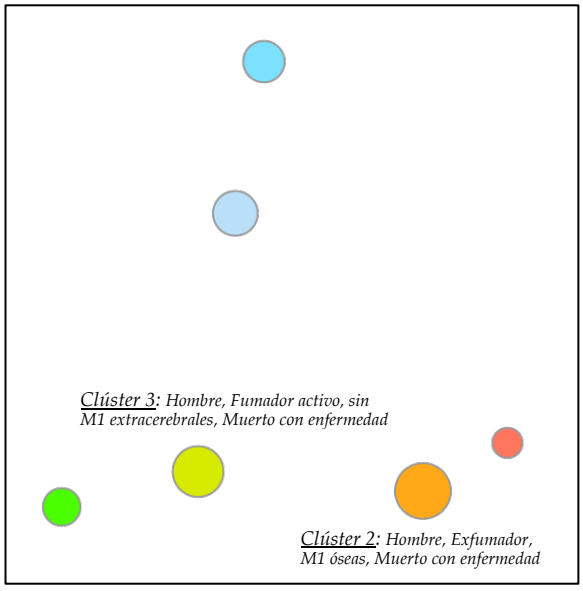

Fig 2. Gráfico de dispersión de los clústeres generados mediante K-means (k=6). El clúster 2 (naranja) agrupa principalmente a varones con adenocarcinoma, metástasis óseas, y exfumadores, mientras que el clúster 3 (amarillo) incluye principalmente a varones, fumadores activos y sin metástasis extracerebrales.

3.3. Predicción de Supervivencia

3.3.1. Árbol de decisión individual

El árbol de decisión individual mostró un rendimiento moderado en la clasificación de los pacientes. El modelo identificó correctamente el 67.02% de los casos, con una alta sensibilidad (79.10%), lo que indica una buena capacidad para identificar correctamente los pacientes positivos.
La precisión obtenida fue del 75.71%, con un F1-score del 0.7794 y un AUC.ROC del 0.6114. Ésta última refleja la capacidad limitada del modelo para discriminar entre clases.

En la *Tabla 2* se puede ver la matriz de confusión obtenida.

TABLA 2
MATRIZ DE CONFUSIÓN

Actual vs Predicho	Muerto con enfermedad	Vivo con enfermedad	Vivo sin enfermedad	Total
Muerto con enfermedad	53 (VP)	11 (FN)	3 (FN)	67
Vivo con enfermedad	16 (FP)	9 (VN)	0 (VN)	25
Vivo sin enfermedad	1 (FP)	1 (VN)	0 (VN)	2
Predicho	70	21	3	94

VP: Verdadero Positivo, FN : Falso Negativo, FP: Falso Positivo, VN: Verdadero Negativo

3.3.2 Random forest

Se entrenaron dos configuraciones del modelo Random Forest: una con 50 árboles y otra con 100 árboles, utilizando una proporción del 80% de los datos para entrenamiento y el 20% para validación.

El modelo de 50 árboles mostró un rendimiento sólido en la clasificación, con una exactitud del 74.47%, lo que indica que más de tres cuartas partes de las predicciones fueron correctas. El modelo demostró una alta sensibilidad (92.54%) con una precisión fue del 76.54%, lo que sugiere que, aunque se detectaron muchos casos positivos, algunos falsos positivos también estuvieron presentes. El F1-Score fue de 0.8378, confirmando el buen rendimiento general del modelo. Finalmente, el AUC-ROC alcanzó un valor de 0.7336, lo que indica una capacidad moderada para discriminar entre las clases positivas y negativas.
Por otro lado, el modelo de 100 árboles presentó una exactitud y una sensibilidad discretamente inferior, del 71.28% y 91.05%, respectivamente. En cuanto a la precisión, el valor fue del 74.39%, sugiriendo que, aunque el modelo detectó muchos casos positivos, también resultaron algunos falsos positivos. El F1-Score fue de 0.8188, lo que refleja un equilibrio razonable entre la precisión y la sensibilidad. Finalmente, el AUC-ROC fue de 0.7601, lo que sugiere una mejor capacidad para distinguir entre clases positivas y negativas en comparación con el modelo de 50 árboles.

