Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Inteligencia artificial

Predicción de la respuesta al tratamiento del cáncer de pulmón mediante redes neuronales artificiales

Trabajo fin de estudio presentado por:	Anna Gracia Colmenarejo
Tipo de trabajo:	Desarrollo Software
Director/a:	Lucía Prieto Santamaría
Codirector/a:	Almudena Ruiz Iniesta
Fecha:	10/07/2024

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mi tutora, Lucía Santamaría Prieto, por su orientación, apoyo constante y valiosos consejos durante todo el desarrollo de este trabajo de fin de máster. Su dedicación y compromiso han sido fundamentales para completar esta investigación. Agradezco también al *National Cancer Institute* de Estados Unidos por proporcionar los datos del estudio *PLCO* a través del *Cancer Data Access System*, que han sido esenciales para llevar a cabo este estudio. A mi madre y a mis amigos, por su apoyo incondicional, comprensión y paciencia a lo largo de este proceso.

Resumen

Este trabajo de fin de máster se centra en el desarrollo de un modelo predictivo de aprendizaje profundo basado en redes neuronales multicapa para realizar una clasificación binaria de la respuesta al tratamiento en pacientes con cáncer de pulmón, específicamente cáncer de pulmón de células no pequeñas. Se utilizan los datos del estudio Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) proporcionados por el Cancer Data Access System (CDAS) del National Cancer Institute (NCI) de Estados Unidos. El modelo predictivo se entrena con una amplia variedad de datos clínicos, incluyendo información de diagnóstico, resultados de pruebas, datos demográficos, hábitos y antecedentes personales y familiares de los pacientes. El objetivo del modelo es predecir la probabilidad de mortalidad del paciente a causa del cáncer de pulmón permitiendo evaluar la efectividad de los tratamientos realizados. El modelo ha alcanzado una exactitud de 0.71, lo que demuestra su eficacia y sugiere una prometedora línea de investigación futura.

Palabras Clave: cáncer de pulmón, cáncer de pulmón de células no pequeñas, NSCLC, redes neuronales, PLCO.

Abstract

This master's thesis focuses on the development of a predictive deep learning model based on multilayer neural networks to perform a binary classification of the treatment response in patients with lung cancer, specifically non-small cell lung cancer. The study uses data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) provided by the Cancer Data Access System (CDAS) of the National Cancer Institute (NCI) of the United States. The predictive model is trained with a wide variety of clinical data, including diagnostic information, test results, demographic data, habits, and personal and family histories of the patients. The model aims to predict the probability of patient mortality due to lung cancer, allowing for the evaluation of the effectiveness of the treatments administered. The model has achieved an accuracy of 0.71, demonstrating its effectiveness and suggesting a promising direction for future research.

Keywords: lung cancer, non-small cell lung cancer, NSCLC, neural networks, PLCO.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

Cada año, la Sociedad Estadounidense del Cáncer estima el número de nuevos casos de cáncer y muertes en los Estados Unidos. En 2024, se prevé que se producirán 2,001,140 nuevos casos y 611,720 muertes por cáncer en los Estados Unidos. Además, posiciona el cáncer de pulmón como uno de los más comunes tanto en hombre como mujeres y la principal causa de muerte por cáncer en Estados Unidos. Para este año, se predicen 234,580 nuevos casos y 125,070 muertes por cáncer de pulmón (Siegel et al., 2024).

En respuesta a la creciente carga de la enfermedad y la necesidad de mejorar el índice de supervivencia, surge una nueva forma de enfrentar el cáncer la medicina personalizada. La medicina personalizada consiste precisamente en tratar cada tumor de manera única y distintiva, en lugar de usar tratamientos más genéricos. Para esto, es necesario conocer de la forma más precisa posible qué alteraciones, mutaciones o desequilibrios sufren las células que conforman ese cáncer y cuáles son las que le permiten crecer sin control alguno (*ECUSA News & Views*, 2023).

En este contexto, la inteligencia artificial, y específicamente el aprendizaje profundo, ha emergido como una herramienta poderosa para analizar grandes conjuntos de datos clínicos, genéticos y moleculares en busca de patrones y correlaciones complejas. El aprendizaje profundo se basa principalmente en redes neuronales artificiales profundas que pueden aprender automáticamente características relevantes de los datos y realizar predicciones precisas sobre la respuesta al tratamiento en pacientes con cáncer de pulmón. Se presentan como una gran revolución para todos los campos en los que se apliquen debido a su gran capacidad de trabajar volúmenes muy elevados de datos y extraer información de ellos, tarea que para un ser humano podría resultar muy costosa.

Este trabajo se propone un modelo redes neuronales artificiales enfocado en la predicción de la respuesta al tratamiento en pacientes con cáncer de pulmón de células no pequeñas. Al comprender mejor los factores que influyen en la respuesta individual al tratamiento, se espera mejorar los resultados clínicos y la calidad de vida de los pacientes, allanando el camino hacia una medicina más precisa y personalizada en el tratamiento del cáncer de pulmón.

1.1. Motivación

El cáncer de pulmón de células no pequeñas representa un desafío clínico y científico significativo debido a su alta incidencia, agresividad y limitada eficacia de los tratamientos disponibles. Su complejidad molecular y su constante evolución dificultan la obtención de nuevos tratamientos y la personalización de estos.

La medicina personalizada, que busca adaptar los tratamientos a las características individuales de cada paciente, ha surgido como una prometedora estrategia para mejorar los resultados clínicos en la oncología.

La convergencia de la medicina personalizada con los avances en aprendizaje profundo ofrece nuevas oportunidades para abordar este desafío. El crecimiento exponencial de las técnicas y las aplicaciones de las redes neuronales artificiales pueden revolucionar la forma en que se llega a un diagnóstico y al tratamiento de enfermedades. Las redes neuronales tienen la capacidad de analizar grandes volúmenes de datos clínicos, genéticos y moleculares para identificar patrones complejos y predecir respuestas individuales al tratamiento.

Este Trabajo de Fin de Máster tiene como objetivo explorar el potencial de las redes neuronales en la predicción de la respuesta al tratamiento recibido en pacientes con cáncer de pulmón con el fin de avanzar hacia una medicina más precisa y personalizada.

La relevancia de este enfoque radica en su capacidad para mejorar los resultados clínicos y revolucionar la toma de decisiones en el ámbito médico. Al proporcionar predicciones más precisas sobre la eficacia de diferentes opciones terapéuticas, se puede reducir el tiempo y los recursos dedicados a tratamientos subóptimos, así como evitar la progresión de la enfermedad y sus consecuencias asociadas.

1.2. Estructura del trabajo

En el apartado 2 de esta memoria se proporciona el contexto necesario para este trabajo, abarcando aspectos relacionados con el cáncer de pulmón incluyendo sus diferentes tipos, factores de riesgo, métodos de diagnóstico y opciones de tratamiento. Además, se realiza un análisis del estado del arte sobre estudios previos que han aplicado técnicas de inteligencia artificial en investigaciones médicas sobre el cáncer de pulmón.

En el apartado 3 se establecen los objetivos específicos que este trabajo pretende alcanzar.

En el apartado 4, se presenta el proceso de análisis y preparación de los datos utilizados en el modelo, así como la implementación del modelo predictivo propuesto y su posterior evaluación.

Finalmente, en el apartado 5, se presentan las conclusiones obtenidas del trabajo.

2. Contexto y estado del arte

En este capítulo, se proporcionará un análisis del contexto y el estado actual de la investigación en el ámbito del cáncer de pulmón y las redes neuronales.

2.1.Contexto del problema

El Instituto Nacional de Salud (NIH) de los Estados Unidos define el cáncer como una enfermedad por la que algunas células del cuerpo se empiezan a multiplicar sin control alguno. Es una enfermedad genética y los cambios en los genes que controlan el funcionamiento de cómo se forman y multiplican las células, causan el cáncer. Existen más de 100 tipos diferentes de cáncer que llevan el nombre del órgano tejido en el que se forman.

El cáncer de cada persona es una combinación única de cambios genéticos. A medida que el cáncer sigue creciendo, ocurren otros cambios. Incluso dentro del mismo tumor, es posible que las diversas células tengan cambios genéticos distintos.

2.1.1. Cáncer de pulmón

El cáncer de pulmón es un tipo de cáncer que se origina en las células que revisten los bronquios y otras partes de los pulmones. Los diagnósticos de cáncer de pulmón se dividen principalmente en dos grupos: cáncer de pulmón de células pequeñas (SCLC) y cáncer de pulmón de células no pequeñas (NSCLC). El SCLC suele comenzar en los bronquios, las vías respiratorias que conducen desde la tráquea hacia los pulmones, y se caracteriza por su rápido crecimiento y diseminación a otras partes del cuerpo, incluidos los ganglios linfáticos. El NSCLC representa la mayoría de los diagnósticos de cáncer de pulmón y generalmente crece a un ritmo más lento que el SCLC. Este tipo de cáncer se desarrolla lentamente y suele presentar pocos o ningún síntoma hasta estar en una etapa avanzada. (Ciupka, 2020)

Este tipo de cáncer, como muchos otros, se asocia a diversos factores de riesgo, algunos modificables y otros no. Entre los factores modificables, el más importante es el tabaquismo, responsable de aproximadamente el 80% de las muertes por cáncer de pulmón. Pero no toda la gente que fuma padece este cáncer, respirar el humo de otras personas, la exposición al gas radón, al asbesto y a otros agentes carcinógenos en el ambiente también incrementan el riesgo de desarrollar esta enfermedad (American Cancer Society, 2024). Otros factores

incluyen la contaminación del aire y el consumo de ciertos suplementos dietéticos, como el betacaroteno en fumadores, que puede aumentar el riesgo en lugar de reducirlo.

Por otro lado, existen factores de riesgo no modificables, como la exposición previa a radioterapia en el tórax, el historial personal o familiar de cáncer de pulmón, y la edad avanzada. La radioterapia previa en el pecho, especialmente en personas tratadas por otros tipos de cáncer, también incrementa el riesgo. La genética también juega un papel en familias con una fuerte historia de cáncer de pulmón.

La forma más eficaz de reducir el riesgo de cáncer de pulmón es evitar el consumo de tabaco y la exposición al humo "de segunda mano". Dejar de fumar antes de que se desarrolle un cáncer permite que el tejido pulmonar dañado se repare gradualmente, reduciendo significativamente el riesgo, sin importar la edad o la duración del hábito. Es importante minimizar la exposición al gas radón mediante la evaluación y tratamiento de los niveles de radón en el hogar o lugar de trabajo. Adoptar una dieta saludable, rica en frutas y verduras, puede contribuir a reducir el riesgo de cáncer de pulmón, aunque su efecto protector es considerablemente menor en comparación con el daño causado por fumar.

El cáncer de pulmón suele ser asintomático hasta que se ha diseminado, aunque algunas personas con etapas tempranas pueden presentar síntomas. Consultar al médico al notar los primeros síntomas puede facilitar un diagnóstico precoz, lo que incrementa la eficacia del tratamiento. Los síntomas comunes del cáncer de pulmón incluyen una tos persistente o que empeora, esputo con sangre, dolor en el pecho, ronquera, pérdida del apetito, pérdida de peso inexplicada, dificultad para respirar, fatiga, infecciones respiratorias recurrentes y aparición de sibilancias. Si el cáncer se disemina a otras partes del cuerpo, puede causar dolor óseo, alteraciones neurológicas, ictericia o hinchazón de ganglios linfáticos. (American Cancer Society, 2024)

El diagnostico se realiza mediante imágenes extraídas de radiografías del tórax, resonancia magnética y tomografías computarizadas, mediante biopsias del tejido del pulmón y mediante muestras de esputo o sangre. Generalmente, el diagnóstico comienza con una radiografía de tórax, donde los tumores pulmonares suelen aparecer como masas blancas o grisáceas. No obstante, esta prueba no puede proporcionar un diagnóstico definitivo, ya que no siempre distingue entre el cáncer y otras afecciones, como los abscesos pulmonares. Si la radiografía sugiere la presencia de cáncer, se refiere al paciente a un especialista para pruebas

adicionales. La siguiente prueba habitual es una tomografía computarizada (TC), que utiliza rayos X y una computadora para crear imágenes detalladas del interior del cuerpo. Si los resultados de la TC indican cáncer, se puede realizar una tomografía por emisión de positrones (PET-TC), que muestra áreas con células cancerosas activas para ayudar en el diagnóstico y la planificación del tratamiento. Si la TC sugiere cáncer en la parte central del pecho, se puede realizar una broncoscopia, que permite al médico observar las vías respiratorias y tomar una muestra de células (biopsia). En algunos casos, se realiza una ecografía endobronquial (EBUS), una técnica más avanzada que combina la broncoscopia con una ecografía para localizar los ganglios linfáticos y obtener biopsias (*Lung Cancer - Diagnosis*, 2017).

El tratamiento del cáncer de pulmón varía según el tipo de cáncer y su grado de diseminación. En el caso del cáncer de pulmón de células no pequeñas, los pacientes pueden ser tratados mediante cirugía, quimioterapia, radioterapia, terapia dirigida o una combinación de estas opciones. Por otro lado, los pacientes con cáncer de pulmón de células pequeñas (SCLC) suelen recibir tratamiento con radioterapia y quimioterapia.

2.1.1.1. Cáncer de pulmón de células no pequeñas (NSCLC)

El cáncer de pulmón de células no pequeñas es el tipo de cáncer de pulmón más común. Entre el 80% y el 85% de los cánceres de pulmón son de este tipo. Es una enfermedad por la que se forman células malignas (cancerosas) en los tejidos del pulmón (*Tratamiento del cáncer de pulmón de células no pequeñas*, 2024). Fumar es el factor de riesgo principal de este tipo de cáncer de pulmón.

Además, existen diferentes tipos dentro de este que vienen nombrados por los tipos de células cancerosas que se multiplican en cada uno. Los principales subtipos de cáncer de pulmón son el adenocarcinoma, el carcinoma de células escamosas y el carcinoma de células grandes. Estos subtipos, que se originan a partir de diferentes tipos de células pulmonares, se agrupan como cáncer de pulmón de células no pequeñas porque su tratamiento y pronóstico suelen ser similares (American Cancer Society, 2024).

El adenocarcinoma está frecuentemente localizado en la periferia del pulmón, asociado a las glándulas que secretan moco. El carcinoma de células escamosas se encuentra habitualmente en la región central del pulmón, adyacente a los bronquios principales, y asociado al tabaquismo. Y, por último, el carcinoma de células grandes que puede originarse en cualquier

región del pulmón, destacándose por su heterogeneidad celular y su potencial de crecimiento rápido (*Tratamiento del cáncer de pulmón de células no pequeñas*, 2024).

El sistema de estadificación que se emplea con más frecuencia para el cáncer de pulmón no microcítico (NSCLC) es el sistema TNM del American Joint Committee on Cancer (AJCC) que se basa en tres piezas clave de información: tamaño y extensión del tumor principal (T), propagación a los ganglios (nódulos) linfáticos adyacentes (N) y la propagación (metástasis) a sitios distantes (M) (American Cancer Society, 2019). Esta clasificación divide el NSCLC en cuatro estadios principales:

- **Estadio I**: indica un tumor localizado en el pulmón, sin evidencia de diseminación a los ganglios linfáticos regionales o a otros órganos distantes. El estadio I se subdivide en estadio IA y IB.
- **Estadio II**: se caracteriza por la presencia de un tumor de mayor tamaño o por la invasión a estructuras adyacentes, o la afectación de los ganglios linfáticos cercanos, pero sin metástasis a distancia. Se subdivide en estadio IIA y IIB.
- **Estadio III**: es este el cáncer se ha diseminado a los ganglios linfáticos mediastínicos o estructuras adyacentes, pero sin evidencia de metástasis a órganos distantes. Se subdivide en estadio IIIA, IIIB y IIIC.
- **Estadio IV**: Representa la fase más avanzada del NSCLC, en la que hay metástasis a distancia. Este estadio se divide en IVA y IVB.

