UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA
ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS
ESPECIALIDAD: INGENIERÍA INFORMÁTICA Y DE SISTEMAS

TESIS

OPTIMIZACIÓN DE UN MODELO ENSEMBLE MEDIANTE INGENIERÍA DE CARACTERÍSTICAS Y SHAP PARA LA PREDICCIÓN DE SUPERVIVENCIA EN PACIENTES CON CÁNCER DE PULMÓN UTILIZANDO LA BASE DE DATOS SEER, 2015-2023

PRESENTADO POR:

BACH. FRAN JHOEL YAPO HUARAYA

BACH. MANUEL EDUARDO QUISPE CONDORI

PARA OPTAR AL TÍTULO PROFESIONAL DE INGENIERO INFORMÁTICO Y DE SISTEMAS

ASESOR:

RAÚL HUILLCA HUALLPARIMACHI

CUSCO – PERÚ

2026

Resumen

El cáncer de pulmón constituye un importante problema de salud pública debido a su elevada incidencia y mortalidad. La presente investigación tiene como objetivo optimizar el rendimiento predictivo de un modelo ensemble mediante ingeniería de características y selección con SHAP para predecir la supervivencia de pacientes con cáncer de pulmón registrados en la base de datos Surveillance, Epidemiology, and End Results (SEER) durante el periodo 2015–2023. El estudio adopta un enfoque cuantitativo, retrospectivo, longitudinal y computacional, con un diseño no experimental y nivel predictivo. La metodología comprende la extracción y depuración de registros, el preprocesamiento de variables clínicas y demográficas, la construcción de características derivadas y la optimización de un modelo ensemble de supervivencia. Se considerará la censura a la derecha y se controlará la fuga de información mediante particiones independientes de entrenamiento, validación y prueba. El rendimiento se evaluará principalmente mediante el índice de concordancia (C-index), complementado con el área bajo la curva dependiente del tiempo y métricas de calibración. Finalmente, se utilizarán valores SHAP para interpretar las contribuciones de las características a las predicciones. Se espera aportar un procedimiento reproducible para el análisis predictivo de la supervivencia en pacientes con cáncer de pulmón.

Palabras clave: Cáncer de pulmón, Análisis de supervivencia, Modelos ensemble, SHAP, SEER.

Abstract

Lung cancer is a major public health problem due to its high incidence and mortality. This research aims to optimize the predictive performance of an ensemble model through feature engineering and SHAP-based feature selection to predict survival in lung cancer patients registered in the Surveillance, Epidemiology, and End Results (SEER) database during the 2015–2023 period. The study adopts a quantitative, retrospective, longitudinal, and computational approach, with a non-experimental design and predictive scope. The methodology includes record extraction and cleaning, preprocessing of clinical and demographic variables, construction of derived features, and optimization of an ensemble survival model. Right censoring will be considered, and information leakage will be controlled through independent training, validation, and test partitions. Model performance will be evaluated primarily using the concordance index (C-index), complemented by the time-dependent area under the curve and calibration metrics. Finally, SHAP values will be used to interpret the contributions of features to the model predictions. The study is expected to provide a reproducible procedure for predictive survival analysis in lung cancer patients.

Keywords: Lung cancer, Survival analysis, Ensemble models, SHAP, SEER.

Índice general

Resumen **ii**

Abstract **iii**

Índice de figuras **v**

Índice de tablas **vi**

Índice de algoritmos **vii**

Índice de abreviaturas y acrónimos **viii**

Introducción **1**

I. Planteamiento del problema **4**

1.1. Descripción de la realidad problemática 4

1.1.1. Glosario y fundamentación conceptual de la realidad problemática . 8

1.2. Formulación del problema 10

1.2.1. Problema general 10

1.2.2. Problemas específicos 10

1.3. Delimitación de la investigación 11

1.3.1. Delimitación espacial y poblacional 11

1.3.2. Delimitación temporal 11

1.3.3. Delimitación teórica y tecnológica 11

1.4. Justificación de la investigación 12

1.4.1. Justificación teórica 12

1.4.2. Justificación práctica 12

1.4.3. Justificación metodológica 13

1.4.4. Justificación tecnológica 13

1.5. Objetivos de la investigación 14

1.5.1. Objetivo general 14
1.5.2. Objetivos específicos 14

II. Marco teórico **15**

2.1. Antecedentes de la investigación 15
2.1.1. Antecedentes internacionales 15
2.1.2. Antecedentes nacionales 17
2.1.3. Antecedentes regionales 18
2.1.4. Síntesis de los antecedentes y vacío de investigación 19
2.2. Bases teóricas 20
2.2.1. Cáncer de pulmón y datos de SEER 20
2.2.2. Aprendizaje automático sobre datos tabulares y modelos ensemble . 23
2.2.3. Análisis de supervivencia 24
2.2.4. Censura a la derecha y su tratamiento 25
2.2.5. Riesgos competitivos 26
2.2.6. Ingeniería de características 26
2.2.7. Selección de características y SHAP 26
2.2.8. Optimización de hiperparámetros 27
2.2.9. Validación y evaluación en datos independientes 28
2.2.10. Métricas de rendimiento 28
2.2.11. Interpretabilidad de los modelos 29
2.2.12. Fuga de información 30
2.2.13. Reproducibilidad 30
2.3. Definición de términos básicos 30

III. Diseño metodológico y experimental **33**

3.1. Tipo, nivel y diseño de investigación 33
3.1.1. Enfoque de investigación 33
3.1.2. Tipo y nivel de investigación 33
3.1.3. Diseño de investigación 34

3.2. Hipótesis de investigación 34
3.2.1. Hipótesis general 35
3.2.2. Hipótesis específicas 35
3.3. Población, muestra y unidad de análisis 35
3.4. Fuente de datos y periodo disponible 36
3.5. Criterios de inclusión y exclusión 37
3.6. Definición operacional del desenlace 38
3.6.1. Origen temporal, tiempo y estado 38
3.6.2. Tipo de desenlace y censura 38
3.6.3. Horizontes de evaluación y seguimientos incompletos 39
3.7. Predictores y control de la fuga de información 39
3.7.1. Momento de predicción y regla de admisibilidad 39
3.7.2. Auditoría de las variables de desenlace 40
3.7.3. Conjuntos de predictores 40
3.8. Preprocesamiento e ingeniería de características 40
3.8.1. Ingeniería de características 43
3.8.2. Selección de características y ablación 44
3.9. Pipeline experimental 44
3.10. Modelo ensemble y configuración de referencia 46
3.11. Optimización de hiperparámetros 46
3.11.1. Función objetivo y selección 46
3.11.2. Estrategia y presupuesto 47
3.11.3. Selección del modelo final 47
3.12. Partición y validación 48
3.12.1. Partición principal 48
3.12.2. Validación cruzada interna 49
3.12.3. Uso del conjunto de prueba y robustez 49
3.13. Métricas de evaluación 49
3.14. Comparación e incertidumbre estadística 51

3.15. Interpretabilidad mediante SHAP 52
3.16. Herramientas y reproducibilidad 52
3.17. Consideraciones éticas 53
3.18. Correspondencia entre objetivos, procedimientos e hipótesis 54

IV. Experimentación **56**

Referencias bibliográficas **57**

Anexos **63**

Anexo 1. Matriz de consistencia de la investigación 63

Índice de figuras

Índice de tablas

2.1. Síntesis comparativa de los antecedentes internacionales: muestra, modelos y aporte 21
2.2. Síntesis comparativa de los antecedentes nacionales: muestra, modelos y aporte 22
2.3. Síntesis comparativa de los antecedentes regionales: muestra, enfoque y aporte 22

3.1. Conjuntos de análisis y regla de formación de cada uno 36
3.2. Predictores admisibles al momento basal y su tratamiento 41
3.3. Variables excluidas del modelo por fuga de información o por derivación del desenlace 42
3.4. Espacios de búsqueda pre-registrados para la optimización del modelo ensemble 48
3.5. Métricas de evaluación pre-definidas por función 50
3.6. Correspondencia entre objetivos, procedimientos del Capítulo III e hipótesis 55