En la *Tabla 3* se puede observar la matriz de confusión para las dos configuraciones.

TABLA 3

MATRIZ DE CONFUSIÓN (50 árboles)

Actual vs Predicho	Muerto con enfermedad	Vivo con enfermedad	Vivo sin enfermedad	Total
Muerto con enfermedad	62 (VP)	5 (FN)	0 (FN)	67
Vivo con enfermedad	17 (FP)	8 (VN)	0 (VN)	25
Vivo sin enfermedad	2 (FP)	0 (VN)	0 (VN)	2
Predicho	81	13	0	94

VP: Verdadero Positivo, FN : Falso Negativo, FP: Falso Positivo, VN: Verdadero Negativo

MATRIZ DE CONFUSIÓN (100 árboles)

Actual vs Predicho	Muerto con enfermedad	Vivo con enfermedad	Vivo sin enfermedad	Total
Muerto con enfermedad	61 (VP)	6 (FN)	0 (FN)	67
Vivo con enfermedad	19 (FP)	6 (VN)	0 (VN)	25
Vivo sin enfermedad	2 (FP)	0 (VN)	0 (VN)	2
Predicho	82	12	0	94

VP: Verdadero Positivo, FN : Falso Negativo, FP: Falso Positivo, VN: Verdadero Negativo

En la *figura 3* se observa la curva ROC de las dos configuraciones del modelo Random Forest (RF).

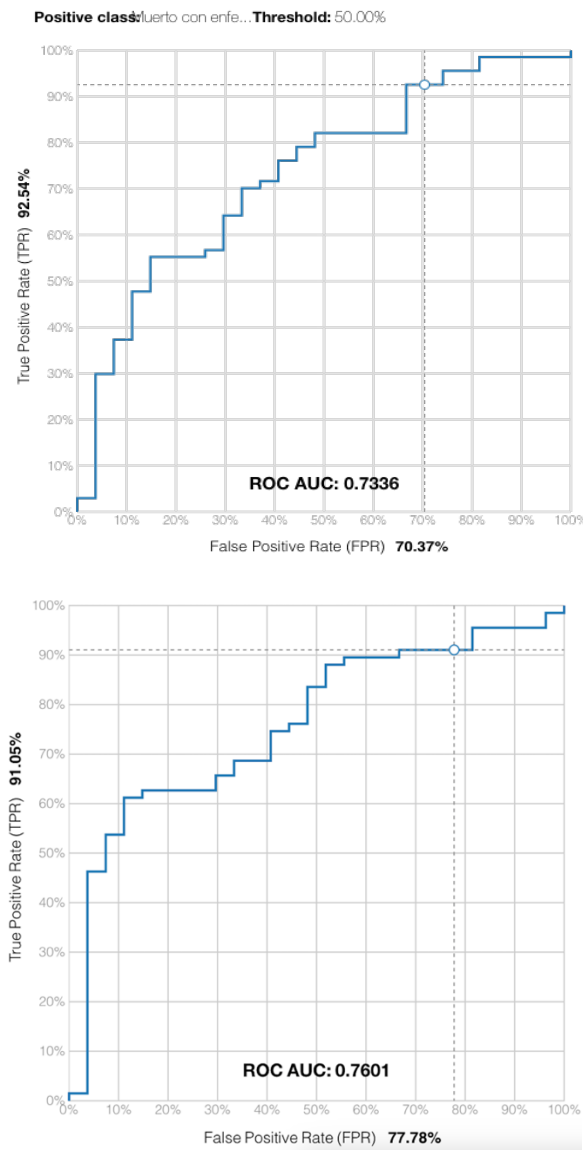

Fig 3. Arriba, la Curva ROC del RF con 50 árboles. Abajo, la Curva ROC del RF con 100 árboles