La estadificación precisa del NSCLC es fundamental para la planificación del tratamiento, que puede incluir cirugía, radioterapia, quimioterapia, inmunoterapia o una combinación de estos métodos, dependiendo del estadio del cáncer al momento del diagnóstico (Ettinger et al., 2022).

2.1.1.2. Tratamientos

Para combatir el cáncer, existen distintos tratamientos. Los principales son la quimioterapia, la radioterapia, el uso de cirugías, la inmunoterapia y terapias dirigidas o combinadas. Sin embargo, decidir un tratamiento para un paciente depende de diferentes factores como el tipo, el estadio y la salud del paciente, entre otros.

La quimioterapia utiliza medicamentos para destruir las células cancerosas, deteniendo su capacidad de crecer y dividirse. Puede ser administrada por vía oral, intravenosa, o directamente en la cavidad corporal afectada (*Non-Small Cell Lung Cancer Treatment - NCI*, 2024). La quimioterapia puede administrarse de diferentes formas dependiendo del tipo y estadio del cáncer, y puede ser sistémica o regional. Se utiliza para curar el cáncer, reducir su tamaño antes de la cirugía o radioterapia, eliminar células cancerosas restantes después de otros tratamientos o aliviar síntomas en casos avanzados. Debido a que también puede afectar a las células sanas de crecimiento rápido (como las del cabello), la quimioterapia puede causar efectos secundarios como pérdida de cabello, náuseas, fatiga, y mayor riesgo de infecciones.

La radioterapia utiliza radiación de alta energía (rayos X) para destruir las células cancerosas. Puede ser externa o interna. La elección del tipo de radioterapia depende del tipo, estadio y localización del cáncer (*Non-Small Cell Lung Cancer Treatment - NCI*, 2024). Al igual que la quimioterapia, puede ser utilizada sola o en combinación con otros tratamientos para curar el cáncer, reducir su tamaño antes de la cirugía, eliminar células cancerosas restantes o paliar síntomas. Los efectos secundarios dependen de la parte del cuerpo tratada, e incluyen irritación de la piel, fatiga, y problemas en el área tratada, como dificultad para tragar si se trata el cuello.

La cirugía implica la extracción física del tumor y, a veces, de los tejidos circundantes y los ganglios linfáticos. Existen varias técnicas quirúrgicas, desde la cirugía abierta tradicional hasta procedimientos mínimamente invasivos como la laparoscopia y la cirugía robótica. La elección del tipo de cirugía depende del tamaño y la ubicación del tumor, así como de la capacidad pulmonar del paciente (*Non-Small Cell Lung Cancer Treatment - NCI*, 2024). Puede ser utilizada para diagnosticar, estadiar, y tratar el cáncer. En algunos casos, la cirugía puede ser suficiente para eliminar el cáncer por completo. Dependiendo del tipo de cirugía y de la parte del cuerpo operada, los efectos secundarios pueden incluir dolor, infección, y pérdida de función en el área tratada.

La inmunoterapia estimula el sistema inmunológico del paciente para que reconozca y destruya las células cancerosas. Este enfoque emplea sustancias producidas por el cuerpo o sintetizadas en un laboratorio para potenciar, dirigir o restaurar las defensas naturales del organismo contra las células cancerosas (*Non-Small Cell Lung Cancer Treatment - NCI*, 2024). Existen diferentes tipos de inmunoterapia, incluyendo inhibidores de puntos de control

inmunitario, terapias con células T, y vacunas contra el cáncer. Pueden ser utilizadas para tratar diversos tipos de cáncer, especialmente aquellos que no responden bien a otros tratamientos. Aunque la inmunoterapia puede causar menos efectos secundarios que la quimioterapia, aún puede provocar respuestas autoinmunes, donde el sistema inmunológico ataca tejidos sanos, causando inflamación y otros síntomas.

La terapia dirigida utiliza medicamentos diseñados para identificar y atacar específicamente las células cancerosas sin dañar significativamente las células normales. Estas terapias se basan en las características genéticas del cáncer y pueden ser muy eficaces en ciertos tipos de cáncer. En el tratamiento del cáncer de pulmón de células no pequeñas avanzado, metastásico o recurrente, se utilizan principalmente cuatro tipos de terapias dirigidas: anticuerpos monoclonales, inhibidores de tirosina quinasa, inhibidores del mTOR e inhibidores de KRAS G12C (*Non-Small Cell Lung Cancer Treatment - NCI*, 2024). Los efectos secundarios pueden incluir diarrea, problemas hepáticos, y reacciones cutáneas, pero generalmente son menos severos que los de la quimioterapia.

A menudo, los tratamientos para el cáncer combinan varias de las opciones anteriores para aumentar su eficacia. La combinación de tratamientos puede aumentar los efectos secundarios, y los médicos deben equilibrar cuidadosamente los beneficios y los riesgos.

2.1.2. Aprendizaje automático

El aprendizaje automático por definición se presenta como aquel que, tras solucionar un problema, reconoce la situación problemática ante la cual ofrecer la solución aprendida. Este aprendizaje autónomo debe ofrecer la habilidad de dar una respuesta adecuada y modificarla según las condiciones variables. Siendo así necesario que el agente (inteligencia o sistema) sea capaz de adaptarse de forma dinámica y sin entrenamiento previo a las situaciones cambiantes.

En el contexto de la inteligencia artificial, el aprendizaje automático es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos capaces de aprender de los datos y mejorar su rendimiento de manera autónoma con el tiempo. A diferencia de los sistemas tradicionales de programación, donde se siguen instrucciones explícitas para cada tarea, el aprendizaje automático permite que los sistemas aprendan patrones y relaciones a partir de grandes volúmenes de datos, de manera que puedan ofrecer

soluciones adecuadas a problemas previamente desconocidos. Este tipo de aprendizaje se caracteriza por su capacidad de reconocer situaciones problemáticas y adaptar sus respuestas de forma dinámica, sin requerir entrenamiento previo para cada posible escenario. El objetivo es que el sistema sea capaz de generalizar a partir de los ejemplos observados, tomando decisiones basadas en la experiencia acumulada y ajustando continuamente su estrategia de acuerdo con las condiciones variables del entorno (Mitchell, 2013).

Según Goodfellow (Ian Goodfellow et al., 2016), estos algoritmos pueden ser supervisados, no supervisados o de refuerzo, dependiendo de si el modelo aprende a partir de ejemplos etiquetados, datos no etiquetados, o mediante la interacción con un entorno, respectivamente.

2.1.2.1. Aprendizaje automático supervisado

El aprendizaje automático supervisado es una de las técnicas más comunes en el campo del aprendizaje automático, donde el modelo se entrena utilizando un conjunto de datos etiquetados, es decir, datos en los que las respuestas correctas ya se conocen. Durante el entrenamiento, el modelo aprende a mapear las entradas (características o atributos) a las salidas correctas (etiquetas), con el objetivo de minimizar el error de predicción y mejorar su capacidad de generalización para nuevos datos (Bishop, 2006). Este tipo de aprendizaje es ampliamente utilizado en aplicaciones prácticas, como reconocimiento de voz, predicción de mercados financieros, diagnóstico médico, y muchas otras áreas donde la precisión en la predicción es fundamental.

En el aprendizaje supervisado, se abordan principalmente dos tipos de problemas:

- **Problemas de clasificación:** el objetivo es asignar una etiqueta a una nueva observación en función de las características proporcionadas. El modelo aprende a separar las diferentes clases basándose en los datos de entrenamiento proporcionados. Los modelos de clasificación se utilizan para predecir categorías discretas o clases (R. Tibshirani, Friedman, Jerome, et al., 2001).
- **Problemas de regresión:** el objetivo es predecir un valor numérico continuo a partir de las características de entrada. En este caso, el modelo aprende una función que aproxima la relación entre las variables de entrada y el valor de salida continuo. Por ejemplo, se puede utilizar para predecir el precio de una casa basándose en

características como el tamaño, la ubicación y el número de habitaciones (Gareth et al., 2022)

2.1.2.2. Problemas de clasificación

Existen diferentes tipos de problemas de clasificación según el número de clases y la naturaleza de los datos:

- **Clasificación binaria**: implica distinguir entre dos posibles clases. Es el tipo de problema de clasificación más sencillo y común. Ejemplos típicos incluyen la detección de spam (correo electrónico clasificado como "spam" o "no spam"), diagnóstico médico (presencia o ausencia de una enfermedad), y la predicción de aprobación de crédito (aprobado o rechazado) (Murphy, 2012).
- **Clasificación multiclase**: problemas en los que hay más de dos posibles clases. Por ejemplo, la clasificación de imágenes en categorías como "gato", "perro" o "pájaro", o la categorización de tipos de cáncer en base a datos histológicos. En estos problemas, el modelo debe aprender a distinguir entre múltiples clases que pueden tener características muy similares (R. Tibshirani, Friedman, Jerome, et al., 2001).
- **Clasificación multietiqueta**: cada instancia de entrada puede pertenecer a más de una clase simultáneamente. Este tipo de problemas requiere modelos que sean capaces de manejar la correlación entre múltiples etiquetas y proporcionar una salida para cada una de ellas.
- **Clasificación desbalanceada**: problemas en los que la distribución de las clases no es uniforme; una clase puede ser mucho más frecuente que otra. Un ejemplo es la detección de fraudes financieros, donde las transacciones fraudulentas son mucho menos frecuentes que las transacciones legítimas. O un fallo en una cadena de producción en una fábrica.

2.1.2.3. Aprendizaje profundo: Redes Neuronales

El aprendizaje profundo (*deep learning* en inglés) es una subdisciplina del aprendizaje automático que utiliza redes neuronales artificiales con múltiples capas para aprender representaciones de datos a diferentes niveles de abstracción. Este enfoque ha ganado popularidad debido a su capacidad para modelar relaciones complejas y no lineales, y su éxito

en aplicaciones como reconocimiento de voz, procesamiento de imágenes, traducción automática y, más recientemente, en el ámbito médico y de salud (LeCun et al., 2015).

Las redes neuronales artificiales (*ANN*, por sus siglas en inglés) son modelos computacionales inspirados en la estructura y el funcionamiento del cerebro humano. Están compuestas por unidades de procesamiento llamadas neuronas, que están organizadas en capas: una capa de entrada, una o más capas ocultas, y una capa de salida. Cada neurona recibe entradas, las procesa mediante una función de activación, y transmite la salida a las neuronas de la siguiente capa (Ian Goodfellow et al., 2016).

En el contexto del *deep learning*, las redes neuronales profundas se diferencian por tener muchas capas ocultas interconectadas, lo que permite que el modelo aprenda características a múltiples niveles de abstracción. Este tipo de arquitecturas es especialmente útil para tareas complejas, como el reconocimiento de patrones en datos de alta dimensionalidad, como imágenes o secuencias temporales.

Las principales arquitecturas de redes neuronales son:

- **Redes Neuronales Convolucionales (CNN):** están diseñadas específicamente para procesar datos con una estructura de cuadrícula, como imágenes. Utilizan capas convolucionales para extraer características locales mediante filtros que se aplican a lo largo de la imagen, detectando patrones como bordes, texturas o formas específicas. Esto las hace altamente efectivas en tareas de visión por computadora, como la clasificación de imágenes, detección de objetos y segmentación (Krizhevsky et al., 2012).
- **Redes Neuronales Recurrentes (RNN)**: están diseñadas para manejar datos secuenciales, donde el orden de los datos importa, como en el procesamiento de texto o señales de tiempo. Utilizan conexiones recurrentes que permiten que la salida de una neurona se retroalimente como entrada para la misma o una neurona anterior, permitiendo que la red conserve "memoria" de los pasos anteriores (Hochreiter & Schmidhuber, 1997).
- **Redes Generativas Antagónicas (GAN)**: compuestas por dos redes neuronales que compiten entre sí: una red generadora, que intenta crear datos falsos que se asemejen a los datos reales, y una red discriminadora, que intenta distinguir entre datos reales y generados. Este enfoque ha sido revolucionario en la generación de imágenes,

videos, y otros tipos de datos sintéticos, así como en la mejora de datos de entrenamiento para otros modelos (Goodfellow et al., 2014).

Las redes neuronales profundas han demostrado un rendimiento sobresaliente en tareas donde los datos son complejos, no lineales y de alta dimensionalidad. Su capacidad para aprender representaciones jerárquicas permite una mayor generalización en comparación con los métodos tradicionales de aprendizaje automático. Sin embargo, su éxito depende de la disponibilidad de grandes cantidades de datos y de un poder computacional significativo para entrenar modelos eficaces, lo que puede ser una limitación en algunos contextos.

2.1.2.4. Elementos de una red neuronal

Una red neuronal artificial está compuesta por varios elementos fundamentales que colaboran para procesar la información y aprender de los datos. Estos elementos incluyen los siguientes:

- **Neurona**: es la unidad básica de procesamiento de una red neuronal, inspirada en las neuronas biológicas del cerebro. Cada neurona recibe una o más entradas, las procesa mediante una función de activación, y produce una salida que se transmite a las neuronas de la siguiente capa. Las neuronas en la red están conectadas entre sí de forma que cada salida de una neurona puede ser la entrada de otra.
- **Pesos**: son parámetros que determinan la influencia de una entrada específica en la neurona. Cada conexión entre dos neuronas tiene un peso asociado que se ajusta durante el entrenamiento de la red. Los pesos son esenciales para que la red neuronal aprenda, ya que estos parámetros se ajustan iterativamente para minimizar el error de las predicciones del modelo.
- **Sesgo (bias)**: es un término adicional que se suma a la entrada ponderada de una neurona antes de aplicar la función de activación. Este parámetro permite que la red modele funciones más complejas y desplaza la función de activación a la izquierda o a la derecha, ayudando a que el modelo se ajuste mejor a los datos de entrenamiento.
- **Función de activación**: introduce no linealidades en la red, permitiendo que aprenda y modele relaciones complejas. Entre las funciones de activación más comunes se encuentran la función sigmoide, la tangente hiperbólica (tanh), y la rectificada lineal unitaria (ReLU). Estas funciones transforman la suma ponderada de las entradas y el

sesgo en una salida que puede ser utilizada por las neuronas en la siguiente capa (Ian Goodfellow et al., 2016).

- **Capas**: una red neuronal está organizada en capas. La capa de entrada recibe los datos de entrada; las capas ocultas procesan esta información mediante la combinación de pesos, sesgos y funciones de activación; y la capa de salida proporciona la predicción o clasificación final del modelo. Las redes neuronales profundas contienen múltiples capas ocultas, lo que permite al modelo aprender características jerárquicas y complejas de los datos.