A.1. Matriz de consistencia: Optimización de un modelo ensemble mediante ingeniería de características y SHAP para la predicción de supervivencia en pacientes con cáncer de pulmón utilizando la base de datos SEER, 2015-2023 64

Índice de algoritmos

III.1. Pipeline experimental de diez etapas 45

Índice de abreviaturas y acrónimos

- **AUC**: área bajo la curva (p.).
- **C-index**: índice de concordancia (p.).
- **CIE-O-3**: Clasificación Internacional de Enfermedades para Oncología, tercera edición (p.).
- **GBDT**: *gradient-boosted decision trees* (p.).
- **IA**: inteligencia artificial (p.).
- **IBS**: *Integrated Brier Score* (p.).
- **NCI**: *National Cancer Institute* (p.).
- **NSCLC**: cáncer de pulmón de células no pequeñas (p.).
- **RSF**: *Random Survival Forest* (p.).
- **SCLC**: cáncer de pulmón de células pequeñas (p.).
- **SEER**: *Surveillance, Epidemiology, and End Results* (p.).
- **SHAP**: *SHapley Additive exPlanations* (p.).
- **TNM**: tumor, ganglios y metástasis (p.).
- **TRIPOD+AI**: guía de reporte de modelos predictivos clínicos que utilizan regresión o aprendizaje automático (p.).

Introducción

El cáncer de pulmón constituye una de las principales causas de mortalidad por neoplasias en el mundo. Las estimaciones globales más recientes registran aproximadamente 2,48 millones de casos nuevos y 1,8 millones de defunciones anuales, con una supervivencia a cinco años que permanece baja en la mayoría de los países (Bray et al., 2024). En el Perú, las estimaciones para 2024 reportan 3 727 casos nuevos y 3 100 defunciones anuales por esta enfermedad (International Agency for Research on Cancer, 2026).

El pronóstico asistencial del cáncer de pulmón descansa tradicionalmente en la estadificación anatómica TNM. Aunque este sistema es fundamental en el diagnóstico inicial, presenta limitaciones intrínsecas para capturar interacciones multivariadas complejas como la histología, la localización tumoral, el grado de diferenciación, la extensión regional, la presencia de metástasis y las características demográficas del paciente.

Por ello, la presente investigación adopta la base de datos SEER, administrada por el NCI de Estados Unidos, que reúne registros poblacionales estructurados con información demográfica, diagnóstico, características tumorales y seguimiento longitudinal de la supervivencia (National Cancer Institute, 2025).

Para superar las restricciones de los enfoques tradicionales, los modelos de IA sobre datos tabulares han demostrado capacidad discriminativa en oncología. Una revisión sistemática y metaanálisis reciente, que sintetiza 12 estudios y más de 210 000 pacientes, reportó un estadístico C promedio de 0.78 para los modelos de aprendizaje automático frente a 0.70 de la regresión logística, aunque con heterogeneidad metodológica considerable (Didier et al., 2024).

En esa misma línea, otras investigaciones recientes han aplicado modelos basados en árboles de decisión, bosques de supervivencia y redes neuronales profundas para estimar la supervivencia de pacientes con cáncer de pulmón bajo censura a la derecha (Astley et al., 2024; Jin et al., 2023; Wiegrebe et al., 2024; Yuan et al., 2025).

No obstante estos avances, la literatura presenta vacíos en el tratamiento conjunto de la representación de las características, la optimización de modelos ensemble y la interpretación

de las predicciones. Con frecuencia se reportan configuraciones puntuales sin documentar el efecto de la ingeniería de características ni el de la búsqueda de hiperparámetros; además, la interpretación mediante SHAP suele presentarse de forma descriptiva y no se integra explícitamente al análisis de las variables que contribuyen a la predicción (Lundberg & Lee, 2017; Salih et al., 2024).

Ante este escenario, la investigación delimita su problema en la siguiente interrogante: ¿de qué manera optimizar el rendimiento predictivo de un modelo ensemble mediante ingeniería de características y selección con SHAP para la predicción de supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015–2023? Para responder a esta problemática, se propone construir variables derivadas a partir de los registros clínicos y demográficos, desarrollar y optimizar un modelo ensemble de supervivencia, y analizar la contribución de las características mediante SHAP.

Científicamente, el estudio aporta evidencia empírica sobre el efecto de la ingeniería de características y de la optimización de hiperparámetros en datos oncológicos con censura a la derecha, así como sobre la interpretación de un modelo ensemble mediante valores de Shapley. En el plano metodológico, la investigación implementa las directrices internacionales del estándar TRIPOD+AI (Collins et al., 2024). Con ello se adopta un protocolo que separa las particiones de entrenamiento, validación y prueba, controla los sesgos por fuga de información y verifica la explicabilidad clínica mediante el análisis de valores SHAP (Salih et al., 2024).

Desde una perspectiva tecnológica, se diseña un marco de experimentación reproducible para el procesamiento de datos oncológicos de gran volumen. Este flujo permite cuantificar el equilibrio operativo entre rendimiento predictivo, interpretabilidad y costo computacional de la búsqueda de hiperparámetros en arquitecturas ensemble basadas en árboles de decisión y en modelos de supervivencia (Bischl et al., 2023; Grinsztajn et al., 2022).

El documento se organiza en cuatro capítulos y en los anexos: el Capítulo I expone el planteamiento del problema, la delimitación, la justificación y los objetivos; mientras que el Capítulo II desarrolla los antecedentes y las bases teóricas del estudio.

Por su parte, el Capítulo III detalla el diseño metodológico y experimental, con la formulación de las hipótesis, la definición de la cohorte SEER, del desenlace de supervivencia,

de los predictores, de la ingeniería de características, del pipeline de optimización, de la validación, de las métricas, de la interpretabilidad mediante SHAP y de las consideraciones éticas. Finalmente, el Capítulo IV corresponde a la experimentación y, tras las referencias bibliográficas, los anexos recogen la matriz de consistencia de la investigación.

Capítulo I

Planteamiento del problema

1.1 Descripción de la realidad problemática

El cáncer de pulmón representa uno de los principales problemas de salud pública a nivel mundial debido a su elevada incidencia y mortalidad. Las estimaciones de GLOBOCAN para 2022 registraron aproximadamente 2,48 millones de casos nuevos y 1,8 millones de defunciones por esta enfermedad en el mundo (Bray et al., 2024). La Organización Mundial de la Salud señala que el cáncer constituye una de las principales causas de muerte prematura y que el cáncer de pulmón concentra una parte importante de esa carga, especialmente en países con transiciones epidemiológicas y demográficas aceleradas (World Health Organization, 2025).

En el Perú, las estimaciones de GLOBOCAN 2024 reportan 3 727 casos nuevos y 3 100 defunciones por cáncer de pulmón (International Agency for Research on Cancer, 2026), lo que evidencia la necesidad de fortalecer las herramientas analíticas que permitan estudiar la evolución de esta enfermedad y apoyar la planificación sanitaria.

El comportamiento clínico del cáncer de pulmón es heterogéneo y la supervivencia de los pacientes depende de diversos factores, entre ellos la edad, las características demográficas, el tipo histológico, el estadio al momento del diagnóstico, la extensión de la enfermedad y los tratamientos recibidos. Aunque el sistema de estadificación TNM permite clasificar la enfermedad y orientar las decisiones terapéuticas, no representa por sí solo todas las características que pueden influir en la supervivencia individual.

Variables como la localización del tumor, el grado de diferenciación, la presencia de metástasis y las condiciones demográficas del paciente también pueden estar relacionadas con la evolución clínica. Por ello, el análisis conjunto de información clínica y poblacional permite estudiar el pronóstico desde una perspectiva más amplia.

La predicción de supervivencia en pacientes con cáncer constituye un problema complejo, debido a que no basta con determinar si un paciente sobrevive o fallece. Es

necesario considerar el tiempo transcurrido hasta el evento de interés y los casos en los que no se conoce el desenlace definitivo durante el periodo de observación.