3.3.3 Red Neuronal (Deepnet)

El modelo identificó correctamente más del 70% de las instancias (exactitud del 72.34%). La sensibilidad fue del 82.09%, lo que indica que fue capaz de detectar la mayoría de los casos positivos. La precisión alcanzó un valor de 79.71%, lo que sugiere un buen equilibrio entre la detección de verdaderos positivos y la minimización de falsos positivos. El F1-Score fue de 0.8088, reflejando un buen rendimiento general del modelo. Por último, el AUC-ROC de 0.7253 muestra una capacidad moderada del modelo para discriminar entre clases positivas y negativas.

En la *Tabla 4* se puede observar la matriz de confusión. En la *figura 4* se observa la curva ROC.

TABLA 4
MATRIZ DE CONFUSIÓN

Actual vs Predicho	Muerto con enfermedad	Vivo con enfermedad	Vivo sin enfermedad	Total
Muerto con enfermedad	55 (VP)	12 (FN)	0 (FN)	67
Vivo con enfermedad	13 (FP)	11 (VN)	1 (VN)	25
Vivo sin enfermedad	1 (FP)	1 (VN)	0 (VN)	2
Predicho	69	24	1	94

VP: Verdadero Positivo, FN : Falso Negativo, FP: Falso Positivo, VN: Verdadero Negativo

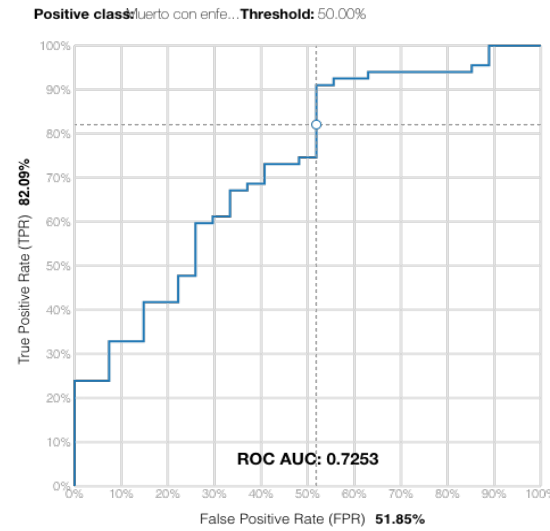

Fig 4. Curva ROC del modelo Deepnet

4. DISCUSIÓN

En general, el perfil de pacientes con cáncer de pulmón y metástasis cerebrales que se observa en este estudio es congruente con los datos de la literatura (11,12). La mayoría de los pacientes eran varones con una edad media de 65 años y una clara asociación con el tabaquismo. La histología predominante fue el adenocarcinoma, lo que concuerda con su mayor tendencia a diseminarse al sistema nervioso central (13).

Un hallazgo sorprendente fue la baja prevalencia de comorbilidades respiratorias, particularmente de enfermedad pulmonar obstructiva crónica (EPOC). Esta baja prevalencia podría explicarse por el infradiagnóstico generalizado de la EPOC , que en España alcanza hasta el 75%, según el estudio EPISCAN II (14). Esto sugiere que la verdadera prevalencia de EPOC en esta población podría estar subestimada, limitando la capacidad de explorar más a fondo la interacción entre ambas patologías en el contexto de las metástasis cerebrales.

En cuanto a las alteraciones moleculares, el 20.64% de los pacientes presentaron mutaciones en EGFR, lo cual concuerda con lo reportado en la literatura (15). Las frecuencias de otras mutaciones, como las asociadas a KRAS, ALK y cMET, también se alinean con estudios previos. Sin embargo, la incidencia de reordenamientos de ROS1 fue inesperadamente alta (5.96% frente a 1-2% reportado en otras series) (15). Esta discrepancia podría deberse a la inclusión de datos sintéticos en el conjunto de datos, lo que posiblemente distorsionó la representación real de esta mutación.