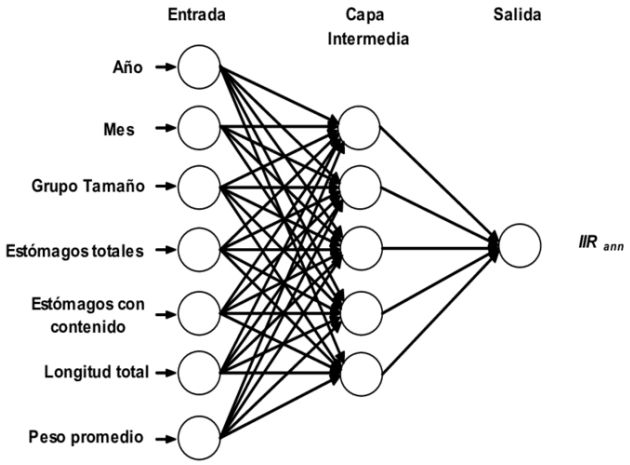


Ilustración 1. Ejemplo de red neuronal simple

Fuente: ResearchGate (Contreras, 2008)

2.1.2.5. Entrenamiento de redes neuronales

El entrenamiento de redes neuronales es el proceso mediante el cual una red ajusta sus parámetros internos (pesos y sesgos) para aprender a realizar tareas específicas, como la clasificación de imágenes, el reconocimiento de voz o la predicción de series temporales. Este proceso se lleva a cabo utilizando un conjunto de datos de entrenamiento, que proporciona ejemplos de entradas y sus correspondientes salidas deseadas. El objetivo del entrenamiento es minimizar la diferencia entre las salidas predichas por la red y las salidas reales, a través de un proceso iterativo de ajuste de parámetros (Ian Goodfellow et al., 2016).

Al comenzar el entrenamiento, los pesos de la red neuronal se inicializan de manera aleatoria o utilizando técnicas específicas de inicialización que ayudan a mejorar la convergencia del modelo, como la inicialización de He o Xavier (Glorot & Bengio, 2006). En cada iteración de entrenamiento, las entradas del conjunto de datos de entrenamiento se pasan a través de la

red (propagación hacia delante, en inglés, *forward propagation*). En esta fase, las neuronas calculan sus salidas utilizando una función de activación que introduce no linealidades, permitiendo que la red aprenda relaciones complejas entre las variables de entrada y las salidas deseadas. Las salidas finales de la red se comparan con las salidas esperadas para calcular el error (diferencia entre la predicción y el valor real) utilizando una función de pérdida o coste (Nielsen, 2015). El siguiente paso es la retropropagación (*backpropagation*). Este proceso comienza calculando el gradiente del error con respecto a cada peso de la red, utilizando el método de diferenciación automática. Luego, estos gradientes se propagan hacia atrás a través de la red, desde la capa de salida hasta la capa de entrada. Este procedimiento permite actualizar los pesos de la red en la dirección que minimiza el error (Rumelhart et al., 1986). Tras calcular los gradientes, los pesos de la red se ajustan utilizando un optimizador, siendo el más común el algoritmo de descenso de gradiente. Este método ajusta los pesos en la dirección opuesta al gradiente del error, con un tamaño de paso determinado por la tasa de aprendizaje.

El entrenamiento de una red neuronal se realiza en varias épocas, donde en cada época la red ve todo el conjunto de datos de entrenamiento una vez. A través de múltiples épocas, el modelo ajusta progresivamente sus pesos para minimizar la función de pérdida. El proceso continúa hasta que el modelo alcanza un rendimiento satisfactorio o hasta que se cumple un criterio de parada, como un número máximo de épocas o una mejora mínima del error. El proceso de entrenamiento es iterativo y requiere una combinación de técnicas avanzadas de optimización, selección de hiperparámetros adecuados y estrategias para mitigar problemas comunes, lo cual es fundamental para asegurar que el modelo de red neuronal aprenda de manera efectiva y generalice bien en nuevos datos.

2.1.2.6. Problemas comunes en redes neuronales

Las redes neuronales, aunque poderosas, enfrentan varios problemas comunes que pueden afectar su rendimiento y eficacia. Uno de los principales desafíos es el sobreajuste (*overfitting*), que ocurre cuando la red aprende demasiado bien los detalles y el ruido del conjunto de datos de entrenamiento, lo que resulta en un modelo que funciona bien en datos de entrenamiento, pero tiene un bajo rendimiento en datos nuevos o no vistos. Otro problema frecuente es el subajuste (*underfitting*), donde la red es demasiado simple para capturar las relaciones complejas entre las variables de entrada y salida, generalmente debido a un

modelo con pocas neuronas, capas o una falta de entrenamiento adecuado (Ian Goodfellow et al., 2016).

Otro desafío importante es el desvanecimiento o explosión del gradiente, especialmente en redes neuronales profundas, donde los gradientes pueden volverse extremadamente pequeños o grandes durante el proceso de retropropagación, dificultando que los pesos de las capas más profundas se actualicen de manera efectiva. Este problema afecta particularmente a las redes neuronales recurrentes durante el entrenamiento en secuencias largas (Hochreiter & Schmidhuber, 1997). Además, la selección de hiperparámetros como la tasa de aprendizaje, el número de neuronas, o el número de capas, puede ser complicada y requiere un ajuste cuidadoso para evitar la convergencia lenta o la falta de convergencia del modelo (Bengio, 2012).

También se presentan problemas de interpretabilidad, ya que las redes neuronales, especialmente las profundas, suelen ser consideradas "cajas negras", dificultando la comprensión de cómo el modelo toma decisiones y, por lo tanto, reduciendo la confianza de los usuarios en su aplicabilidad, especialmente en áreas sensibles como la medicina o las finanzas (Samek et al., 2017).

2.1.2.7. Evaluación de redes neuronales

La evaluación de redes neuronales es un proceso crucial para entender la efectividad y el rendimiento de un modelo en tareas de predicción. Esta evaluación se realiza utilizando diversas métricas y técnicas que ayudan a determinar cómo de bien una red neuronal está funcionando con respecto a los objetivos específicos del problema que se está abordando. Para medir el correcto comportamiento de las redes (y otros algoritmos de inteligencia artificial) existen diferentes métricas de evaluación:

- **Exactitud (*accuracy*)**: una de las métricas más comunes y mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Aunque es útil para tener una idea general del rendimiento, su utilidad se reduce en casos de desbalance de clases, donde una alta exactitud puede ser engañosa si las clases están desigualmente representadas (Ian Goodfellow et al., 2016).
- **Precisión (*precision*):** ratio de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo. La precisión es particularmente útil cuando el costo

de los falsos positivos es alto, como en aplicaciones médicas donde un diagnóstico incorrecto puede llevar a consecuencias graves.

- **Sensibilidad (*recall*):** Es la proporción de verdaderos positivos identificados correctamente sobre el total de verdaderos positivos existentes. Esta métrica es especialmente importante cuando es crítico minimizar los falsos negativos, como en el diagnóstico de enfermedades graves (Murphy, 2012).
- **F1 *Score*:** Es la media armónica de la precisión y la sensibilidad, proporcionando un equilibrio entre ambas métricas. Es especialmente útil cuando existe un desbalance de clases, ya que combina las ventajas de la precisión y la sensibilidad en un solo valor (Powers, 2020).
- **Área Bajo la Curva ROC (ROC-AUC)**: mide la capacidad del modelo para distinguir entre clases. Un AUC cercano a 1 indica un excelente rendimiento del modelo, mientras que un AUC de 0.5 sugiere que el modelo no es mejor que un clasificador aleatorio (Fawcett, 2006).

Evaluar correctamente las redes neuronales es fundamental para asegurar que el modelo pueda generalizar bien a nuevos datos. Además, permite ajustar hiperparámetros como el número de capas, el tamaño del lote o la tasa de aprendizaje, entre otros, para mejorar la eficiencia y la precisión del modelo en la tarea específica que se está abordando.

2.1.3. Herramientas técnicas

Existen diversas herramientas con las que realizar la creación de un modelo predictivo, implementar redes neuronales, analizar los datos y la evaluación de los resultados obtenidos. A continuación, se detallan las principales herramientas:

- **Python** es un lenguaje de programación gratuito de uso general, pero que es la herramienta principal en ciencia de datos y en la implementación de modelos de inteligencia artificial. Se ha convertido en la primera opción para desarrollar estos algoritmos debido a su agilidad, flexibilidad y sobre todo, por la gran cantidad de bibliotecas disponibles para la realización de esta tarea.
- **Jupyter Notebook** es un entorno interactivo en que se crean documentos de código, visualizaciones y texto. Se utiliza para desarrollar algoritmos a la vez que se documentan los experimentos realizados.

- **Scikit-Learn** es una librería de aprendizaje automático en Python que proporciona herramientas simples y eficientes para el modelado predictivo, incluyendo algoritmos de clasificación, regresión, modelado estadístico, selección de características y validación cruzada.
- **Pandas** es una librería de Python especializada en el tratado y análisis de datos. Proporciona estructuras de datos de alto rendimiento, llamados DataFrames, que permiten manipular, filtrar, agregar y limpiar datos de manera sencilla y eficiente.
- **Tensorflow** es una librería de código abierto de aprendizaje automático, especialmente utilizada para aprendizaje profundo. El nombre deriva de las operaciones que las redes neuronales realizan sobre vectores multidimensionales de datos a los que se refiere como tensores.
- **Keras** es una librería especializada en la creación de redes neuronales de código abierto escrita en Python. Al usar Keras, los desarrolladores pueden definir rápida y fácilmente las estructuras de las redes ya que funciona como una interfaz de alto nivel dentro de TensorFlow.

2.2. Estado del arte

El artículo "Deep learning predicts lung cancer treatment response from serial medical imaging" (Xu et al., 2019) destaca la aplicación del aprendizaje profundo en la predicción de la respuesta al tratamiento del cáncer de pulmón a partir de imágenes médicas seriadas de tomografía computarizada. Utilizan un modelo de red neuronal convolucional que aprende características relevantes y patrones complejos de las imágenes como la textura del tumor, el tamaño o la forma. Sus resultados mostraron que el modelo es capaz de predecir con alta precisión la respuesta al tratamiento a partir de imágenes e incluso, superar otros enfoques tradicionales.

Por otro lado, el artículo “Modelo de predicción de la respuesta al tratamiento de quimio-radioterapia en pacientes con cáncer de pulmón de células no pequeñas localmente avanzado irresecable mediante la aplicación de radiómica en imágenes de TC” (Rozalen et al., 2023) se centró en el desarrollo de un modelo predictivo de la respuesta al tratamiento de quimio-radioterapia en pacientes con cáncer de pulmón de células no pequeñas. El estudio incluyó a pacientes del Servicio de Oncología Radioterápica del Hospital Universitario Ramón y Cajal, diagnosticados con CPCNP en estadios IIIA, IIIB o IIIC, quienes recibieron tratamiento completo

con quimio-radioterapia entre 2015 y 2022. Para la selección de pacientes se aplicaron criterios estrictos, como la necesidad de al menos dos ciclos de quimioterapia basada en platino y una dosis mínima de radioterapia de 60 Gy. Se recopilaron datos de imágenes TC obtenidas con protocolos estandarizados y segmentadas con software especializado para definir la región de interés del tumor. A continuación, se construyeron siete modelos de clasificación, incluidos Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes Gaussiano (GNB), Random Forest (RF), Gradient Boosting (GB), Regresión Logística (LR) y Perceptrón Multicapa (MLP). Los modelos predictivos mostraron resultados prometedores, donde el Perceptrón Multicapa (MLP) alcanzó un área bajo la curva (AUC) de 0.90, y la Regresión Logística (LR) un AUC de 0.88. El estudio resalta el potencial de la radiómica y el aprendizaje automático para predecir la respuesta al tratamiento, permitiendo un enfoque más personalizado y potencialmente menos invasivo en la gestión del cáncer de pulmón, al identificar de manera anticipada a los pacientes más propensos a responder a ciertos tratamientos. Sin embargo, también reconoce limitaciones, como el tamaño relativamente pequeño de la muestra y la posible heterogeneidad de los datos, sugiriendo la necesidad de estudios adicionales con cohortes más amplias (Rozalen et al., 2023).

El siguiente artículo: "Assessing eligibility for lung cancer screening using parsimonious ensemble machine learning models: A development and validation study" (Callender et al., 2023) presenta un estudio sobre modelos de aprendizaje automático de ensamblado para determinar la elegibilidad en el cribado de cáncer de pulmón. Este estudio tiene como objetivo desarrollar modelos de predicción simples pero efectivos que mantengan la precisión de los modelos más complejos, facilitando la adopción generalizada de programas de cribado personalizado. Para el desarrollo del modelo, se utilizaron datos de 216,714 fumadores del Reino Unido y 26,616 fumadores de alto riesgo de los Estados Unidos. Los modelos predijeron dos resultados a 5 años: el diagnóstico de cáncer de pulmón y la muerte por cáncer de pulmón. El estudio demuestra que los modelos de aprendizaje automático de ensamblado, con un número reducido de predictores, pueden simplificar la implementación del cribado de cáncer de pulmón basado en el riesgo. Sin embargo, una limitación importante es que estos modelos no han sido validados fuera de las cohortes del Reino Unido y EE. UU. Se requiere más validación en diferentes contextos antes de su implementación generalizada.

En la misma línea de investigación, el artículo: “OWL: an optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations” (Pan et al., 2023) se centra en el desarrollo y validación de un modelo optimizado de predicción de riesgo de cáncer de pulmón llamado OWL. Este modelo fue desarrollado utilizando la técnica de *machine learning XGBoost* con datos de un gran conjunto de participantes del Biobanco del Reino Unido y fue validado de manera independiente con datos de otros estudios, incluidos PLCO y *National Lung Screening Trial* (NLST). OWL mostró un rendimiento discriminativo superior en comparación con otros modelos de predicción de riesgo haciendo ver que es un avance significativo en la predicción del riesgo de cáncer de pulmón, demostrando una precisión y utilidad clínica mejoradas en comparación con los modelos existentes.

Por último, el artículo "Explainable Machine Learning for Lung Cancer Screening Models" (Kobylińska et al., 2022) aborda la creciente necesidad de modelos de aprendizaje automático que sean explicables y comprensibles, especialmente en el contexto del cribado de cáncer de pulmón. Este estudio aplica técnicas de inteligencia artificial explicable (XAI) para analizar y comparar tres modelos utilizados comúnmente para predecir el riesgo de cáncer de pulmón: BACH, PLCOm2012 y LCRAT. El estudio utiliza un conjunto de datos del *Domestic Lung Cancer Database* que incluye más de 34,000 pacientes de la población polaca. Los resultados muestran que las diferentes técnicas de XAI permiten entender qué variables son más importantes en cada modelo, cómo influyen en las predicciones y cómo se transforman estos datos en resultados predictivos. Por ejemplo, variables como la edad, los años fumados y el tiempo desde que se dejó de fumar tienen diferentes niveles de importancia en los modelos BACH, PLCOm2012 y LCRAT. El uso de técnicas XAI puede mejorar significativamente la comprensión de los modelos de cribado de cáncer de pulmón, facilitando su adopción en la práctica clínica. Esta mayor transparencia es clave para una mejor colaboración entre los algoritmos de aprendizaje automático y los profesionales de la salud, lo que puede llevar a un enfoque más personalizado y efectivo en la detección temprana del cáncer de pulmón.

2.3. Conclusiones

En conjunto, los estudios revisados muestran varias oportunidades y desafíos en el desarrollo de modelos predictivos para el cáncer de pulmón. Mientras que los enfoques basados en aprendizaje profundo y modelos optimizados como OWL muestran un gran potencial en

términos de precisión y aplicabilidad clínica, también se destaca la necesidad de que estos modelos sean interpretables y estén validados en diversas poblaciones para asegurar su efectividad en contextos reales.

Este trabajo pretende aprovechar estos avances aplicando redes neuronales al *dataset* PLCO para contribuir a mejorar las pautas de tratamiento y la medicina personalizada en el cáncer de pulmón. Una mejora en el tratamiento del paciente puede impactar positivamente en la mejora de la supervivencia de uno de los tipos de cáncer más mortales, mediante la optimización de la toma de decisiones clínicas basada en datos, y avanzar hacia un enfoque más preciso y personalizado en la atención del paciente.