Estos últimos casos se conocen como observaciones censuradas y requieren métodos estadísticos y computacionales apropiados para el análisis de supervivencia. Ignorar la censura o reducir el desenlace a una categoría binaria puede descartar información longitudinal valiosa y producir sesgos de selección (Denfeld et al., 2023; Wiegrebe et al., 2024).

Los modelos tradicionales de supervivencia, como el modelo de riesgos proporcionales de Cox, han sido ampliamente utilizados porque ofrecen una interpretación directa de los coeficientes y una base estadística consolidada. Sin embargo, pueden presentar limitaciones cuando existen relaciones no lineales, interacciones complejas y estructuras de datos de alta dimensionalidad.

En este contexto, los métodos de aprendizaje automático ofrecen alternativas para identificar patrones complejos y mejorar la capacidad predictiva, siempre que su desarrollo se realice con procedimientos de validación adecuados y con un control riguroso de la fuga de información (Didier et al., 2024; van de Mortel & van Wingen, 2025).

Los modelos ensemble permiten combinar diferentes modelos o algoritmos con el propósito de obtener predicciones más robustas. Su fundamento consiste en agregar las predicciones de varios aprendices para reducir la varianza, el sesgo o ambos, dependiendo de la estrategia de combinación empleada.

No obstante, el rendimiento de un ensemble depende de la calidad de los datos, la selección de variables, la representación de las características y la configuración de sus hiperparámetros. Por esta razón, su construcción no garantiza por sí misma una mejora predictiva y requiere una evaluación experimental que compare configuraciones de referencia frente a configuraciones optimizadas (Bischl et al., 2023; Grinsztajn et al., 2022).

La ingeniería de características permite transformar, combinar y construir variables a partir de los datos originales para representar mejor los factores relacionados con la supervivencia. Un proceso adecuado puede mejorar la calidad de la información que recibe el modelo, reducir redundancias y facilitar la identificación de patrones relevantes.

En el contexto oncológico, la creación de categorías clínicas interpretables, la armoni-

zación de códigos histológicos, el tratamiento explícito de valores faltantes y la construcción de variables derivadas a partir de la edad, el estadio y la extensión tumoral constituyen decisiones metodológicas que deben documentarse y evaluarse (Büyükkeçeci & Okur, 2023; Nijman et al., 2022).

Por otro lado, los modelos complejos pueden presentar dificultades de interpretación. SHAP (*SHapley Additive exPlanations*) proporciona herramientas para explicar las contribuciones de las características a las predicciones de un modelo. Mediante sus valores es posible analizar la importancia global de las variables y la contribución de estas en predicciones individuales, a partir de un marco teórico basado en la teoría de juegos cooperativos (Lundberg & Lee, 2017; Salih et al., 2024).

En esta investigación, SHAP podrá utilizarse para analizar la importancia de las características y, si el procedimiento metodológico lo contempla, apoyar una selección de características basada en su contribución predictiva. Es importante no presentar SHAP como un algoritmo que mejora automáticamente el rendimiento del modelo: su función principal es la interpretación y explicación de las predicciones.

La base de datos SEER constituye una fuente relevante para investigaciones sobre cáncer, debido a que reúne información poblacional sobre características demográficas, diagnóstico, características tumorales y supervivencia. Es administrada por el NCI de Estados Unidos y su documentación describe los dominios, los periodos de disponibilidad y las reglas de codificación de las variables (National Cancer Institute, 2025). Su utilización permite estudiar patrones asociados con la supervivencia de pacientes con cáncer de pulmón a partir de registros estructurados y de cobertura poblacional.

Sin embargo, trabajar con SEER también presenta desafíos. Es necesario conocer la documentación de las variables, los periodos de disponibilidad, los criterios de inclusión, los datos faltantes, las categorías codificadas y las limitaciones del seguimiento.

Asimismo, debe definirse correctamente el evento de interés y el tiempo de supervivencia, considerando las observaciones censuradas y diferenciando la supervivencia global de la supervivencia específica por cáncer. No se debe asumir que todas las variables están disponibles para todos los años del periodo 2015–2023 ni que todos los registros cuentan con

el mismo tiempo de seguimiento.

El periodo de diagnóstico 2015–2023 debe delimitarse de manera rigurosa. No se debe asumir que todos los pacientes cuentan con cinco años de seguimiento. El periodo de diagnóstico no es equivalente al periodo de seguimiento, por lo que será necesario especificar la versión de SEER utilizada, la fecha de corte del seguimiento y las reglas de construcción de la variable de tiempo. La censura administrativa, la pérdida de seguimiento y las diferencias en la disponibilidad de variables por año son aspectos que deben reportarse antes de interpretar cualquier métrica de rendimiento (Collins et al., 2024; National Cancer Institute, 2025).

En el contexto peruano también existen investigaciones clínicas relacionadas con el cáncer de pulmón. Gálvez-Nino et al. (2023) analizaron resultados reales de la terapia anti-EGFR en pacientes peruanos con cáncer de pulmón de células no pequeñas avanzado y mutación EGFR, aportando información sobre una población clínica específica.

Asimismo, Gago et al. (2024) desarrollaron un sistema basado en aprendizaje automático para el diagnóstico y la predicción de supervivencia de pacientes con cáncer de mama, mientras que Mamani Barrios (2023) estudió algoritmos de aprendizaje automático para predecir anemia infantil en Cusco. Aunque estos trabajos abordan problemas y poblaciones diferentes, muestran la aplicación de métodos computacionales en el análisis de información sanitaria y la necesidad de adaptar las técnicas a las características de cada fuente de datos.

En la región Cusco, los boletines de la Gerencia Regional de Salud registraron 226 notificaciones de tumores malignos de tráquea, bronquios y pulmón durante el periodo 2021–2025 (Gerencia Regional de Salud Cusco, 2026). Por su parte, Quispe Rodríguez (2022) estudió las características clínico-epidemiológicas y sociodemográficas de 78 pacientes con cáncer de pulmón atendidos en el Hospital Antonio Lorena durante 2015–2021.

Estas fuentes evidencian la importancia del problema en Cusco, aunque sus datos epidemiológicos y descriptivos no constituyen por sí mismos modelos predictivos de supervivencia. La presente investigación no utiliza esos registros locales, sino la base poblacional SEER, con el propósito de disponer de una fuente estructurada y auditable para el desarrollo metodológico.

Además, la construcción de modelos predictivos presenta dificultades relacionadas

con la heterogeneidad de los registros, los datos faltantes, la selección de variables disponibles al momento de la predicción y el riesgo de incorporar información posterior al diagnóstico. La inclusión de variables que revelan acontecimientos futuros puede producir una fuga de información y generar estimaciones de rendimiento artificialmente elevadas (van de Mortel & van Wingen, 2025).

Por ello, la evaluación debe separar adecuadamente los datos utilizados para el entrenamiento y la optimización de aquellos destinados a la prueba independiente, siguiendo criterios internacionales de desarrollo y reporte de modelos predictivos (Collins et al., 2024; Riley et al., 2024).

Ante esta problemática, la presente investigación plantea optimizar un modelo ensemble mediante ingeniería de características y selección con SHAP para predecir la supervivencia de pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015–2023. Para ello, se propone analizar y preprocesar los datos clínicos y demográficos, construir variables predictivas relevantes, desarrollar y optimizar el modelo ensemble, y evaluar su rendimiento mediante métricas apropiadas para análisis de supervivencia.

Asimismo, se analizarán las contribuciones de las características mediante SHAP, con el propósito de interpretar los factores que influyen en las predicciones y de aportar evidencia sobre el comportamiento del modelo en un conjunto de prueba independiente.

1.1.1 Glosario y fundamentación conceptual de la realidad problemática

A continuación se precisan los términos clínicos y computacionales empleados en el planteamiento:

1. **Estadios clínicos del cáncer (I, II, III y IV):** los estadios tempranos (I y II) comprenden neoplasias confinadas al parénquima pulmonar, susceptibles de cirugía curativa. Los estadios avanzados engloban el estadio III (invasión de ganglios mediastínicos o estructuras torácicas) y el estadio IV (metástasis a distancia). La disponibilidad y la codificación del estadio en SEER se verifican antes de incorporarlo como predictor.