El tratamiento más común para las metástasis cerebrales (M1) fue la radioterapia holocraneal, seguida de la radioterapia focal, la radiocirugía y la cirugía. Este hallazgo es coherente con la práctica clínica histórica, donde la radioterapia holocraneal ha sido el estándar. Sin embargo, con los avances en el diagnóstico molecular y el desarrollo de terapias dirigidas, se ha observado un aumento en la supervivencia de los pacientes con CPCNP, lo que ha impulsado la adopción de tratamientos locales más precisos y menos tóxicos, aunque la estandarización de estos sigue siendo controversial debido a la heterogeneidad de los pacientes y la variabilidad en la presentación clínica (6).

Este escenario subraya la importancia de utilizar técnicas como el machine learning para identificar patrones complejos y guiar la selección del tratamiento más adecuado.

El análisis no supervisado mediante clustering no mostró patrones claros que permitieran una diferenciación efectiva de los pacientes según su supervivencia. La variabilidad dentro de los clústeres refleja la complejidad inherente a esta población, lo que sugiere la necesidad de incorporar más variables o utilizar métodos más avanzados para mejorar la segmentación y predicción

Por otro lado, el análisis supervisado con tres modelos de machine learning (Árbol de Decisión, Random Forest y Deepnet) mostró que el Random Forest con 100 árboles fue el más adecuado para predecir la supervivencia, aunque su rendimiento fue moderado, con un AUC-ROC que no alcanzó valores superiores a 0.8. Esta moderación en el rendimiento puede atribuirse a varias limitaciones, como el tamaño de la muestra y la naturaleza de los datos sintéticos, que pueden haber afectado la calidad de las predicciones.
El Deepnet, aunque competitivo, no superó al Random Forest en cuanto a la capacidad para predecir correctamente la supervivencia. El Árbol de Decisión, si bien más fácil de interpretar, no fue capaz de ofrecer un rendimiento comparable en este conjunto de datos.

El uso de modelos como el Random Forest y el Deepnet podría ayudar a identificar patrones clínicos complejos que influyen en la supervivencia de los pacientes, proporcionando una herramienta potencialmente útil para la toma de decisiones clínicas. Sin embargo, el desempeño moderado de los modelos en general, con AUC-ROC por debajo de 0.8, indica que aún hay margen para mejorar las predicciones, tal vez incorporando más variables o ajustando los hiperparámetros de los modelos.

4.1. Limitaciones

La principal limitación de este estudio es el tamaño muestral, incluso con la generación de datos sintéticos. Un conjunto de datos más amplio y real probablemente habría mejorado el rendimiento de los modelos y su generalización.

En cuanto a la calidad de los datos, aunque el manejo de valores faltantes fue mínimo y se decidió no aplicar imputación, esto podría haber influido en los resultados.
Asimismo, los modelos entrenados presentan ciertas limitaciones. Con un ajuste manual de hiperparámetros y un mayor conocimiento de la materia, es posible que los resultados hubieran sido mejores.

5. CONCLUSIONES

Este estudio ha demostrado que el uso de machine learning puede ser útil para predecir la supervivencia en pacientes con CPCNP y metástasis cerebrales. Entre los tres modelos evaluados, el Random Forest con 100 árboles mostró el mejor rendimiento, con un AUC-ROC de 0.7601, superando al Árbol de Decisión y al modelo Deepnet.

Aunque los resultados son prometedores, el pequeño tamaño muestral y el uso de datos sintéticos representan limitaciones importantes.

Este trabajo proporciona un punto de partida para el desarrollo de modelos predictivos más robustos que puedan ser utilizados en la práctica clínica. Futuros estudios deberán enfocarse en ampliar el conjunto de datos y optimizar los modelos para mejorar la precisión y la capacidad de discriminación, con el fin de apoyar la toma de decisiones clínicas en el tratamiento de estos pacientes.