3. Objetivos concretos y metodología de trabajo

3.1. Objetivo general

El objetivo general de este trabajo es desarrollar un modelo predictivo basado en redes neuronales para predecir la respuesta al tratamiento en pacientes con cáncer de pulmón de células no pequeñas, utilizando los datos del estudio PLCO proporcionados por el CDAS. Se espera obtener una métrica de precisión superior del 0.60.

3.2. Objetivos específicos

Los objetivos específicos son:

1. Analizar y preparar los datos del estudio PLCO para su uso en la red neuronal, incluyendo la limpieza, normalización, y selección de variables predictoras relevantes que puedan influir en la respuesta al tratamiento.
2. Diseñar y desarrollar un modelo de redes neuronales capaz de aprender de los datos clínicos, demográficos y de pruebas diagnósticas de los pacientes, para predecir la probabilidad de respuesta al tratamiento.
3. Entrenar y optimizar el modelo mediante el uso de técnicas avanzadas de aprendizaje automático, como el ajuste de hiperparámetros y la validación cruzada, para mejorar su precisión y generalización.
4. Evaluar el rendimiento del modelo utilizando métricas estándar como la precisión, el área bajo la curva (AUC), la sensibilidad y la especificidad, para validar su efectividad en la predicción de la respuesta al tratamiento.
5. Analizar los resultados obtenidos del modelo predictivo, interpretando su desempeño y comparándolo con otros enfoques existentes en la literatura, para identificar sus ventajas y limitaciones y proponer posibles mejoras.

3.3. Metodología del trabajo

Para el desarrollo de este trabajo se ha utilizado el método *Cross-Industry Standard Process for Data Mining* (CRISP-DM), un estándar ampliamente aceptado en el campo de la ciencia de datos y la minería de datos que representa el ciclo de vida de los proyectos de análisis de datos. Fue desarrollada en 1996 por un consorcio de empresas con el objetivo de crear un marco común y flexible que pudiera ser aplicado en diversos dominios y tipos de datos.

Esta metodología proporciona un marco estructurado y flexible que consta de seis fases iterativas. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario (*SPSS Modeler Subscription*, 2021).

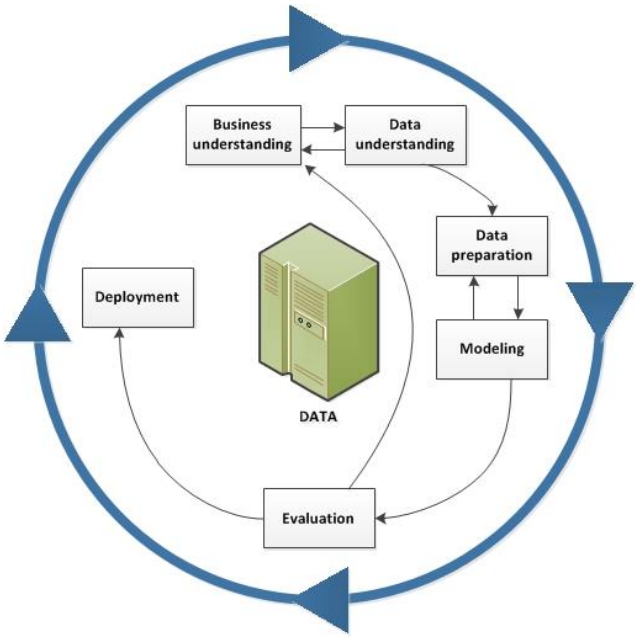


Ilustración 2. Ciclo de vida del método CRISP-DM

Fuente: IMB (*SPSS Modeler Subscription*, 2021)

A continuación, se explican las seis fases del modelo CRISP-DM:

1. **Comprensión del negocio**: es la primera fase del proyecto, se centra en comprender los objetivos y requisitos del proyecto desde la perspectiva del negocio. Se define el problema que se desea resolver, los objetivos del análisis, las necesidades de los *stakeholders*, y los criterios de éxito del proyecto y también se crea un plan preliminar de los pasos a seguir. Es crucial desarrollar una comprensión profunda del contexto del negocio para alinear los resultados del proyecto con los objetivos organizacionales.
2. **Comprensión de los datos**: se recopilan los datos necesarios para el proyecto y se realiza una exploración preliminar para entender su contenido, calidad y relevancia. El objetivo es obtener una comprensión clara de los datos disponibles y su potencial para alcanzar los objetivos establecidos en la fase anterior.
3. **Preparación de los datos**: en esta etapa se construye el *dataset* final, a partir de los datos obtenidos en la fase anterior, que será utilizado en el modelado. Esta fase suele ser la más larga del proceso, ya que implica tareas de limpieza de datos, selección de características relevantes, creación de nuevas variables, codificación de variables

categóricas, normalización o estandarización de los datos, y manejo de valores faltantes. El objetivo es asegurar que los datos estén en un formato adecuado y de alta calidad para alimentar los algoritmos de modelado.

4. **Modelado**: En esta fase se seleccionan y aplican las técnicas de modelado más apropiadas para el problema en cuestión. Implica el desarrollo de uno o más modelos predictivos utilizando algoritmos de aprendizaje automático o minería de datos. Se ajustan los parámetros de los modelos para optimizar su rendimiento, y se prueban diferentes enfoques para identificar el más eficaz. El modelado es un proceso iterativo que puede requerir volver a fases anteriores para mejorar los resultados.
5. **Evaluación**: se debe evaluar los resultados utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto (*SPSS Modeler Subscription*, 2021).Para ello, se utilizan métricas de rendimiento adecuadas, como precisión, sensibilidad, especificidad, área bajo la curva (AUC). Además, se revisa el proceso en su totalidad para verificar que se han cumplido todos los criterios de éxito definidos en la fase de comprensión del negocio. Esta evaluación permite determinar si el modelo es suficientemente bueno para ser desplegado o si requiere ajustes adicionales.
6. **Despliegue**: la fase final del método consta en la poner el modelo en un entorno de producción. También incluye la documentación del proyecto, la transferencia de conocimiento a los usuarios finales, y la definición de planes de mantenimiento y monitoreo para asegurar que el modelo siga siendo efectivo a futuro.

4. Desarrollo específico de la contribución

4.1. Comprensión del negocio

En la fase de comprensión del negocio se llevaron a cabo varias investigaciones preliminares para definir los objetivos y el enfoque del proyecto. En primer lugar, se realizó una investigación sobre el uso de técnicas de aprendizaje profundo en la detección, diagnóstico, medicina personalizada y predicción de respuesta al tratamiento en el cáncer de pulmón. Esta revisión permitió identificar las aplicaciones actuales de las redes neuronales en el ámbito oncológico, sus ventajas, limitaciones y las áreas donde se podría mejorar la precisión y la personalización del tratamiento.

Asimismo, fue necesario realizar un estudio detallado sobre esta enfermedad, sus diferentes tipos (especialmente el cáncer de pulmón de células no pequeñas), los factores de riesgo, los métodos de diagnóstico y los tratamientos disponibles. Esta comprensión básica fue crucial para establecer un marco adecuado para la aplicación de técnicas de aprendizaje automático y para asegurar que el enfoque del modelo fuera clínicamente relevante y alineado con las necesidades reales de los pacientes de cáncer de pulmón.

Como resultado de estas investigaciones, se establecieron los objetivos del proyecto, que se han detallado en los apartados 3.1 y 3.2, centrados en el desarrollo de un modelo predictivo de redes neuronales que permita mejorar la precisión en la predicción de la respuesta al tratamiento en pacientes con cáncer de pulmón de células no pequeñas, utilizando los datos del estudio PLCO proporcionados por el CDAS.

4.2. Comprensión de los datos

Tras una búsqueda exhaustiva de bases de datos adecuadas para el desarrollo del proyecto, se decidió trabajar con el conjunto de datos del estudio PLCO. El conjunto de datos PLCO es parte de un estudio longitudinal y multicéntrico iniciado por el Instituto Nacional del Cáncer de Estados Unidos con el objetivo de investigar la efectividad de las técnicas de cribado en la reducción de la mortalidad por cáncer de próstata, pulmón, colorrectal y ovario. Los datos no son de acceso abierto, ya que contiene información confidencial sobre pacientes; por lo tanto, es necesario solicitar un permiso especial para su uso en investigaciones. Para este trabajo, se

obtuvo la aprobación correspondiente a través del CDAS del NCI, bajo el número de proyecto PLCO-1542.

Este conjunto fue seleccionado porque proporciona información detallada sobre pacientes con cáncer de pulmón, incluyendo datos sobre el diagnóstico, tratamientos administrados, procedimientos médicos realizados, mortalidad, antecedentes médicos, y otras variables relevantes que permiten estudiar la respuesta al tratamiento.

El estudio PLCO incluye información de aproximadamente 150,000 pacientes, tanto hombres como mujeres, de 55 a 74 años que padecen cáncer de próstata, pulmón, colon y/o ovario. El conjunto de datos está compuesto por varios archivos CSV, cada uno de los cuales contiene información específica y detallada sobre diferentes aspectos relacionados con los pacientes incluidos en el estudio.

A continuación, se explica la información que contiene cada archivo CSV del estudio:

- **lung_data_mar22_d032222.csv - General**: es el *dataset* de información general donde cada columna presenta una característica sobre los pacientes como la edad, sexo, raza, hábitos de fumar, presencia de enfermedades preexistentes, historial familiar de cáncer, entre otras. Este *dataset* es crucial para entender las características demográficas y clínicas de los pacientes y sus posibles factores de riesgo.
- **lung_proc_data_mar22_d032222.csv – Procedimientos de diagnóstico**: incluye detalles de los procedimientos diagnósticos realizados en los pacientes, tales como biopsias, tomografías computarizadas (CT), resonancias magnéticas (MRI), y otras pruebas diagnósticas utilizadas para evaluar la presencia y el estado del cáncer de pulmón.
- **lung_med_data_mar22_d032222.csv – Complicaciones médicas**: recoge información sobre las complicaciones médicas que los pacientes han experimentado durante el estudio como reacciones alérgicas, infecciones, fiebres, dolores, neumonías, etc.
- **lung_screen_data_mar22_d032222.csv - Cribado**: contiene datos de las pruebas de cribado realizadas a los pacientes, como radiografías de tórax, tomografías computarizadas de baja dosis, y otros procedimientos utilizados para la detección temprana del cáncer de pulmón.

- **lung_scrsub_data_mar22_d032222.csv – Anomalías de cribado**: registra los hallazgos anómalos detectados durante los procedimientos de cribado, proporcionando información sobre lesiones sospechosas, nódulos pulmonares, y otras anormalidades que podrían indicar la presencia de cáncer.
- **lung_trt_data_mar22_d032222.csv - Tratamientos**: incluye datos detallados sobre los diferentes tratamientos recibidos por los pacientes, como quimioterapia, radioterapia, inmunoterapia, los días que se ha tardado en recetar, y otros enfoques terapéuticos. Este *dataset* es esencial para analizar la relación entre los tratamientos aplicados y los resultados clínicos de los pacientes.

Para cada *dataset* se realizó un análisis profundo extrayendo así información importante sobre posibles variables predictoras, posibles variables objetivo y valores nulos.

4.2.1. *Dataset* General

El *dataset* general contiene 154,887 entradas o filas y 251 variables o columnas que proporcionan información detallada sobre los pacientes participantes en el estudio. Cada paciente está anonimizado y es identificado de forma única mediante la variable ***PLCO_id***, lo que garantiza la privacidad y confidencialidad de los datos. Mediante el análisis de esta variable, se ha verificado que no existen entradas duplicadas en este *dataset*, lo que asegura la integridad de los datos para su uso en el modelo predictivo.

Como se ha mencionado anteriormente, este estudio no es exclusivo de cáncer de pulmón por lo que los pacientes deben ser filtrados y solo seleccionar los que padezcan esta enfermedad. Este filtro se realiza con la variable ***lung_cancer*** que para pacientes enfermos de cáncer de pulmón tendrá valor de 1.

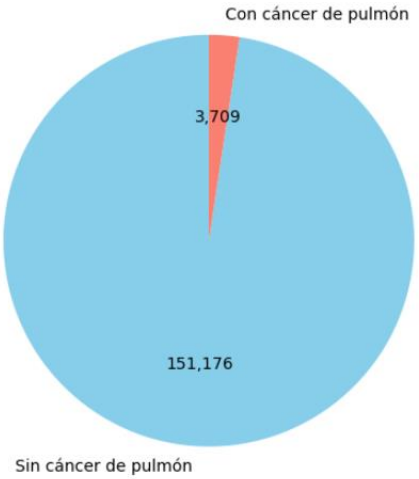


Ilustración 3. Comparativa de pacientes enfermos de cáncer de pulmón y pacientes sin cáncer de pulmón

Fuente: Elaboración propia con Python

De los 151,177 participantes del estudio 3,710 sufren cáncer de pulmón, estos valores muestran una reducción muy grande de los pacientes con los que se podrá entrenar el modelo.

Como este trabajo está centrado en los pacientes de NSCLC, de estos 3,710 se escogerán solo los que padezcan este subtipo de cáncer mediante la variable **lung_cancer_type**. Para pacientes con NSCLC valdrá 1 y para pacientes con SCLC valdrá 2.

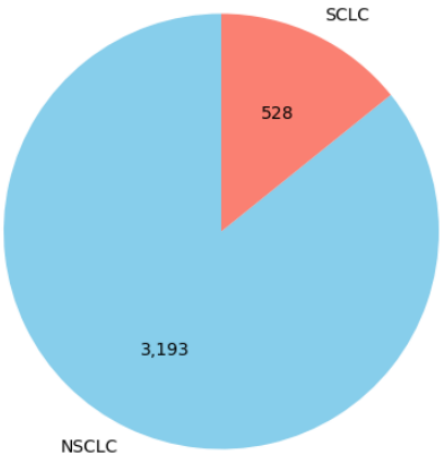


Ilustración 4. Comparativa de pacientes con NSCLC y SCLC

Finalmente, del *dataset* general se extraen 3,193 pacientes de cáncer de pulmón de células no pequeñas para entrenar y validar el modelo.

Dado que no todas las 251 variables son útiles para el modelo de predicción y serían demasiadas para definir una a una, se ha realizado un proceso de selección de las variables

más relevantes que podrían potencialmente influir en la respuesta al tratamiento del cáncer de pulmón. El criterio para mantener estas variables ha sido que la información que aporten sea relevante para el modelo, que no sea redundante con otras variables y que el número de nulos sea menor del 20% del total de pacientes con NSCLC. Por lo tanto, según estos criterios las variables seleccionadas son las que a continuación se enumeran: *num_cancl, intstatl_cat, lung_stage, lung_topography, lung_grade, lung_histtype_cat, curative_pneuml, curative_wsll, curative_chemol, curative_radl, neoadjuvantl, lung_cancer_first, lung_exitdays, lung_cancer_diagdays, biopllink0-3, educat, marital, occupat, pipe, cigar, sisters, brothers, asp, ibup, xray_history, asppd, ibuppd, hyperten_f, hearta_f, stroke_f, emphys_f, bronchit_f, diabetes_f, polyps_f, arthrit_f, osteopor_f, divertic_f, gallblad_f, race7, smoked_f, smokea_f, rsmoker_f, cigpd_f, filtered_f, cig_stat, cig_stop, cig_years, bmi_curc, weight_f, height_f, colon_comorbidity, liver_comorbidity, fh_cancer, lung_fh, lung_fh_cnt, d_dthl, mortality_exitage, ph_any_trial, ph_lung_trial, Sex, Agelevel, fstcan_exitstat, fstcan_exitage, fstcan_exitdays.*

4.2.2. *Dataset* Procedimientos de Diagnóstico

El *dataset* de procedimientos de diagnóstico presentó ciertos desafíos para su integración con el *dataset* general, debido a la presencia de múltiples entradas para un mismo paciente. Esta duplicación ocurría porque cada entrada correspondía a un procedimiento específico al que el paciente se había sometido. Por ejemplo, un mismo paciente podría tener varias filas en el *dataset* si se había realizado un examen físico, una biopsia, o cualquier otro procedimiento diagnóstico, lo que generaba múltiples registros para el mismo identificador de paciente (*PLCO_id*). Esta situación complicaba la fusión con el *dataset* general, donde cada paciente debe estar representado por una única entrada.