2. **Histopatología:** subtipo celular del tumor determinado por biopsia. Comprende el carcinoma de células pequeñas (SCLC), de proliferación rápida y alta agresividad, y

el carcinoma de células no pequeñas (NSCLC), que incluye el adenocarcinoma y el carcinoma epidermoide, con patrones biológicos y de supervivencia distintos.

3. **Localización tumoral:** sitio anatómico primario del tumor, codificado mediante la topografía de la CIE-O-3. Para el cáncer de pulmón se consideran los códigos correspondientes a tráquea, bronquios y pulmón, de acuerdo con la documentación de SEER.

4. **Grado tumoral:** medida de diferenciación celular del tumor. Su codificación y disponibilidad varían según el año de diagnóstico y el tipo histológico, por lo que se armoniza en categorías clínicas amplias antes de su uso.

5. **Extensión regional y metástasis:** indicadores de diseminación tumoral a ganglios regionales y a órganos distantes. Se construyen a partir de las variables de estadio y extensión disponibles en la versión de SEER seleccionada.

6. **Censura a la derecha:** condición en la cual el seguimiento clínico finaliza antes de observar la muerte del paciente. Los clasificadores binarios que excluyen estas observaciones introducen un sesgo de selección; los modelos de supervivencia incorporan el tiempo parcial de seguimiento en la verosimilitud (Denfeld et al., 2023).

7. **Supervivencia global:** tiempo desde el diagnóstico hasta la muerte por cualquier causa, con censura a la derecha en los pacientes vivos al último contacto. Se diferencia de la supervivencia específica por cáncer, que considera únicamente las muertes atribuibles a la neoplasia.

8. **Calibración probabilística:** grado de concordancia entre la probabilidad predicha por el algoritmo y la frecuencia observada del evento. Se evalúa con el *Brier Score* y sus variantes dependientes del tiempo (Zhu et al., 2025).

9. **Fuga temporal (*look-ahead bias*):** sesgo producido cuando se incluyen como predictoras variables que cronológicamente ocurrieron después del momento basal del diagnóstico, como el estado vital final o el tiempo de seguimiento, lo que infla artificialmente el rendimiento en laboratorio (van de Mortel & van Wingen, 2025).

10. **Modelo ensemble:** modelo que combina las predicciones de dos o más aprendices para obtener una estimación agregada. La combinación puede realizarse mediante promediado, votación, potenciación de gradiente o apilamiento, y su rendimiento debe contrastarse experimentalmente.

11. **Interpretabilidad con SHAP:** conjunto de técnicas que descomponen una predicción en contribuciones aditivas de las características, permitiendo analizar la importancia global y local de las variables (Lundberg & Lee, 2017; Salih et al., 2024).

1.2 Formulación del problema

1.2.1 Problema general

¿De qué manera **optimizar el rendimiento predictivo** de un **modelo ensemble mediante ingeniería de características y selección con SHAP** para la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**?

1.2.2 Problemas específicos

PE1: ¿Cómo realizar el **preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** para identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**?

PE2: ¿De qué manera **desarrollar y optimizar** un **modelo ensemble mediante algoritmos de aprendizaje automático** para mejorar el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**?

PE3: ¿En qué medida la aplicación de **SHAP** permite **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**?

1.3 Delimitación de la investigación

1.3.1 Delimitación espacial y poblacional

La investigación utilizará registros de pacientes diagnosticados con cáncer de pulmón disponibles en la base de datos SEER, administrada por el NCI de Estados Unidos. La población estará conformada por los registros que cumplan los criterios de inclusión y exclusión establecidos para el estudio. La identificación de casos se realizará utilizando los códigos de topografía y morfología correspondientes, de acuerdo con la clasificación CIE-O-3 y la documentación oficial de SEER (National Cancer Institute, 2025).

El tamaño de la cohorte de esta investigación es un resultado del procedimiento de selección y no una cifra de partida. En el Capítulo III se documentan los criterios de inclusión y exclusión, la depuración de registros y la construcción de los conjuntos de análisis. La unidad de análisis será el registro oncológico del paciente, y la definición de unicidad se establecerá de acuerdo con los identificadores disponibles en la versión de SEER utilizada.

1.3.2 Delimitación temporal

Se considerarán los diagnósticos registrados entre los años 2015 y 2023. Se deberá especificar la versión de SEER utilizada y la fecha de corte del seguimiento. No se debe afirmar que todos los registros cuentan con cinco años de seguimiento, porque el periodo de diagnóstico no es equivalente al periodo de seguimiento. La cobertura efectiva, la fecha de corte, la proporción de censura y la mediana de seguimiento se verificarán antes de modelar y se reportarán como parte de la descripción de los datos.

1.3.3 Delimitación teórica y tecnológica

La investigación se centrará en el preprocesamiento de datos estructurados, la ingeniería de características, el desarrollo y la optimización de modelos ensemble, la evaluación de supervivencia y la interpretación mediante SHAP. Se considerarán modelos de análisis de supervivencia que manejan censura a la derecha, así como algoritmos de aprendizaje automático adecuados para datos tabulares. La comparación se realizará dentro de un procedi-

miento común de optimización y evaluación en datos independientes, bajo las directrices de TRIPOD+AI (Collins et al., 2024).

No se incluirá procesamiento de imágenes médicas, análisis genómico ni automatización del diagnóstico clínico. Tampoco se utilizarán variables posteriores al diagnóstico que comprometan la validez temporal del modelo. La optimización se limitará a estrategias de búsqueda sobre un espacio de hiperparámetros explícito, aplicadas exclusivamente sobre las particiones de entrenamiento y validación.

1.4 Justificación de la investigación

1.4.1 Justificación teórica

La investigación aporta evidencia sobre el efecto de la ingeniería de características y de la optimización de hiperparámetros en modelos ensemble que predicen la supervivencia de pacientes con cáncer de pulmón bajo censura a la derecha. La literatura consultada reporta el rendimiento de configuraciones puntuales, por lo que no siempre permite determinar cuánto de ese rendimiento proviene de la estructura del modelo, cuánto de la representación de las variables y cuánto de la búsqueda efectuada sobre los datos.

Frente a ello, el presente estudio permite contrastar, de manera cuantitativa, si las mejoras observadas durante la optimización persisten cuando el modelo se evalúa en datos independientes. Asimismo, este documento no asume que la optimización o el uso de SHAP mejoren necesariamente los indicadores: tal mejora es objeto de contraste experimental.

1.4.2 Justificación práctica

Los servicios de salud con recursos limitados requieren criterios verificables antes de incorporar un modelo predictivo a su operación. El procedimiento propuesto produce evidencia comprobable sobre cuándo un modelo ensemble optimizado rinde mejor que su configuración de referencia, con qué costo de búsqueda y con qué nivel de calibración, lo cual es indispensable para decidir si su adopción está justificada. Asimismo, el análisis mediante SHAP orienta a los profesionales clínicos sobre las variables que el modelo considera relevantes, sin sustituir el juicio médico ni establecer relaciones causales.

1.4.3 Justificación metodológica

El estudio incorpora las directrices de TRIPOD+AI (Collins et al., 2024) y las recomendaciones para el desarrollo y la validación de modelos predictivos clínicos (Efthimiou et al., 2024; Riley et al., 2024). Con ello se establece un protocolo experimental que:

1. controla la fuga de información, garantizando que las predicciones utilicen variables disponibles en el momento basal del diagnóstico (van de Mortel & van Wingen, 2025);

2. separa las particiones de entrenamiento, validación y prueba, de modo que la búsqueda de hiperparámetros y la selección de características no consulten los datos de prueba;

3. evalúa la discriminación y la calibración probabilística mediante el índice de concordancia, el área bajo la curva dependiente del tiempo, el *Brier Score* y el *Integrated Brier Score* (Zhu et al., 2025);

4. analiza la interpretabilidad del modelo ensemble con valores SHAP, sin atribuirles carácter causal (Salih et al., 2024).