6. REFERENCIAS

[1] Ferlay J, Ervik M, Lam F, et al. "Global Cancer Observatory: Cancer Today (Version 1.0)." *International Agency for Research on Cancer*, 2024. Accessed February 1, 2024. Available at: https://gco.iarc.who.int/today. (URL link)

[2] Sociedad Española de Oncología Médica (SEOM), "Las Cifras del Cáncer en España 2024," https://seom.org/images/publicaciones/informes-seom-de-evaluacion-de-farmacos/LAS_CIFRAS_2024.pdf. (2024). (URL link)

[3] L.E. Hendriks, K.M. Kerr, J. Menis, T.S. Mok, U. Netle, A. Pasaro, S. Peters, D. Planchard, E.F. Smit, B.J. Solomon, G. Veronesi, M. Reck, "Non-oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis,

treatment and follow-up," ESMO Guidelines Committee. (Guideline document)
[4] Barnholtz-Sloan JS, Sloan AE, Davis FG, et al. "Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the Metropolitan Detroit Cancer Surveillance System." *Journal of Clinical Oncology*, 2004; 22:2865. doi: 10.1200/JCO.2004.XX12345. (Conference Paper)
[5] Schouten LJ, Rutten J, Huveneers HA, Twijnstra A. "Incidence of brain metastases in a cohort of patients with carcinoma of the breast, colon, kidney, and lung and melanoma." *Cancer*, 2002; 94:2698. doi: 10.1002/cncr.10552. (Journal Paper)
[6] Le Rhun, E. et al., "EANO–ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up of patients with brain metastasis from solid tumours," Annals of Oncology, Vol. 32, No. 11, pp. 1332-1347.
[7] Patki N, Wedge R, Veeramachaneni K. "The Synthetic Data Vault." *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016. Available at: https://dai.lids.mit.edu/sdv. (Accessed September 2024)
[8] Paesmans, M. "Breath", 2012; 9:112-121.DOI 10.1183/20734735.006911. (Journal Paper)
[9] Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P. "Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group." *American Journal of Clinical Oncology*, 1982; 5:649-655. (Guideline Document)
[10] Detterbeck, F.C. "The eighth edition TNM stage classification for lung cancer: What does it mean on main street?" *Journal of Thoracic Oncology*, 2017; 12(6):707-709. (Article)
[11] Grupo Español de Cáncer de Pulmón (GECP), "Cáncer de pulmón: Incidencia y factores de riesgo," Disponible en: https://www.gecp.org/cancer-de-pulmon-incidencia-y-factores-de-riesgo/ (acceso septiembre 2024).
[12] Sociedad Española de Oncología Médica (SEOM), "Infografía cáncer de pulmón REDECAN," Disponible en: https://seom.org/images/INFOGRAFIA_CANCER_PULMON_REDECAN.PDF (acceso septiembre 2024).
[13] Spandidos Publications. "The role of molecular pathways in clinical oncology ." A vailable at: www .spandidos-publications.com/10.3892/mco.2013.130. (Accessed September 2024). (URL link)
[14] Sociedad Española de Neumología y Cirugía Torácica (SEPAR), "Vinculación entre EPOC y cáncer de pulmón," Nota de prensa (7 marzo 2022), Disponible en: https://www.separ.es/sites/default/files/SEPAR%20NP%20Vinculaci%C3%B3n%20entre%20EPOC%20y%20c%C3%A1ncer%20de%20pulm%C3%B3n_%287%20mar%2022%29.pdf (acceso septiembre 2024).
[15] Asociación Española de Afectados por Cáncer de Pulmón (AEACaP), "Biomarcadores," 2021. Disponible en: https://afectadoscancerdepulmon.com/biomarcadores/ (acceso abril 2022).