Para resolver este problema, se decidió transformar la variable categórica ***proc_catl***, que indica el tipo de procedimiento al que se había sometido cada paciente, en una serie de columnas binarias, donde cada columna representaba un tipo de procedimiento específico. Esta técnica, conocida como **codificación *one-hot*** o codificación de variables categóricas, permitió crear una estructura de datos en la que cada paciente tuviera una sola fila con múltiples columnas que indicaban si el paciente se había sometido o no a cada uno de los posibles procedimientos.

Además, para no perder la información contenida en la variable ***proc_res***, que proporciona el resultado de cada procedimiento, se utilizó una estrategia adicional: en cada nueva columna creada a partir de ***proc_catl***, se almacenó el resultado del procedimiento correspondiente. De este modo, cada columna representaba tanto la realización de un procedimiento específico como su resultado, lo que permitió conservar la riqueza de la información original sin duplicar registros. Los valores que puede adoptar la variable ***proc_res*** son de tipo categórico y son los siguientes: 1 (Negativo), 2 (Anómalo, no sospechoso), 3 (Anómalo, sospechoso de cáncer de pulmón), 4 (Anómalo, diagnóstico de cáncer de pulmón), 5 (Anómalo, sospechoso de metástasis), 6 (Anómalo, metástasis confirmada), 7 (insatisfactorio), 8 (Poco concluyente), 9 (Anómalo, no sospechoso de metástasis). Esta transformación facilitó la integración del *dataset* de procedimientos de diagnóstico con el *dataset* general, proporcionando un conjunto de datos consolidado y adecuado para el entrenamiento del modelo predictivo.

A continuación, se muestran las variables seleccionadas de los procedimientos de diagnóstico:

VARIABLE	**DESCRIPCIÓN**
proc_catl_1_res	Variable categórica que almacena el resultado del procedimiento 'Revisión de registros'.
proc_catl_2_res	Variable categórica que almacena el resultado del procedimiento 'Examen físico/evaluación clínica'.
proc_catl_3_res	Variable categórica que almacena el resultado del procedimiento 'Repeat Screen'.
proc_catl_4_res	Variable categórica que almacena el resultado del procedimiento 'Imaging'.

proc_catl_5_res	Variable categórica que almacena el resultado del procedimiento 'Endoscopias y endoscopios no invasivos'.
proc_catl_6_res	Variable categórica que almacena el resultado del procedimiento 'Endoscopias, biopsias y cirugías invasivas (con anestesia local)'.
proc_catl_7_res	Variable categórica que almacena el resultado del procedimiento 'Cirugías'.
proc_catl_8_res	Variable categórica que almacena el resultado del procedimiento 'Otros'.

Tabla 1. Variables seleccionadas del *dataset* de procedimientos de diagnostico

Fuente: Elaboración propia

4.2.3. *Datasets* Complicaciones Médicas, Cribado y Anomalías de Cribado

Los *datasets* de complicaciones médicas, cribado y anomalías de cribado no contenían un número suficiente de pacientes (<500 pacientes) con diagnóstico de cáncer de pulmón, lo que limitaba significativamente la cantidad de datos relevantes disponibles para su uso en el modelo predictivo. Debido a esta limitación, se determinó que las variables de este *dataset* no serían representativas ni aportarían suficiente valor al modelo, por lo que se decidió no incluirlas en el análisis final.

4.2.4. Dataset Tratamientos

El *dataset* de tratamientos presentó un problema similar al del *dataset* de procedimientos de diagnóstico en cuanto a la duplicación de entradas para un mismo paciente. Cada entrada en este *dataset* correspondía a un tratamiento específico recibido por el paciente, lo que resultaba en múltiples registros por paciente en caso de haber recibido varios tratamientos diferentes. Por ejemplo, un paciente podría haber recibido quimioterapia y radioterapia, generando así múltiples filas asociadas al mismo ***PLCO_id***.

Para abordar esta problemática, se utilizó una estrategia similar a la aplicada en el *dataset* de procedimientos de diagnóstico. La variable categórica ***trt_familyl***, que indica el tipo de tratamiento recibido por cada paciente, fue transformada mediante **codificación *one-hot*** para crear las columnas. Para conservar la información contenida en la variable númerica ***trt_days***, que refleja el número de días que tomó recibir cada tratamiento, se añadió esta información a las columnas correspondientes. Es decir, cada nueva columna creada a partir de ***trt_familyl*** no solo indicaba si el tratamiento fue administrado, sino también los días asociados a ese tratamiento en particular. De este modo, se preserva la información detallada sobre los tratamientos administrados y sus duraciones.

VARIABLE	**DESCRIPCIÓN**
trt_family_1_days	Variable numérica que indica el número de días en los que se tardó en recibir una neumonectomía o bilobectomía.
trt_family_2_days	Variable numérica que indica el número de días en los que se tardó en recibir '*wedge resection, segmental resection, or lobectomy.*
trt_family_3_days	Variable numérica que indica el número de días en los que se tardó en recibir radioterapia.
trt_family_4_days	Variable numérica que indica el número de días en los que se tardó en recibir quimioterapia.
trt_family_5_days	Variable numérica que indica el número de días en los que se tardó en recibir tratamiento no curativo.

Tabla 2. Variables seleccionadas del *dataset* de tratamientos

Fuente: Elaboración propia

4.3. Preparación de los datos

En esta fase, se realizó un proceso de preparación de los datos para asegurar que estuvieran en un formato adecuado para ser utilizados en el modelo predictivo. Este proceso incluyó la imputación de valores faltantes y la fusión de los diferentes *datasets*.

Para algunos pacientes, ciertas columnas contenían valores faltantes. Para manejar estos casos y evitar la pérdida de pacientes en el análisis, se optó por imputar los valores faltantes. En variables numéricas, los valores faltantes fueron reemplazados por la media de cada variable. Esta estrategia permite mantener la distribución original de los datos sin introducir sesgos significativos. Y, en variables categóricas, os valores faltantes fueron reemplazados por la moda (el valor más frecuente) de cada variable categórica. Este enfoque ayuda a preservar la tendencia más común en los datos categóricos sin añadir ruido significativo.

Tras la fusión de todos los *datasets* (general, procedimientos de diagnóstico y tratamientos), se obtuvo un *dataset* final compuesto por 81 columnas (características) y 2,758 filas (pacientes). Este *dataset* consolidado contiene una amplia gama de variables que incluyen información demográfica, clínica, de procedimientos de diagnóstico y tratamientos, proporcionando una base de datos integral para desarrollar el modelo predictivo.

Después de consolidar el *dataset*, se procedió a buscar una variable objetivo para predecir la efectividad del tratamiento. Dado que no existía una variable específica que indicara de manera directa si el tratamiento fue exitoso o no, se optó por utilizar la variable **d_dthl**, que es una variable binaria que indica si la muerte de un paciente fue a causa del cáncer de pulmón o no. La elección de esta variable se fundamenta en la premisa de que, si la muerte del paciente no fue atribuible al cáncer de pulmón, esto podría sugerir que el tratamiento aplicado fue eficaz o adecuado para controlar la enfermedad.

4.3.1. Análisis de la variable d_dthl

La variable d_dthl es una variable binaria que indica la causa de la muerte de los pacientes en el conjunto de datos final. Está codificada como:

- **0**: Muerte no causada por cáncer de pulmón.
- **1:** Muerte causada por cáncer de pulmón.

Al analizar la distribución de las clases de esta variable, se observa un desequilibrio significativo entre las dos categorías. Como se muestra en el gráfico, hay 1,940 pacientes cuya muerte no fue causada por cáncer de pulmón, en comparación con 818 pacientes cuya muerte sí fue causada por esta enfermedad.

Este desequilibrio en la distribución de clases puede tener importantes implicaciones para el entrenamiento del modelo predictivo. Los modelos de aprendizaje automático tienden a ser más precisos en la predicción de la clase mayoritaria cuando hay una distribución de clases desbalanceada. Para abordar este problema se pueden aplicar técnicas de manejo de datos desbalanceados, como el sobremuestreo de la clase minoritaria, el submuestreo de la clase mayoritaria o el uso de algoritmos de ajuste de peso que compensen este desequilibrio con el objetivo de mejorar la capacidad del modelo, para identificar correctamente los casos de muerte causada por cáncer de pulmón.

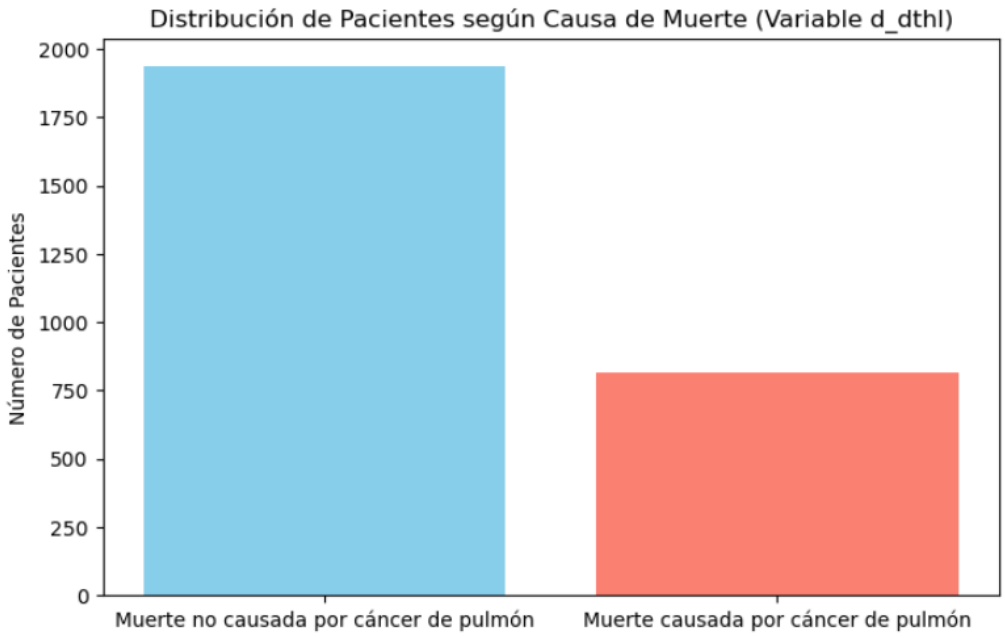


Ilustración 5. Distribución de la variable d_dthl

Fuente: Elaboración propia con Python

4.4. Modelado

Antes de proceder con la arquitectura diseñada se debe mencionar que debido al problema del desbalanceo de clases en la variable objetivo ***d_dthl***; se implementaron tres técnicas diferentes con el fin de identificar la estrategia más eficaz para mejorar el rendimiento del modelo predictivo:

- **Undersampling**: reduce la cantidad de ejemplos de la clase mayoritaria para equilibrar el número de ejemplos en cada clase.
- **Oversampling**: aumenta la cantidad de ejemplos de la clase minoritaria mediante la duplicación de muestras o la generación de nuevas muestras sintéticas.
- **Ajuste de peso**: asigna un peso mayor a la clase minoritaria en la función de pérdida, para que el modelo penalice más los errores cometidos en dicha clase durante el entrenamiento.

4.4.1. Arquitectura de la red

La red neuronal desarrollada está compuesta por diferentes capas y elementos diseñados específicamente para capturar las características complejas del conjunto de datos y proporcionar predicciones precisas.

4.4.1.1. Capas de la red y funciones de activación

La **capa de entrada** de la red neuronal consta de 81 neuronas, igual al número de características del *dataset* final después de la fusión. Esta capa recibe directamente los datos de entrada (características) y transmite esta información a las siguientes capas de la red. No realiza ningún procesamiento adicional más allá de la transmisión de datos; su función es simplemente aceptar la entrada y pasarla a las capas ocultas de la red.

Las **capas densas intermedias** son seis capas totalmente conectadas (*fully connected*) que permiten que la red neuronal aprenda patrones complejos en los datos. Estas capas están conectadas de manera que cada neurona de una capa recibe entradas de todas las neuronas de la capa anterior y envía su salida a todas las neuronas de la siguiente capa:

1. **Primera capa densa**: contiene 128 neuronas activadas con la función *ReLU* y regularizadas con L2.
2. **Segunda capa densa**: contiene 64 neuronas activadas con la función *ReLU* y regularizadas con L2.
3. **Tercera capa densa**: contiene 32 neuronas activadas con la función *ReLU* y regularizadas con L2.
4. **Cuarta, quinta y sexta capas densas**: cada una contiene 16 neuronas activadas con la función *ReLU* y regularizadas con L2.

La función *ReLU* es ampliamente utilizada en redes neuronales profundas porque introduce no linealidades que permiten al modelo aprender patrones complejos. *ReLU* convierte cualquier valor negativo de entrada en cero mientras deja pasar valores positivos sin cambios. Esto ayuda a mitigar el problema del desvanecimiento del gradiente y acelera la convergencia del modelo durante el entrenamiento (Nair & Hinton, 2010). Además, se aplica regularización L2 con un valor de 0.001 en cada capa densa para evitar el sobreajuste del modelo. La regularización L2 penaliza grandes pesos, favoreciendo soluciones más simples y generalizables al minimizar la magnitud de los pesos durante el entrenamiento (Ng, 2004).

La **capa de salida** de la red consta de una única neurona que utiliza la función de activación sigmoide (*sigmoid*). Esta función es ideal para problemas de clasificación binaria porque toma cualquier valor de entrada y lo convierte en un valor de probabilidad entre 0 y 1. Esto permite que la red neuronal indique la probabilidad de que la salida pertenezca a una clase específica (en este caso, si la muerte fue causada por cáncer de pulmón o no).

4.4.1.2. Compilación de la red

Para compilar el modelo, se utiliza el optimizador *Adaptive Moment Estimation* (*Adam*), que combina las ventajas de los algoritmos de descenso de gradiente estocástico (*SGD*) y el método de acumulación de gradientes (*RMSprop*). Adam ajusta automáticamente la tasa de aprendizaje para cada parámetro, lo que permite una convergencia más rápida y estable del modelo. Su eficacia se debe a su capacidad para adaptar las tasas de aprendizaje de forma independiente para cada parámetro, lo que es especialmente útil en modelos complejos como las redes neuronales profundas (Kingma & Ba, 2017).

Por otro lado, la función de pérdida utilizada es la *binary crossentropy*, adecuada para tareas de clasificación binaria. Esta función mide la diferencia entre las predicciones del modelo y las etiquetas reales, calculando la pérdida para cada instancia de entrenamiento. El objetivo del modelo es minimizar esta pérdida durante el entrenamiento, lo que se traduce en una mayor precisión de las predicciones (Murphy, 2012).

4.5. Evaluación del modelo

Para evaluar la eficacia del modelo de red neuronal propuesto, se evaluarán las tres técnicas aplicadas por el desbalanceo de clases por separado con las métricas de exactitud, precisión, sensibilidad, f1 score y la curva ROC-AUC. Esta evaluación comparativa permitirá identificar la

estrategia más adecuada para mejorar la capacidad del modelo de generalizar y hacer predicciones precisas en un escenario con clases desbalanceadas. Además, se realizará una evaluación en el conjunto de entrenamiento para detectar posibles signos de sobreajuste (overfitting) y asegurar que el modelo no esté memorizando los datos de entrenamiento, sino aprendiendo patrones generales aplicables a nuevos datos.