1.4.4 Justificación tecnológica

Desde la perspectiva de la ingeniería de sistemas y ciencia de datos, se diseña un procedimiento computacional reproducible para datos tabulares clínicos de gran volumen. Este flujo permite cuantificar el equilibrio operativo entre rendimiento predictivo, interpretabilidad y costo computacional de la búsqueda de hiperparámetros en arquitecturas ensemble basadas en árboles y en modelos de supervivencia.

Con ello se facilita su eventual integración en sistemas de información hospitalaria con infraestructura estándar. Para tal fin, el espacio de hiperparámetros, la estrategia de búsqueda y el presupuesto se fijan en el Capítulo III (sección 3.11).

1.5 Objetivos de la investigación

1.5.1 Objetivo general

Optimizar el rendimiento predictivo de un **modelo ensemble mediante ingeniería**

de características y selección con SHAP para la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**.

1.5.2 Objetivos específicos

OE1: **Realizar el preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** para identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**.

OE2: **Desarrollar y optimizar** un **modelo ensemble mediante algoritmos de aprendizaje automático** para mejorar el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**.

OE3: **Aplicar SHAP** para **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**.

Capítulo II

Marco teórico

2.1 Antecedentes de la investigación

Los antecedentes se seleccionaron por su pertinencia con la predicción de supervivencia a partir de datos clínicos oncológicos y con los procedimientos de entrenamiento, validación, optimización e interpretación de modelos. Para fines de esta revisión se distingue, en cada estudio, el tipo de desenlace modelado, la población estudiada y el grado de relación con la base de datos SEER. Esta distinción permite identificar qué aportes son directamente aplicables al problema y cuáles constituyen fundamento metodológico.

2.1.1 Antecedentes internacionales

Didier et al. (2024), en su investigación titulada "Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis", desarrollaron una revisión sistemática y metaanálisis sobre la aplicación de algoritmos de aprendizaje automático en la predicción de supervivencia de pacientes con cáncer de pulmón. El trabajo reunió investigaciones que utilizaron diferentes técnicas de aprendizaje automático y métodos estadísticos para estimar el pronóstico de los pacientes.

A partir de esa síntesis, los autores identificaron el potencial de estos algoritmos para la predicción de supervivencia, aunque también señalaron diferencias entre los estudios en cuanto a los datos utilizados, las variables consideradas y los procedimientos de evaluación. Este antecedente es importante porque permite reconocer el desarrollo de modelos predictivos en cáncer de pulmón y la necesidad de evaluar rigurosamente su rendimiento.

Jin et al. (2023), en el estudio "Development and validation of machine learning models to predict survival of patients with resected stage-III NSCLC", desarrollaron y validaron modelos de aprendizaje automático para predecir la supervivencia de pacientes con cáncer de pulmón de células no pequeñas en estadio III sometidos a resección. La investigación utilizó información de la base de datos SEER y comparó modelos de supervivencia con

censura, lo que la convierte en un antecedente metodológico directo para el presente estudio.

Su aporte principal consiste en demostrar la aplicación de diferentes algoritmos para estimar la supervivencia en una población oncológica delimitada y en evidenciar la utilidad de SEER para el desarrollo de modelos predictivos. Sin embargo, sus resultados deben interpretarse considerando la población clínica específica y el contexto de tratamiento en el que fueron desarrollados.

Astley et al. (2024), en el trabajo "Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy", investigaron el uso de aprendizaje profundo explicable para predecir la supervivencia de pacientes con cáncer de pulmón de células no pequeñas que recibieron radioterapia radical. El estudio abordó la utilización de modelos de aprendizaje profundo y mecanismos de explicabilidad para analizar las predicciones.

En ese sentido, este antecedente resulta relevante porque muestra que las redes neuronales pueden emplearse en problemas de supervivencia oncológica y que la interpretación de las predicciones constituye un aspecto importante cuando los modelos se aplican a información clínica. No obstante, la población y el tratamiento considerados delimitan el alcance de sus resultados.

Yuan et al. (2025), en el estudio "Artificial intelligence-assisted machine learning models for predicting lung cancer survival", analizaron la aplicación de modelos de aprendizaje automático asistidos por inteligencia artificial para predecir la supervivencia de pacientes con cáncer de pulmón. La investigación representa un antecedente reciente sobre el empleo de técnicas computacionales para estimar el pronóstico de los pacientes.

Su relevancia se encuentra en la aplicación de diferentes modelos predictivos en información clínica oncológica, lo que permite reconocer la importancia de comparar algoritmos y estudiar su capacidad para realizar predicciones. Asimismo, este trabajo se relaciona con la presente investigación por el uso de aprendizaje automático en la estimación de supervivencia, aunque las diferencias en las fuentes de datos y los procedimientos de validación deben considerarse al comparar sus resultados.

Cardoso et al. (2026), en el artículo "Cross-cancer survival prediction using machine

learning models", desarrollaron modelos de aprendizaje automático destinados a predecir la supervivencia en diferentes tipos de cáncer. El estudio constituye un antecedente sobre el uso de modelos ensemble y técnicas de optimización en datos oncológicos, aunque no emplea la base de datos SEER ni un enfoque centrado en la ingeniería de características y la interpretación mediante SHAP.

En consecuencia, su aporte se relaciona con la comparación sistemática de algoritmos y con la necesidad de documentar los procedimientos de validación, aspectos metodológicos que se retoman en esta investigación.

2.1.2 Antecedentes nacionales

Gálvez-Nino et al. (2023), en el estudio "Real-world outcomes of anti-EGFR therapy in advanced non-small cell lung cancer EGFR mutated in Peru", analizaron resultados reales de la terapia anti-EGFR en pacientes peruanos con cáncer de pulmón de células no pequeñas avanzado y mutación EGFR. El trabajo aporta evidencia clínica nacional sobre el comportamiento de pacientes con características moleculares específicas y los resultados asociados al tratamiento.

Por ello, este antecedente es importante porque aborda directamente el cáncer de pulmón en el contexto peruano y demuestra la relevancia de estudiar información clínica real. Sin embargo, su población se encuentra delimitada por la presencia de una mutación específica y el tratamiento recibido, por lo que no representa necesariamente a todos los pacientes con cáncer de pulmón.

Gago et al. (2024), en el trabajo "Machine learning system for the effective diagnosis and survival prediction of breast cancer patients", desarrollaron un sistema basado en aprendizaje automático para el diagnóstico y la predicción de supervivencia de pacientes con cáncer de mama. Aunque el estudio no aborda el cáncer de pulmón, constituye un antecedente nacional relacionado con la aplicación de modelos de aprendizaje automático en el análisis de información oncológica.

De este modo, su contribución para la presente investigación se encuentra en el uso de técnicas computacionales para apoyar tareas de predicción en salud, considerando que los

modelos deben evaluarse de acuerdo con las características de los datos y el problema clínico estudiado.

Mamani Barrios (2023), en la tesis “Algoritmos de machine learning para predecir la anemia en niños de 6 a 35 meses de edad en Cusco”, estudió la aplicación de algoritmos de aprendizaje automático para predecir una condición de salud utilizando información correspondiente a una población infantil de Cusco. Aunque el problema clínico es diferente al cáncer de pulmón, este trabajo constituye un antecedente académico regional y nacional sobre el uso de técnicas de aprendizaje automático en datos de salud.

Su relación con la presente investigación se encuentra en la aplicación de modelos predictivos a información clínica estructurada, si bien el objetivo, la población y la naturaleza del desenlace son diferentes.

2.1.3 Antecedentes regionales

Quispe Rodríguez (2022), en la tesis “Cáncer de pulmón: características clínico epidemiológicas y sociodemográficas en el Hospital Antonio Lorena del Cusco, 2015–2021”, desarrolló un estudio descriptivo sobre pacientes diagnosticados con cáncer de pulmón en dicho establecimiento de salud. La investigación consideró 78 pacientes y describió características sociodemográficas, epidemiológicas y clínicas, entre ellas la edad, el sexo, los antecedentes de exposición y las características histológicas de la enfermedad.