4.5.1. Pesos de las clases

Para esta primera estrategia, se evaluó el modelo inicialmente con el conjunto de prueba y se obtuvieron los siguientes resultados:

MÉTRICA	**RESULTADO**
Exactitud	0.73
Precisión	0.79
Sensibilidad	0.82
F1 score	0.80

Tabla 3. Resultados del conjunto de prueba. Variante: Pesos de las clases

Fuente: Elaboración propia

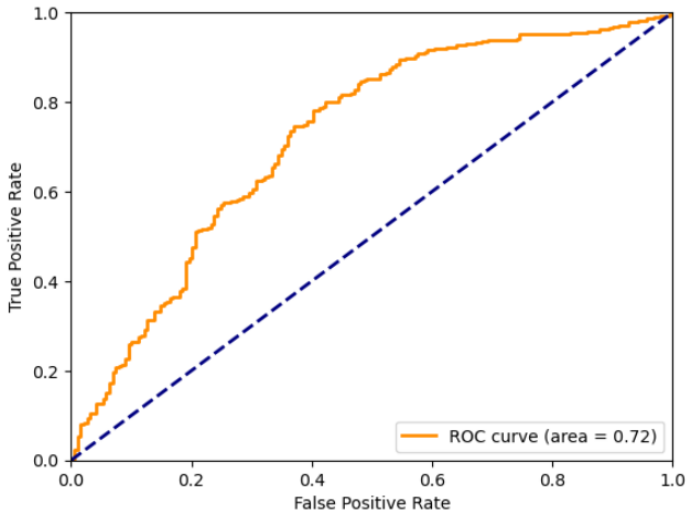


Ilustración 6. Curva ROC conjunto de prueba. Variante: Pesos de las clases

Fuente: Elaboración propia con Python

Estos resultados indican que el modelo tiene un rendimiento razonablemente bueno en el conjunto de prueba, con una precisión de 0.79 y una sensibilidad de 0.82. Esto sugiere que el modelo es capaz de identificar correctamente la mayoría de los casos positivos (muerte causada por cáncer de pulmón) y tiene un buen equilibrio entre precisión y sensibilidad, como lo refleja el F1 score de 0.80.

Para comprobar si el modelo presenta sobreajuste, se realizó una evaluación adicional en el conjunto de entrenamiento. Los resultados obtenidos fueron:

MÉTRICA	**RESULTADO**
Exactitud	0.98
Precisión	0.99
Sensibilidad	0.99
F1 score	0.99

Tabla 4. Resultados del conjunto de entrenamiento. Variante: Pesos de las clases

Fuente: Elaboración propia.

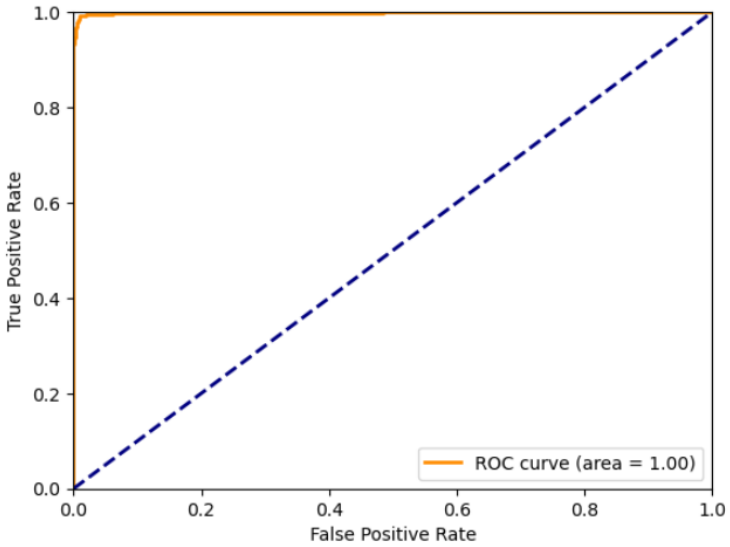


Ilustración 7. Curva ROC del conjunto de entrenamiento. Variante: Pesos de las clases

Fuente: Elaboración propia con Python

Estos valores extremadamente altos en el conjunto de entrenamiento, en comparación con los del conjunto de prueba, indican un posible sobreajuste del modelo. Esto significa que el

modelo ha aprendido muy bien los patrones específicos de los pacientes del conjunto de entrenamiento, pero puede no estar generalizando adecuadamente a los pacientes no vistos no vistos.

Para solucionar este problema se ha reducido el número de épocas a 12 y se ha empleado el uso de la técnica de *Dropout*. Esta es una técnica de regularización que ayuda a reducir el sobreajuste durante el entrenamiento, "apagando" aleatoriamente un porcentaje de neuronas en cada iteración de entrenamiento. En este caso se ha añadido un Dropout de 0.30 en todas las capas intermedias, lo que estará apagando un 30% de neuronas aleatorias en cada capa. Después de estos ajustes se han obtenidos los siguientes resultados:

MÉTRICA	**CONJUNTO DE PRUEBA**	**CONJUNTO DE ENTRENAMIENTO**
Exactitud	0.72	0.78
Precisión	0.82	0.86
Sensibilidad	0.74	0.83
F1 score	0.78	0.84

Tabla 5. Resultados obtenidos después de realizar Dropout y reducir el número de épocas. Variante: Pesos de las clases

Fuente: Elaboración propia.

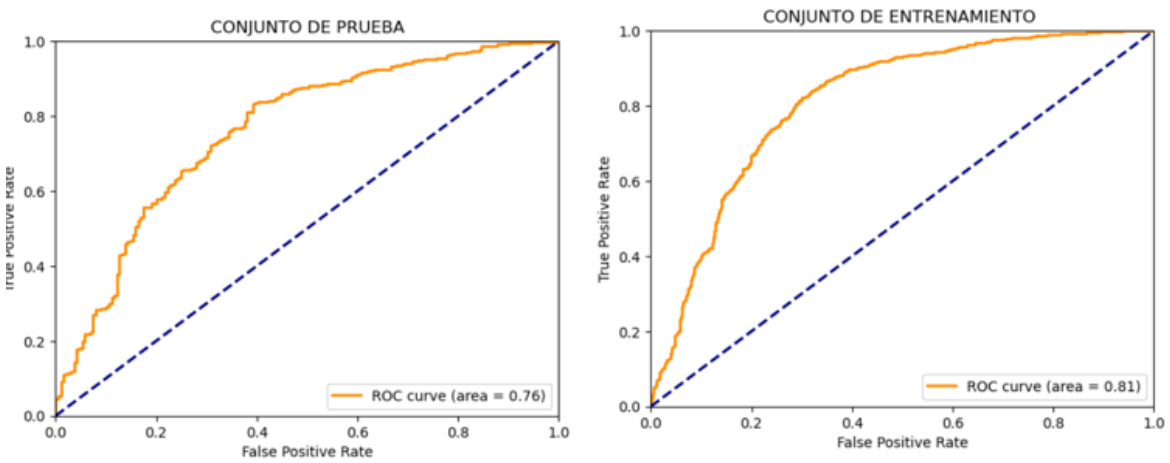


Ilustración 8. Comparativa de la curva ROC del conjunto de prueba y entrenamiento después de realizar Dropout y reducir el número de épocas. Variante: Pesos de las clases.

Fuente: Elaboración propia con Python

Estos resultados muestran una mejora en la capacidad del modelo para generalizar y un menor riesgo de sobreajuste después de los ajustes realizados. Aunque el rendimiento en el conjunto de entrenamiento es todavía ligeramente superior al del conjunto de prueba, las diferencias no son tan pronunciadas como en evaluaciones anteriores, lo que sugiere que las medidas implementadas han contribuido positivamente a mejorar la robustez del modelo.

En conclusión, el modelo obtuvo una exactitud de 0.72 en el conjunto de prueba, lo que indica que el 72% de las predicciones fueron correctas en pacientes no vistos. Con una precisión de 0.82 en el conjunto de prueba, el modelo tiene una buena capacidad para predecir correctamente los casos positivos (muerte causada por cáncer de pulmón) sin muchos falsos positivos. La sensibilidad en el conjunto de prueba es de 0.74, lo que significa que el modelo está identificando correctamente el 74% de los casos positivos reales. El F1 score en el conjunto de prueba es de 0.78, lo que refleja un buen equilibrio entre precisión y sensibilidad.

4.5.2. Oversampling

Para el caso de *oversampling*, se evaluó el modelo inicialmente con el conjunto de prueba y se obtuvieron los siguientes resultados:

MÉTRICA	**RESULTADO**
Exactitud	0.67
Precisión	0.73
Sensibilidad	0.81
F1 score	0.77

Tabla 6. Resultados del conjunto de prueba. Variante: *Oversampling*

Fuente: Elaboración propia.

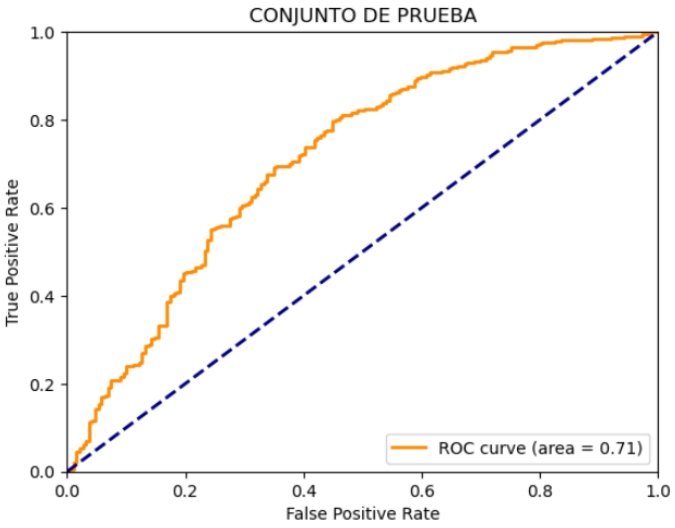


Ilustración 9. Curva ROC del conjunto de prueba. Variante: *Oversampling*

Fuente: Elaboración propia

Estos resultados indican que el modelo logró una exactitud del 0.67, lo que significa que el 67% de las predicciones realizadas fueron correctas en datos no vistos. Aunque la exactitud es moderada, la técnica de *oversampling* parece haber mejorado la sensibilidad, que es de 0.81, lo que implica que el modelo está capturando correctamente el 81% de los casos positivos (muertes causadas por cáncer de pulmón).

Para comprobar si el modelo presenta sobreajuste, se realizó una evaluación adicional en el conjunto de entrenamiento. Los resultados obtenidos fueron:

MÉTRICA	**RESULTADO**
Exactitud	0.97
Precisión	0.99
Sesibilidad	0.95
F1 score	0.97

Tabla 7. Resultados del conjunto de entrenamiento. Variante: *Oversampling*

Fuente: Elaboración propia

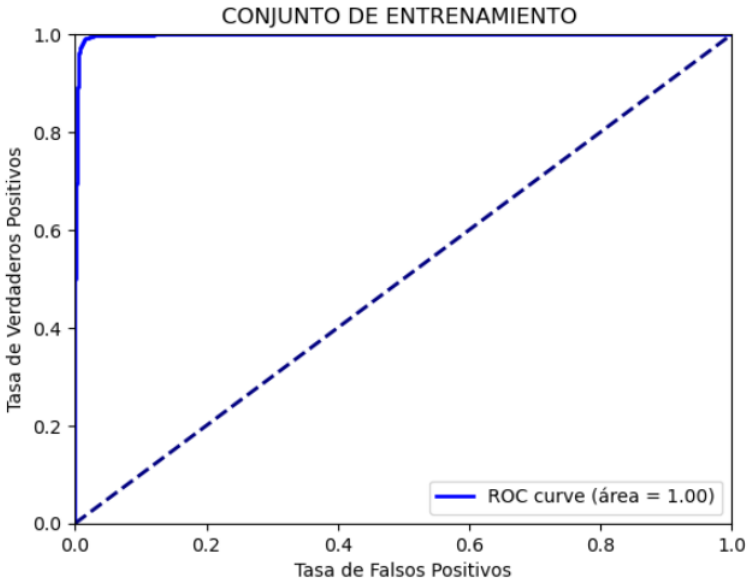


Ilustración 10. Curva ROC del conjunto de entrenamiento. Variante: *Oversampling*

Fuente: Elaboración propia con Python

Como en la estrategia anterior, estos resultados muestran que el modelo tiene un rendimiento muy alto en el conjunto de entrenamiento. La diferencia significativa entre los resultados obtenidos en el conjunto de entrenamiento y los del conjunto de prueba sugiere que el modelo podría estar presentando sobreajuste.

Para mitigarlo se ha seguido la técnica anterior, se ha añadido un *Dropout* de 0.3 en todas las capas intermedias y esta vez se ha reducido el número de épocas a 10.

Los resultados han sido:

MÉTRICA	**CONJUNTO DE PRUEBA**	**CONJUNTO DE ENTRENAMIENTO**
Exactitud	0.72	0.86
Precisión	0.76	0.84
Sensibilidad	0.85	0.89
F1 score	0.80	0.87

Tabla 8. Resultados después de realizar *Dropout* y reducir el número de épocas. Variante: *Oversampling*

Fuente: Elaboración propia

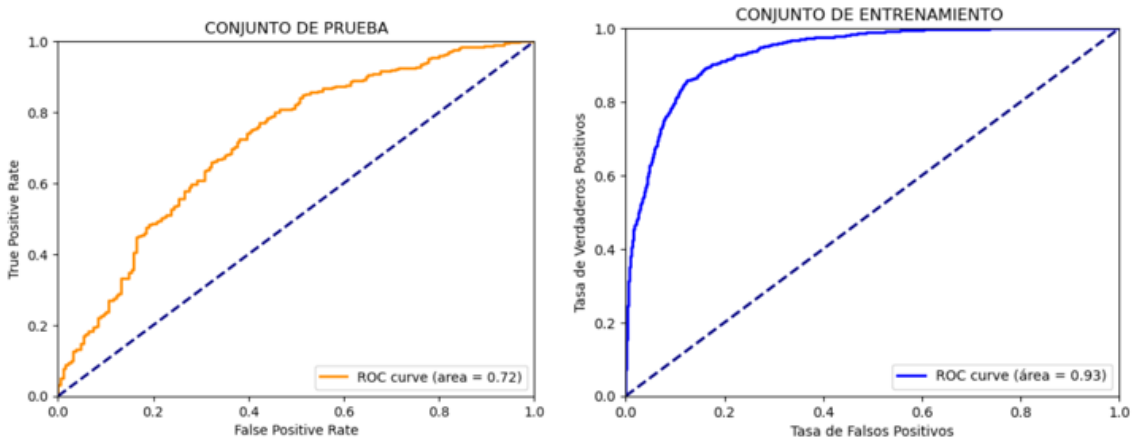


Ilustración 11. Comparativa de la curva ROC del conjunto de prueba y entrenamiento después de realizar *Dropout* y reducir el número de épocas. Variante: *Oversampling*.

Fuente: Elaboración propia

La exactitud del modelo en el conjunto de prueba aumentó a 0.72, lo que indica una mejora en la proporción de predicciones correctas en comparación con el caso anterior. En el conjunto de entrenamiento, la exactitud se redujo a 0.86, lo que sugiere una reducción del sobreajuste. La precisión en el conjunto de prueba es de 0.76, lo que refleja una capacidad aceptable para predecir correctamente los casos positivos sin generar muchos falsos positivos. La sensibilidad del modelo en el conjunto de prueba es alta, lo que indica que el modelo está capturando correctamente la mayoría de los casos positivos.

Los ajustes realizados, la adición de *Dropout* y la reducción del número de épocas, han ayudado a reducir el sobreajuste, como lo evidencia la reducción de la diferencia entre las métricas del conjunto de prueba y el conjunto de entrenamiento.

4.5.3. Undersampling

Finalmente, para el caso de *undersamplig*, se evaluó el modelo inicialmente con el conjunto de prueba y se obtuvieron los siguientes resultados:

MÉTRICA	**RESULTADO**
Exactitud	0.70
Precisión	0.85
Sensibilidad	0.70

F1 score	0.77

Tabla 9. Resultados del conjunto de prueba. Variante: *Undersampling*

Fuente: Elaboración propia

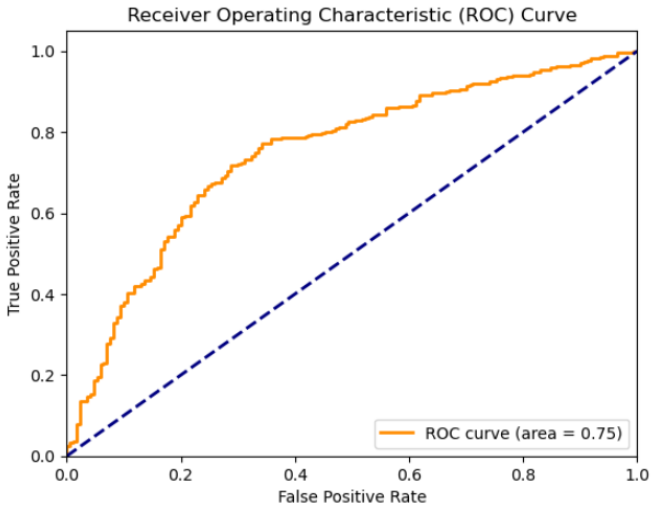


Ilustración 12. Curva ROC del conjunto de prueba. Variante: *Undersampling*

Estos valores indican que el modelo clasificó correctamente el 70% de los ejemplos en el conjunto de prueba. El 85% de las instancias clasificadas como positivas por el modelo son realmente positivas y que el modelo identificó correctamente el 70% de todas las instancias positivas. Un F1 Score de 0.77 sugiere que el modelo tiene un buen balance entre precisión y sensibilidad. El área bajo la curva ROC indica un desempeño aceptable del modelo en la clasificación de ejemplos positivos y negativos, pero aún hay margen de mejora.

Para comprobar si el modelo presenta sobreajuste, se realizó una evaluación adicional en el conjunto de entrenamiento. Los resultados obtenidos fueron:

MÉTRICA	**RESULTADO**
Exactitud	0.95
Precisión	0.93
Sensibilidad	0.97
F1 score	0.95

Tabla 10. Resultados del conjunto de entrenamiento. Variante: *Undersampling*

Fuente: Elaboración propia

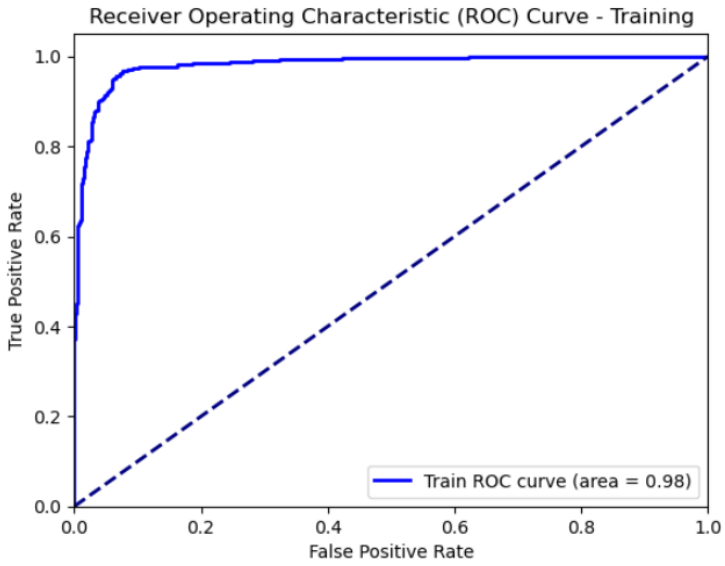


Ilustración 13. Curva ROC conjunto de entrenamiento. Variante: *Undersampling*

Como en los dos ejemplos anteriores, el modelo presenta **sobreajuste,** ya que muestra un rendimiento significativamente mejor en el conjunto de entrenamiento en comparación con el conjunto de prueba.

Para mitigar el sobreajuste observado en el modelo, se ha añadido una capa de ***Dropout*** con una tasa de 0.3 en cada una de las capas intermedias de la red neuronal. Los datos obtenidos han sido:

MÉTRICA	**CONJUNTO DE PRUEBA**	**CONJUNTO DE ENTRENAMIENTO**
Exactitud	0.73	0.74
Precisión	0.88	0.73
Sensibilidad	0.70	0.76
F1 score	0.78	0.74

Tabla 11. Resultados después de realizar *Dropout*. Variante: *Undersampling*

Fuente: Elaboración propia

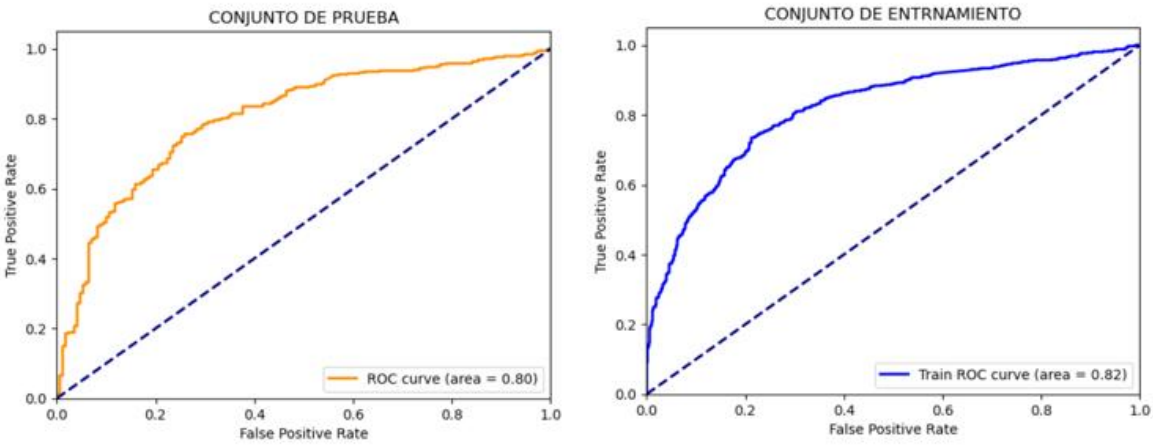


Ilustración 14. Comparativa de la Curva ROC del conjunto de prueba y de entrenamiento. Variante: *Undersampling*

Fuente: Elaboración propia con Python

El modelo alcanzó una exactitud de 0.73 en el conjunto de prueba, lo que representa un buen nivel de rendimiento en la predicción de casos correctos. La exactitud en el conjunto de entrenamiento es muy similar (0.74), lo que sugiere una mejora en la capacidad del modelo para generalizar y una reducción del sobreajuste. La precisión en el conjunto de prueba es notablemente alta (0.88), lo que indica que el modelo predice de manera correcta la mayoría de los casos positivos y minimiza los falsos positivos. En el conjunto de entrenamiento, la precisión es más baja (0.73), lo que sugiere que el modelo no está memorizando excesivamente los datos de entrenamiento. La sensibilidad en el conjunto de prueba es de 0.70, lo que implica que el modelo detecta correctamente el 70% de los casos positivos reales. En el conjunto de entrenamiento, la sensibilidad es ligeramente superior (0.76), pero sigue estando en un rango similar, lo que es un buen indicativo de que el modelo está equilibrado y no está sobreajustado. El F1 score es de 0.78 en el conjunto de prueba y de 0.74 en el conjunto de entrenamiento, lo que refleja un buen equilibrio entre precisión y sensibilidad en ambos conjuntos de datos. Esta coherencia sugiere que el modelo tiene un rendimiento consistente y es capaz de generalizar mejor a datos no vistos.

4.5.4. Comparativa de resultados

A continuación, se presenta una comparación de los resultados obtenidos en cada técnica, considerando las métricas clave de evaluación en los conjuntos de prueba:

MÉTRICA	**AJUSTE PESOS**	**OVERSAMPLING**	**UNDERSAMPLING**

Exactitud	0.72	0.67	0.73
Precisión	0.79	0.73	0.88
Sensibilidad	0.82	0.81	0.70
F1 score	0.80	0.77	0.78
ROC-AUC	0.72	0.73	0.79

Tabla 12. Comparativa de las métricas en las 3 variantes realizadas.

Fuente: Elaboración propia

El ajuste de pesos ofreció un buen equilibrio entre precisión y sensibilidad, sin embargo, se observa una notable diferencia entre las métricas del conjunto de entrenamiento y el de prueba, lo que indica cierto grado de sobreajuste. El *oversampling* aumentó la capacidad del modelo para detectar correctamente casos positivos, pero a costa de una mayor cantidad de falsos positivos y un evidente sobreajuste. El *undersampling* con *Dropout* proporcionó el mejor equilibrio general entre evitar el sobreajuste y mantener una buena precisión y sensibilidad. Las métricas de prueba y entrenamiento son más consistentes, lo que sugiere que esta técnica puede ser la más adecuada para este problema específico.

Por lo tanto, aunque cada técnica tiene sus ventajas, el *undersampling* combinado con *Dropout* parece ser la estrategia más efectiva para este modelo, ya que mejora la generalización sin sacrificar demasiado la precisión y la sensibilidad.

5. Conclusiones y trabajo futuro

5.1. Conclusiones

Este trabajo ha consistido en el desarrollo de un modelo predictivo basado en redes neuronales profundas para predecir la respuesta al tratamiento en pacientes con cáncer de pulmón, utilizando datos del estudio PLCO.

A lo largo del proyecto, se han implementado diversas técnicas para manejar el desbalance de clases y mitigar el sobreajuste, con el fin de mejorar la capacidad del modelo para generalizar a datos no vistos y proporcionar predicciones más precisas y fiables. Se aplicaron tres técnicas principales: ajuste de pesos, *oversampling*, y *undersampling* combinado con *Dropout*. Cada técnica fue evaluada en términos de precisión, sensibilidad, F1 score, y área bajo la curva ROC (AUC). Los resultados indicaron que el *undersampling* con *Dropout* proporcionó el mejor equilibrio general entre evitar el sobreajuste y mantener un rendimiento robusto en el conjunto de prueba. La inclusión de técnicas de regularización, como el *Dropout*, y la reducción del número de épocas de entrenamiento ayudaron a mejorar la capacidad del modelo para generalizar, minimizando el riesgo de sobreajuste observado inicialmente con técnicas como el *oversampling*.

El proceso de desarrollo y evaluación de este modelo predictivo ha permitido profundizar en el uso de redes neuronales profundas y el aprendizaje profundo aplicado a datos médicos complejos. Se ha demostrado la importancia de aplicar técnicas adecuadas de manejo de desbalance de clases y de regularización para mejorar la capacidad de generalización de los modelos predictivos, especialmente en un contexto tan crítico como el tratamiento del cáncer de pulmón.

El enfoque metodológico seguido, basado en el marco CRISP-DM, ha permitido estructurar el trabajo de manera eficiente, desde la comprensión del negocio y los datos, hasta el modelado, evaluación, y el análisis de los resultados obtenidos. Este enfoque ha sido fundamental para asegurar que cada etapa del proyecto estuviera alineada con los objetivos generales.

El aprendizaje obtenido durante el máster y la realización de este trabajo final han sido fundamentales para consolidar conocimientos en técnicas de modelado predictivo, manejo de datos complejos, y aplicación de metodologías estructuradas de ciencia de datos. Además,

se ha aprendido la importancia de adaptar estas técnicas al contexto específico de cada problema, considerando no solo los aspectos técnicos sino también las implicaciones clínicas y éticas.

5.2. Líneas de trabajo futuro

A partir de los resultados obtenidos en este trabajo, surgen varias líneas de investigación futuras que podrían contribuir a mejorar y ampliar el alcance del modelo predictivo desarrollado para predecir la respuesta al tratamiento en pacientes con cáncer de pulmón de células no pequeñas.

Una de las limitaciones de este estudio fue el uso de un único conjunto de datos (PLCO) con un número limitado de pacientes con NSCLC. Futuros trabajos podrían incorporar datos de otros estudios o registros clínicos internacionales para mejorar la generalización del modelo y su aplicabilidad a diferentes poblaciones. Generar un modelo multimodal con datos de imágenes médicas, como tomografías computarizadas y resonancias magnéticas (MRI), utilizando técnicas avanzadas de aprendizaje profundo como redes neuronales convolucionales, podría mejorar la capacidad del modelo para capturar patrones complejos asociados con la respuesta al tratamiento. Y, por último, la inclusión de variables adicionales relacionadas con factores socioeconómicos, ambientales y de calidad de vida podría proporcionar un enfoque más holístico para la predicción de la respuesta al tratamiento, permitiendo un análisis más completo de los factores que afectan los resultados en pacientes con NSCLC.

Realizar validaciones externas con datos de cohortes independientes de otros centros médicos o estudios podría evaluar la robustez y eficacia del modelo en diferentes contextos clínicos. Esta validación adicional ayudaría a determinar la aplicabilidad real del modelo en la práctica clínica.

Desarrollar técnicas que mejoren la interpretabilidad del modelo, como métodos de inteligencia artificial explicable (XAI), permitiría a los profesionales de la salud comprender mejor las predicciones del modelo y, por tanto, aumentar la confianza y aceptación en su uso clínico. Integrar el modelo en un sistema de apoyo a la decisión clínica (CDSS) que asista a los médicos en la toma de decisiones personalizadas para el tratamiento del cáncer de pulmón.

Este sistema podría proporcionar recomendaciones basadas en datos y análisis predictivo, mejorando así la atención al paciente.

Futuros estudios podrían explorar diferentes arquitecturas de redes neuronales y técnicas de optimización de hiperparámetros para mejorar aún más el rendimiento del modelo. El uso de enfoques como la búsqueda bayesiana o algoritmos evolutivos podría ayudar a identificar configuraciones de modelos más efectivas.