En consecuencia, este trabajo constituye un antecedente regional directamente relacionado con el cáncer de pulmón, debido a que permite conocer características de pacientes atendidos en Cusco. Sin embargo, su enfoque descriptivo no contempla la optimización de modelos ensemble ni la interpretación mediante SHAP.

Como complemento al contexto regional, la Gerencia Regional de Salud Cusco publicó información epidemiológica sobre las notificaciones de tumores malignos de tráquea, bronquios y pulmón durante el periodo 2021–2025 (Gerencia Regional de Salud Cusco, 2026). En este periodo se registraron 226 notificaciones correspondientes a dicha localización. Dicha información permite contextualizar la presencia de la enfermedad en la región, aunque no constituye un estudio de modelado predictivo ni una base longitudinal individual para estimar

supervivencia; por ello, se emplea como información epidemiológica de contexto.

El Hospital Regional del Cusco también ha publicado información sobre los casos de tumores malignos registrados en su establecimiento durante diferentes años (Hospital Regional del Cusco, 2025). Estos registros permiten reconocer la presencia de pacientes con cáncer de pulmón dentro de la atención hospitalaria regional.

Sin embargo, los datos corresponden a un establecimiento específico y no deben interpretarse como una representación completa de la población de Cusco. Su utilidad consiste en complementar la descripción del contexto regional y mostrar la importancia de contar con información clínica organizada para desarrollar investigaciones sobre el comportamiento de las enfermedades oncológicas.

2.1.4 Síntesis de los antecedentes y vacío de investigación

Los antecedentes internacionales muestran que el aprendizaje automático y el aprendizaje profundo se han utilizado para estimar la supervivencia de pacientes con cáncer de pulmón mediante diferentes fuentes de información clínica. Las investigaciones de Didier et al. (2024), Jin et al. (2023), Astley et al. (2024) y Yuan et al. (2025) permiten reconocer la diversidad de algoritmos y enfoques aplicados a este problema. El trabajo de Jin et al. (2023) resulta especialmente cercano porque emplea SEER y un desenlace de supervivencia con censura, mientras que Cardoso et al. (2026) aporta evidencia sobre la comparación de modelos en datos oncológicos.

En el ámbito nacional, los trabajos de Gálvez-Nino et al. (2023), Gago et al. (2024) y Mamani Barrios (2023) muestran la existencia de investigaciones relacionadas con el análisis clínico y la aplicación de técnicas computacionales en salud. En el ámbito regional, Quispe Rodríguez (2022) aporta información clínica sobre pacientes con cáncer de pulmón atendidos en Cusco, mientras que los boletines epidemiológicos permiten reconocer la presencia de esta enfermedad en la región.

A partir de estos antecedentes, se identifica la necesidad de desarrollar una investigación que se enfoque en la optimización de un modelo ensemble mediante ingeniería de características y selección con SHAP para estimar la supervivencia de pacientes con cáncer

de pulmón, utilizando la base de datos SEER durante el periodo 2015–2023.

En particular, la investigación considera como aspectos relevantes el tratamiento de datos faltantes, la construcción de variables derivadas, la selección de características, el control de fuga de información, la consideración de observaciones censuradas, la optimización de hiperparámetros y la interpretación de las contribuciones de las variables. De esta manera, se busca estudiar el rendimiento del modelo bajo procedimientos de evaluación adecuados para datos de supervivencia y aportar evidencia sobre las características que más contribuyen a sus predicciones.

En la Tabla 2.1 se presenta la síntesis comparativa de los antecedentes internacionales, detallando el tamaño muestral, los modelos o algoritmos implementados y el aporte metodológico de cada investigación respecto al modelado de supervivencia en cáncer de pulmón.

Por su parte, en la Tabla 2.2 se resume la revisión de los antecedentes nacionales desarrollados en el Perú. En ella se contrastan las poblaciones clínicas analizadas, las técnicas estadísticas o de aprendizaje automático empleadas y su relación metodológica con el presente estudio.

Asimismo, en la Tabla 2.3 se consolidan los antecedentes y reportes epidemiológicos regionales de Cusco. Estos registros permiten dimensionar la presencia hospitalaria y sanitaria del cáncer de pulmón en el ámbito local, diferenciando los estudios descriptivos del modelado predictivo propuesto.

2.2 Bases teóricas

2.2.1 Cáncer de pulmón y datos de SEER

El cáncer de pulmón engloba un conjunto diverso de neoplasias malignas con comportamiento biológico y agresividad disímiles. La clasificación anatomopatológica distingue el NSCLC, que representa cerca del 85 % de los casos y comprende principalmente el adenocarcinoma y el carcinoma epidermoide, y el SCLC, que abarca el 15 % restante y se caracteriza por una proliferación celular rápida y un pronóstico desfavorable (Bray et al., 2024; Didier et al., 2024).

Tabla 2.1
Síntesis comparativa de los antecedentes internacionales: muestra, modelos y aporte

Autor y año	**Muestra (*N*)**	**Modelos / Algoritmos**	**Población, desenlace y aporte**
Didier et al. (2024)	12 estudios; 211 068 pacientes	Regresión logística y modelos de ML (metaanálisis)	Pulmón; supervivencia global. Síntesis del rendimiento comparativo. Brecha: heterogeneidad y ausencia de optimización sistemática.
Jin et al. (2023)	$N = 4\,617$ (SEER)	Red neuronal profunda, RSF, Cox	Pulmón (NSCLC estadio III); supervivencia con censura. Antecedente directo del uso de SEER y de modelos de supervivencia.
Astley et al. (2024)	$N = 471$ (Reino Unido)	DeepSurv, RSF, Cox, LIME	Pulmón (NSCLC); supervivencia con censura y calibración. Emplea Brier Score e interpretabilidad; no utiliza SHAP.
Yuan et al. (2025)	$N = 737$ (China)	Random Forest, SVM, CatBoost	Pulmón; desenlace de supervivencia evaluado con clasificación y calibración. Aporta selección de variables; sin optimización comparada.
Cardoso et al. (2026)	Cohorte oncológica multicáncer	LightGBM, XGBoost, Optuna	Modelos ensemble y optimización en datos oncológicos. Antecedente metodológico; no emplea SEER ni análisis SHAP.

Nota. Elaboración propia bajo la norma APA 7.ª edición a partir de los estudios internacionales citados.

Tabla 2.2
Síntesis comparativa de los antecedentes nacionales: muestra, modelos y aporte

Autor y año	**Muestra (*N*)**	**Modelos / Algoritmos**	**Población, desenlace y aporte**
Gálvez-Nino et al. (2023)	$N = 613$ (INEN)	Kaplan-Meier, prueba log-rank	Pulmón avanzado (NSCLC con mutación EGFR); supervivencia descriptiva. Evidencia clínica nacional directa, sin aprendizaje automático.
Gago et al. (2024)	$569 + 1\,980$ registros (UPC)	Random Forest, DT, SVM, NB, GB	Cáncer de mama; clasificación diagnóstica y pronóstica. Precedente nacional en ingeniería; sin análisis de censura en pulmón.
Mamani Barrios (2023)	ENDES 2019–2020 (UNSAAC)	KNN, SVM, redes neuronales	Anemia infantil; clasificación supervisada. Precedente institucional en la UNSAAC sobre datos de salud; sin desenlace oncológico.

Nota. Elaboración propia bajo la norma APA 7.ª edición a partir de las investigaciones nacionales revisadas.

Tabla 2.3
Síntesis comparativa de los antecedentes regionales: muestra, enfoque y aporte

Fuente y año	**Muestra / Registros**	**Enfoque / Método**	**Población, alcance y aporte**
Quispe Rodríguez (2022)	78 pacientes (Hospital Antonio Lorena)	Estudio descriptivo clínico-epidemiológico	Pulmón; caracterización clínica, histológica y sociodemográfica en Cusco (2015–2021). Evidencia regional directa, sin modelado predictivo.
Gerencia Regional de Salud Cusco (2026)	226 notificaciones (2021–2025)	Vigilancia epidemiológica regional	Contexto sanitario regional; notificaciones de tumores malignos de tráquea, bronquios y pulmón. Fuente de contexto epidemiológico.
Hospital Regional del Cusco (2025)	Casos registrados (enero–agosto de 2025)	Registro de vigilancia institucional	Contexto hospitalario regional sobre atención de neoplasias malignas; complementa la justificación práctica local.

Nota. Elaboración propia bajo la norma APA 7.ª edición a partir de las fuentes regionales de Cusco.

La extensión anatómica se evalúa mediante el sistema TNM, que combina el tamaño y la invasión del tumor primario, la afectación de ganglios regionales y la presencia de metástasis a distancia. Aunque esta clasificación define los estadios y orienta las decisiones terapéuticas, resulta insuficiente como predictor univariado de la supervivencia individual, porque no integra la edad, el tipo histológico, el grado tumoral ni otras características demográficas y clínicas del paciente. Por ello, el modelado predictivo requiere considerar un conjunto amplio de variables y evaluar su contribución conjunta.

La base de datos SEER reúne registros de cáncer de base poblacional y proporciona información sobre características demográficas, fecha de diagnóstico, localización y morfología tumoral, estadio, tratamiento y seguimiento de la supervivencia. Su documentación oficial describe los diccionarios de variables, los códigos de la CIE-O-3, las reglas de codificación y los periodos de disponibilidad (National Cancer Institute, 2025). El acceso a estos registros permite construir cohortes retrospectivas y analizar la supervivencia de pacientes oncológicos en un contexto poblacional amplio.

El uso de SEER exige atender la heterogeneidad de codificación entre periodos, la disponibilidad diferencial de variables y la presencia de valores faltantes o categorías especiales. Asimismo, varias variables de tratamiento o de seguimiento pueden registrarse después del diagnóstico, por lo que su incorporación como predictores debe evaluarse con una regla explícita de admisibilidad temporal (van de Mortel & van Wingen, 2025).

2.2.2 Aprendizaje automático sobre datos tabulares y modelos ensemble

En problemas con datos clínicos estructurados, los algoritmos basados en árboles de decisión potenciados por gradiente (GBDT) son ampliamente utilizados por su manejo de variables heterogéneas, valores faltantes y relaciones no lineales, y siguen siendo competitivos o superiores a las redes profundas en datos tabulares (Boldini et al., 2023; Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022). El principio de *boosting* opera de forma aditiva minimizando una pérdida diferenciable $\mathscr{L}$:

$$\hat{y}_i^{(M)} = \sum_{m=1}^{M} f_m(x_i) \tag{2.1}$$

El algoritmo *LightGBM* introduce el muestreo unilateral basado en gradientes y el empaquetado exclusivo de características, con crecimiento de árbol por hojas, lo que reduce el costo computacional y el consumo de memoria (Microsoft Corporation, 2025). Por su parte, *XGBoost* optimiza el entrenamiento mediante una aproximación de Taylor de segundo orden con regularización ℓ_1 y ℓ_2, y admite objetivos de supervivencia (Barnwal et al., 2022; Wiens et al., 2025).

Los modelos ensemble combinan varios aprendices para obtener una predicción agregada. En datos tabulares, las estrategias más frecuentes son el *bagging* de árboles, la potenciación de gradiente y el apilamiento de modelos. Los bosques aleatorios de supervivencia (RSF) constituyen un caso particular de ensemble basado en árboles, en el que cada árbol se ajusta sobre una muestra bootstrap y los nodos se dividen con un criterio adaptado a la censura.

De esta manera, la combinación de predicciones permite reducir la varianza y capturar interacciones complejas, aunque su rendimiento depende de la calidad de las variables, del número de árboles, de la profundidad y de los hiperparámetros de regularización (Grinsztajn et al., 2022; Wiegrebe et al., 2024).

El apilamiento (*stacking*) combina las predicciones de varios modelos base mediante un metamodelo entrenado sobre las salidas de estos. Su implementación requiere una partición interna que evite que el metamodelo se ajuste sobre predicciones generadas con los mismos datos de entrenamiento, porque ello produciría sobreajuste y fuga de información. Por esta razón, cualquier estrategia de ensemble debe documentar las particiones, el procedimiento de combinación y la forma en que se seleccionan los modelos base (Bischl et al., 2023; Collins et al., 2024).

2.2.3 Análisis de supervivencia

El análisis de supervivencia modela la ocurrencia del evento y el tiempo transcurrido hasta el mismo (Denfeld et al., 2023). El comportamiento temporal se define mediante la función de supervivencia $S(t) = P(T > t)$ y la función de riesgo instantáneo $\lambda(t)$:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right), \quad \lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2.2}$$

El modelo semiparamétrico de Cox, base de los métodos de supervivencia actuales, formula $\lambda(t \mid X) = \lambda_0(t)\exp(\beta^T X)$ y estima los coeficientes mediante la verosimilitud parcial:

$$L_p(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(T_i)} \exp(\beta^T X_j)} \tag{2.3}$$

donde $R(T_i)$ es el conjunto de pacientes en riesgo en el tiempo T_i. La red neuronal de supervivencia *DeepSurv* supera la asunción de linealidad de Cox sustituyendo $\beta^T X$ por una función no lineal $h_\theta(X)$ y minimizando la verosimilitud parcial negativa con regularización (Wiegrebe et al., 2024). Asimismo, los bosques aleatorios de supervivencia (RSF) dividen los nodos maximizando la discrepancia mediante un criterio de *log-rank* adaptado a la censura (Wiegrebe et al., 2024).

2.2.4 Censura a la derecha y su tratamiento

Siendo T^* el tiempo real hasta el evento y C el tiempo de censura a la derecha, se observa el par (T, δ):

$$T = \text{mín}(T^*, C), \quad \delta = \mathbb{I}(T^* \leq C) \tag{2.4}$$

El parámetro δ indica si el evento fue observado ($\delta = 1$) o si el seguimiento se interrumpió antes de observarlo ($\delta = 0$). La estimación de la supervivencia utiliza la información parcial de las observaciones censuradas, mientras que una formulación binaria a un horizonte fijo descarta los casos cuyo seguimiento no alcanza ese horizonte, introduciendo un sesgo de selección (Denfeld et al., 2023; Wiegrebe et al., 2024). Por esta razón, el desenlace principal de esta investigación es la supervivencia global con censura a la derecha.

2.2.5 Riesgos competitivos

Cuando el evento de interés es la muerte por cáncer y existen otras causas de muerte que impiden observarlo, las muertes por otras causas no son censuras ordinarias, sino eventos competitivos (Monterrubio-Gómez et al., 2024). Ignorar esta estructura puede sobreestimar la probabilidad acumulada del evento de interés. La presente investigación adopta la supervivencia global como desenlace principal y reserva el análisis de riesgos competitivos como análisis secundario exploratorio, condicionado a que la calidad de la codificación de la causa de muerte permita distinguir las categorías (van Geloven et al., 2022).

2.2.6 Ingeniería de características

La ingeniería de características comprende el conjunto de transformaciones que convierten las variables originales en representaciones más adecuadas para el modelo. Incluye la armonización de códigos, la agrupación de categorías poco frecuentes, la construcción de variables derivadas, la codificación de variables categóricas y el tratamiento de valores faltantes (Büyükkeçeci & Okur, 2023). En datos oncológicos, una representación adecuada puede facilitar la identificación de relaciones no lineales entre la edad, el estadio, la histología y la supervivencia.

La ingeniería de características debe realizarse dentro de cada partición de entrenamiento y repetirse en cada pliegue de validación. Calcular transformaciones con todos los datos antes de separar las particiones introduce fuga de información y produce estimaciones de rendimiento optimistas (van de Mortel & van Wingen, 2025). Por ello, los parámetros de las transformaciones se aprenden exclusivamente con los datos de entrenamiento y se aplican después a validación y prueba.