Referencias bibliográficas

American Cancer Society. (2019, octubre 1). *Etapas del cáncer de pulmón no microcítico* [Etapas del cáncer de pulmón no microcítico]. Etapas del cáncer de pulmón no microcítico. https://www.cancer.org/es/cancer/tipos/cancer-de-pulmon/deteccion-diagnostico-clasificacion-por-etapas/clasificacion-por-etapas-no-microcitico.html

American Cancer Society. (2024, enero 29). *Lung Cancer Risk Factors | Smoking & Lung Cancer* [Lung Cancer Risk Factors | Smoking & Lung Cancer]. Lung Cancer Risk Factors | Smoking & Lung Cancer. https://www.cancer.org/cancer/types/lung-cancer/causes-risks-prevention/risk-factors.html

American Cancer Society. (2024, enero 29). *Lung Cancer Signs & Symptoms | Early Signs of Lung Cancer* [Lung Cancer Signs & Symptoms | Early Signs of Lung Cancer]. Lung Cancer Signs & Symptoms | Early Signs of Lung Cancer. https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/signs-symptoms.html

American Cancer Society. (2024, enero 29). *What Is Lung Cancer? | Types of Lung Cancer* [What Is Lung Cancer?]. What Is Lung Cancer? https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html

Bengio, Y. (2012). *Practical recommendations for gradient-based training of deep architectures* (arXiv:1206.5533). arXiv. https://doi.org/10.48550/arXiv.1206.5533

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Callender, T., Imrie, F., Cebere, B., Pashayan, N., Navani, N., Schaar, M. van der, & Janes, S. M. (2023). Assessing eligibility for lung cancer screening using parsimonious ensemble machine learning models: A development and validation study. *PLOS Medicine*, *20*(10), e1004287. https://doi.org/10.1371/journal.pmed.1004287

Ciupka, N. W. B. (2020, noviembre 4). SCLC vs. NSCLC: What's the Difference? | NFCR Lung Cancer. *NFCR*. https://www.nfcr.org/blog/small-cell-lung-cancer-vs-non-small-cell-lung-cancer-whats-the-difference/

Contreras, F. (2008). *Esquema típico de una red neuronal artificial de...* ResearchGate. https://www.researchgate.net/figure/Figura-2-Esquema-tipico-de-una-red-neuronal-artificial-de-retropropagacion_fig1_260673350

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., Bruno, D. S., Chang, J. Y., Chirieac, L. R., D'Amico, T. A., DeCamp, M., Dilling, T. J., Dowell, J., Gettinger, S., Grotz, T. E., Gubens, M. A., Hegde, A., Lackner, R. P., Lanuti, M., … Hughes, M. (2022). Non–Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network*, *20*(5), 497-530. https://doi.org/10.6004/jnccn.2022.0025

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2022). An introduction to statistical learning with applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, *6*(1), 87-87. https://doi.org/10.1080/24754269.2021.1980261

Glorot, X., & Bengio, Y. (2006). *Understanding the difficulty of training deep feedforward neural networks*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ian Goodfellow, Yoshua Bengio, & Aaron Courville. (2016). *Deep Learning* [Deep Learning]. Deep Learning. https://www.deeplearningbook.org/

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kobylińska, K., Orłowski, T., Adamek, M., & Biecek, P. (2022). Explainable Machine Learning for Lung Cancer Screening Models. *Applied Sciences*, *12*(4), Article 4. https://doi.org/10.3390/app12041926

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lung cancer—Diagnosis. (2017, octubre 23). Nhs.Uk. https://www.nhs.uk/conditions/lung-cancer/diagnosis/

Mitchell, T. M. (2013). *Machine learning* (Nachdr.). McGraw-Hill.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nair, V., & Hinton, G. E. (2010). *Rectified Linear Units Improve Restricted Boltzmann Machines*.

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. *Twenty-First International Conference on Machine Learning - ICML '04*, 78. https://doi.org/10.1145/1015330.1015435

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com

Non-Small Cell Lung Cancer Treatment—NCI (nciglobal,ncienterprise). (2024, abril 26). [pdqCancerInfoSummary]. https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

Pan, Z., Zhang, R., Shen, S., Lin, Y., Zhang, L., Wang, X., Ye, Q., Wang, X., Chen, J., Zhao, Y., Christiani, D. C., Li, Y., Chen, F., & Wei, Y. (2023). OWL: An optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations. *eBioMedicine*, *88*. https://doi.org/10.1016/j.ebiom.2023.104443

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation* (arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

Rozalen, G., Seiffert, A. P., Gómez, E. J., Martín-Pinacho, J. J., & Sánchez-González, P. (2023). *Modelo de predicción de la respuesta al tratamiento de quimio- radioterapia en pacientes con cáncer de pulmón de células no pequeñas localmente avanzado irresecable mediante la aplica- ción de radiómica en imágenes de TC*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (arXiv:1708.08296). arXiv. https://doi.org/10.48550/arXiv.1708.08296

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, *74*(1), 12-49. https://doi.org/10.3322/caac.21820

SPSS Modeler Subscription. (2021, agosto 17). https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Tibshirani, R., Friedman, Jerome, J., & Hastie, T. (2001). *Valerie and Patrick Hastie*.

Tratamiento del cáncer de pulmón de células no pequeñas (nciglobal,ncienterprise). (2024, agosto 16). [pdqCancerInfoSummary]. https://www.cancer.gov/espanol/tipos/pulmon/paciente/tratamiento-pulmon-celulas-no-pequenas-pdq

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., H Mak, R., & Hugo J W L Aerts. (2019, abril 22). *Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging—PubMed*. https://pubmed.ncbi.nlm.nih.gov/31010833/

Anexo A. Código fuente y datos analizados

Código del modelo de red neuronal con undersampling.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.utils import class_weight
from sklearn.metrics import precision_score, recall_score, f1_score,
roc_curve, auc
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.regularizers import l2
from tensorflow.keras.metrics import Precision, Recall
from imblearn.under_sampling import RandomUnderSampler
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Cargar el dataset
df = pd.read_csv('lung_final.csv')
df = df.drop(columns=['plco_id'])

# Variables categóricas
categorical_columns = ['num_cancl','lung_cancer',
                       'intstatl_cat','lung_topography',
                       'lung_stage', 'lung_grade', 'lung_histtype_cat',
                       'curative_pneuml', 'curative_wsll',
                       'curative_chemol', 'curative_radl',
                       'neoadjuvantl', 'lung_cancer_first',
                       'biopllink0', 'biopllink1', 'biopllink3', 'educat',
                       'marital',
                       'occupat', 'pipe', 'cigar', 'sisters', 'brothers',
                        'asp', 'ibup',
                       'xray_history', 'asppd', 'ibuppd', 'hyperten_f',
                       'hearta_f',
                       'stroke_f', 'emphys_f', 'bronchit_f', 'diabetes_f',
                       'polyps_f',
                       'arthrit_f', 'osteopor_f', 'divertic_f',
                       'gallblad_f', 'race7',
                       'smoked_f', 'smokea_f', 'rsmoker_f', 'cigpd_f',
                       'filtered_f',
                       'cig_stat', 'cig_stop', 'bmi_curc',
                       'colon_comorbidity',
                       'liver_comorbidity', 'fh_cancer', 'lung_fh',
                       'mortality_exitage',
                       'ph_any_trial', 'ph_lung_trial', 'sex', 'agelevel',
                       'fstcan_exitstat', 'proc_catl_1_res',
                       'proc_catl_2_res', 'proc_catl_3_res',
                       'proc_catl_4_res', 'proc_catl_5_res',
                       'proc_catl_6_res', 'proc_catl_7_res',
                       'proc_catl_8_res']
```

```
# Separar características de la variable objetivo
X = df.drop(columns=['d_dthl'])
y = df['d_dthl']

# Definir el preprocesamiento para características categóricas (one-hot
encoding)
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'),
categorical_columns)
    ],
    remainder='passthrough'
)


pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('scaler', StandardScaler(with_mean=False))
])


# Dividir en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
stratify=y, random_state=42)

# Undersampling
undersampler = RandomUnderSampler(random_state=0)
X_train_resampled, y_train_resampled = undersampler.fit_resample(X_train,
y_train)

# Convertir y_train a numpy array para evitar problemas de indexación
y_train_array = np.array(y_train_resampled)
# Ajustar y transformar datos
X_train_transformed = pipeline.fit_transform(X_train_resampled)
X_test_transformed = pipeline.transform(X_test)

# Red neuronal
model = Sequential([
    Dense(128, input_shape=(X_train_transformed.shape[1],),
activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(64, activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(32, activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(16, activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(16, activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(16, activation='relu', kernel_regularizer=l2(0.001)),
    Dropout(0.3),
    Dense(1, activation='sigmoid')
])
model.compile(optimizer='adam', loss='binary_crossentropy',
metrics=['accuracy'])
# Entrenar el modelo
```

```
model.fit(X_train_transformed, y_train_array, epochs=15, batch_size=64,
verbose=1)

# Evaluar el modelo con el conjunto de prueba
y_pred_proba = model.predict(X_test_transformed)
y_pred = (y_pred_proba > 0.5).astype("int32").flatten()
accuracy = np.mean(y_pred == np.array(y_test))
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')
print(f'Precisión: {precision:.2f}')
print(f'Recall: {recall:.2f}')
print(f'F1 Score: {f1:.2f}')

# Curva ROC-AUC conjunto de prueba
fpr, tpr, _ = roc_curve(y_test, y_pred_proba)
roc_auc = auc(fpr, tpr)

plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area =
{roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()

# Evaluar el modelo en el conjunto de entrenamiento
y_train_pred_proba = model.predict(X_train_transformed)
y_train_pred = (y_train_pred_proba > 0.5).astype("int32").flatten()
accuracy_train = np.mean(y_train_pred == y_train_array)
precision_train = precision_score(y_train_array, y_train_pred)
recall_train = recall_score(y_train_array, y_train_pred)
f1_train = f1_score(y_train_array, y_train_pred)
print(f'Accuracy: {accuracy_train:.2f}')
print(f'Precision: {precision_train:.2f}')
print(f'Recall: {recall_train:.2f}')
print(f'F1 Score: {f1_train:.2f}')

# Curva ROC-AUC en el conjunto de entrenamiento
fpr_train, tpr_train, _ = roc_curve(y_train_array, y_train_pred_proba)
roc_auc_train = auc(fpr_train, tpr_train)

plt.figure()
plt.plot(fpr_train, tpr_train, color='blue', lw=2, label=f'Train ROC curve
(area = {roc_auc_train:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve - Training')
plt.legend(loc="lower right")
```

```
plt.show()
```

Debido a que los datos utilizados en este trabajo contienen información sensible de pacientes, no es posible mostrarlos directamente ya que no son de acceso público ni de código abierto. Sin embargo, los datos se han obtenido a través de la plataforma del CDAS, que regula el acceso a conjuntos de datos confidenciales como el del estudio PLCO. En este anexo, se incluye el acuerdo firmado con el CDAS que autoriza el uso de estos datos para fines de investigación.

DATA TRANSFER AGREEMENT

For PLCO Data

Please complete the information below:	
CDAS PROJECT NUMBER:	PLCO-1542
PROJECT TITLE:	AI Modeling Project for Predicting Treatment Response in Non-Small Cell Lung Cancer
RECIPIENT:	Universidad Internacional de La Rioja
RECIPIENT LEAD INVESTIGATOR:	Anna Gracia Colmenarejo

The National Cancer Institute (NCI) and the RECIPIENT hereby enter into this Agreement for the transfer of data collected in the course of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (DATA) to RECIPIENT through NCI's Cancer Data Access System (CDAS). Collectively or individually, NCI and RECIPIENT shall also be referred to as Parties or Party. This Agreement is effective as of the date of the last signature below (Effective Date).

In consideration of NCI providing DATA to RECIPIENT, RECIPIENT hereby agrees to the following terms and conditions:

1. DATA WILL NOT BE USED TO TREAT OR DIAGNOSE HUMAN SUBJECTS. RECIPIENT will use DATA in compliance with all applicable local, state, and/or federal laws and regulations, including but not limited to those for the protection of human subjects.
2. RECIPIENT must not use DATA for any study other than the approved Research Plan, attached as **Attachment 1**, unless RECIPIENT obtains the written consent of NCI by way of a new approved application through CDAS or by written and signed amendment to this Agreement. RECIPIENT grants NCI the right to publicly disclose the Research Plan, including titles, summaries or any other information contained therein as well as the names and contact information for the investigators conducting the research.
3. The DATA will be used solely by RECIPIENT LEAD INVESTIGATOR and RECIPIENT's faculty, employees, fellows, students, and agents that have a need to use, or provide a service in respect of, the DATA in connection with the Research Plan and whose obligations for using the DATA are consistent with the terms of this Agreement.
4. The DATA will not be further distributed to others without NCI's written consent. The RECIPIENT shall refer any request for the DATA to NCI.
5. Personally identifiable information will not be provided. If DATA being provided are coded, RECIPIENT will not request the key to the code. RECIPIENT must not attempt to learn the identity of or to contact the human subjects from which DATA were obtained, their physicians, or the collection sites for DATA. In the event that personally identifiable information is inadvertently transferred, RECIPIENT agrees to immediately destroy the personally identifiable information and report the circumstances to NCI. The DATA may be protected by the Federal Privacy Act and/or a Certificate of Confidentiality.

6. DATA are the property of NCI and are made available as a service to the research community. RECIPIENT will not claim. infer. or imply ownership of DATA or any endorsement of RECIPIENT'S activities or products by the U.S. Government, DHHS, NIH, NCI, or NCI employees.

7. NCI MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF DATA WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS. Unless prohibited by law, RECIPIENT assumes all liability for claims for damages against it by third parties which may arise from its use, storage, or disposal of DATA.

8. RECIPIENT will acknowledge NCI as the source of DATA in all publications and presentations by including language substantially similar to the following: "The authors thank the National Cancer Institute for access to NCI's data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial". Each publication and presentation should reference the CDAS Project Number.

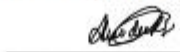
9. RECIPIENT must submit a description of each publication resulting from its use of DATA to the following website: https://cdas.cancer.gov/projects/plco/1542/PLCO-1542/discussion/. RECIPIENT agrees that NCI may publicly disclose this description.

10. This Agreement shall be in effect for five (5) years from the Effective Date. At the end of these five (5) years, if RECIPIENT is still using DATA for the approved Research Plan, RECIPIENT may seek an amendment to extend the term of this Agreement. This Agreement may be terminated by either Party for any reason by providing written notice to the other Party at least thirty (30) days prior to the desired termination date. Upon expiration or earlier termination of this Agreement or if RECIPIENT's use of DATA is complete, RECIPIENT must destroy DATA and upon NCI's request, confirm in writing as to such destruction. The RECIPIENT may retain one (1) copy of the DATA to the extent necessary to comply with the records retention requirements under any law, and for the purposes of research integrity and verification.

SIGNATURES BEGIN ON THE NEXT PAGE

ACCEPTED AND AGREED

FOR THE RECIPIENT (Universidad Internacional de La Rioja)



08/05/2024

(Authorized Signatory for Recipient) — Date

Printed Name: Almudena Ruiz Iniesta

Title: Academic Coordinator of the Master's Degree in Artificial Intelligence

Address: Av. de la Paz 137

26006, Logroño (La Rioja)

Read and Acknowledged by Recipient Lead Investigator:

Signature 

Name: Anna Gracia Colmenarejo — Date: May 8th, 2024

CDAS DTA for PLCO-1542
NCI Reference No._____

Page 3 of 3
July 27, 2021 Version

Project Information

Title:

AI Modeling Project for Predicting Treatment Response in Non-Small Cell Lung Cancer

Summary:

This project represents the culmination of my Master's program in Artificial Intelligence at International University of La Rioja. It seeks to bridge the gap between cutting-edge technology and personalized medical care. The primary goal is to construct an advanced AI model capable of predicting treatment responses in lung cancer patients. By harnessing the power of machine learning, this model aims to contribute the field of oncology by providing clinicians with valuable insights into individual patient outcomes. Through meticulous data analysis and algorithm development, the project endeavors to pave the way for more effective and personalized treatment strategies in the fight against cancer.

Aims:

Develop a robust deep learning model leveraging machine learning techniques to predict the response to treatment in patients diagnosed with non-small cell lung cancer (NSCLC).

Implement state-of-the-art algorithms for feature selection, dimensionality reduction, and predictive modeling to enhance the accuracy and generalizability of the AI model.

Evaluate the performance of the AI model using rigorous cross-validation techniques.

Investigate and collaborate in the advancement of personalized medicine by leveraging deep learning techniques to contribute to the field of oncology, particularly in the realm of non-small cell lung cancer (NSCLC) treatment response prediction.