2.2.7 Selección de características y SHAP

La selección de características busca reducir la dimensionalidad y conservar las variables con mayor capacidad predictiva, evitando el sobreajuste y mejorando la estabilidad del modelo (Büyükkeçeci & Okur, 2023). Los métodos se agrupan en filtros, envolventes e integrados. En modelos clínicos, la selección debe realizarse dentro de la partición de

entrenamiento y repetirse en cada pliegue de validación, porque una selección efectuada con todos los datos filtra información del conjunto de prueba (Takefuji, 2025; van de Mortel & van Wingen, 2025).

SHAP proporciona una descomposición aditiva de la predicción de un modelo en contribuciones de sus características, basada en los valores de Shapley de la teoría de juegos cooperativos (Lundberg & Lee, 2017; Salih et al., 2024). Para un modelo f y una instancia x, la explicación se expresa como:

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j \tag{2.5}$$

donde ϕ_0 es el valor base y ϕ_j es la contribución de la característica j a la predicción. Los valores SHAP permiten analizar la importancia global mediante el promedio de las magnitudes de las contribuciones y la importancia local en predicciones individuales. Su cálculo exacto puede ser costoso; por ello, para modelos basados en árboles se emplean implementaciones eficientes como TreeSHAP. La atribución SHAP describe cómo el modelo utiliza las variables y no establece relaciones causales entre ellas y el desenlace.

Cuando SHAP se utiliza para apoyar la selección de características, el procedimiento debe definirse explícitamente y ejecutarse únicamente con datos de entrenamiento o validación. El conjunto de prueba no debe emplearse para seleccionar variables ni para ajustar hiperparámetros. Además, cualquier mejora atribuida a la selección basada en SHAP debe demostrarse mediante una comparación experimental adecuada, porque SHAP no mejora automáticamente el rendimiento del modelo.

2.2.8 Optimización de hiperparámetros

La optimización de hiperparámetros es el procedimiento mediante el cual se selecciona, entre un conjunto de configuraciones posibles de un modelo, la que produce el mejor rendimiento según un criterio definido sobre los datos de validación (Bischl et al., 2023). Sus componentes son: el espacio de búsqueda, la función objetivo, el presupuesto de búsqueda y la estrategia de búsqueda, que puede ser exhaustiva, aleatoria o adaptativa, como la búsqueda bayesiana y las variantes de validación sucesiva empleadas en herramientas como *Optuna*

(Optuna Developers, 2025).

Dos riesgos condicionan el diseño: un presupuesto insuficiente no llega a explorar el espacio y reporta una configuración arbitraria; un presupuesto elevado sin partición separada de validación conduce a sobreajuste de selección, es decir, a elegir la configuración que mejor se ajusta al ruido de la validación (Tetko et al., 2024). Por ello, la optimización se ejecuta exclusivamente sobre las particiones de entrenamiento y validación, y la configuración ganadora se congela antes de evaluarla en el conjunto de prueba.

2.2.9 Validación y evaluación en datos independientes

La separación en particiones cumple funciones distintas: el conjunto de entrenamiento ajusta los pesos o los árboles; el de validación guía la selección de hiperparámetros y de características; y el de prueba produce la estimación final del rendimiento de la configuración ya elegida (Efthimiou et al., 2024). Mezclar estas funciones produce estimaciones optimistas, porque la información del conjunto de prueba habría participado en la selección.

Cuando los datos tienen estructura temporal, la partición aleatoria simple puede filtrar información futura hacia el entrenamiento; en esos casos se prefieren esquemas que respeten el orden cronológico del registro, además de la validación cruzada convencional cuando el tamaño de la muestra lo permite (Riley et al., 2024). El estándar internacional TRIPOD+AI (Collins et al., 2024) prescribe reportar explícitamente el diseño de desarrollo y validación, el manejo de datos faltantes, las métricas con sus incertidumbres y las decisiones de modelado, condiciones que se adoptan como referencia de esta investigación.

2.2.10 Métricas de rendimiento

Discriminación. La capacidad de ordenar a los pacientes según su riesgo se mide mediante el C-index de Harrell, cuyas limitaciones de interpretación y uso se discuten en Hartman et al. (2023) y Sonabend et al. (2022), y cuya descomposición se analiza en Alabdallah et al. (2024):

$$C = \frac{\sum_{i \neq j} \mathbb{I}(T_i < T_j) \cdot \mathbb{I}(\hat{R}_i > \hat{R}_j) \cdot \delta_i}{\sum_{i \neq j} \mathbb{I}(T_i < T_j) \cdot \delta_i} \tag{2.6}$$

Calibración. La precisión de las probabilidades predichas se evalúa con el ***Brier Score*** dependiente del tiempo (Wiegrebe et al., 2024; Zhu et al., 2025), ponderado por la función de censura $\hat{G}(t)$:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(0-\hat{S}(t \mid X_i))^2 \cdot \mathbb{I}(T_i \leq t, \delta_i = 1)}{\hat{G}(T_i^-)} + \frac{(1-\hat{S}(t \mid X_i))^2 \cdot \mathbb{I}(T_i > t)}{\hat{G}(t)}\right] \tag{2.7}$$

cuya integral define el IBS. El AUC dependiente del tiempo se estima con métodos que incorporan la censura y la correlación entre observaciones (Beyene & Chen, 2024). La evaluación de calibración puede complementarse con la pendiente y la intersección de calibración en horizontes clínicamente interpretables.

La **métrica principal de contraste** se fija antes de ejecutar los experimentos: el C-index de Harrell sobre el conjunto de prueba. El área bajo la curva dependiente del tiempo, el *Brier Score* y el IBS se establecen como métricas de discriminación y calibración complementarias. No se utiliza la exactitud (*accuracy*) como métrica principal, porque no representa adecuadamente el tiempo hasta el evento ni la censura. Todas las métricas se reportan en el conjunto de prueba independiente.

2.2.11 Interpretabilidad de los modelos

Para garantizar la interpretabilidad clínica, se emplean los valores SHAP (Lundberg & Lee, 2017; Salih et al., 2024), basados en la teoría de juegos cooperativos, que calculan la contribución marginal aditiva de cada variable a la predicción individual. Su uso persigue dos finalidades: identificar los factores de riesgo clínico de mayor peso y verificar que los predictores relevantes correspondan a información disponible en el momento basal. La atribución SHAP describe cómo el modelo utiliza las variables; no establece relaciones causales entre ellas y el desenlace.

2.2.12 Fuga de información

La fuga de información (*data leakage* o *look-ahead bias*) ocurre cuando una variable disponible solo después del momento en que se emite la predicción participa en el entrena-

miento o en la evaluación. En datos clínicos oncológicos el caso típico son las variables de seguimiento o de desenlace, como el estado vital final y el tiempo de supervivencia, que elevan el rendimiento en laboratorio sin ser utilizables en la práctica (van de Mortel & van Wingen, 2025).

Por ello, su control se aborda mediante la definición explícita del momento basal, la verificación de las marcas de tiempo de cada variable y la exclusión de todo lo que no esté disponible en ese instante; esta auditoría forma parte del objetivo específico de preparación de datos (OE1).

2.2.13 Reproducibilidad

La reproducibilidad exige que otro investigador, con los mismos datos y el mismo procedimiento, obtenga resultados equivalentes. En este trabajo se concreta en: registro de la versión de SEER y de los criterios de selección; fijación de semillas y de las particiones; documentación del espacio y del presupuesto de búsqueda; publicación del código de la canalización de experimentos, y reporte conforme a TRIPOD+AI (Collins et al., 2024). Toda cifra presentada como resultado de un experimento deberá provenir de la ejecución documentada de dicho procedimiento; los resultados no se anticipan en este plan de tesis.

2.3 Definición de términos básicos

- **Configuración de referencia (línea base):** valores iniciales de los hiperparámetros de un modelo, elegidos por defecto o documentados por sus autores, con los cuales se compara el rendimiento del modelo optimizado.
- **Conjunto de entrenamiento, validación y prueba:** divisiones mutuamente excluyentes de los datos destinadas, respectivamente, a ajustar el modelo, a guiar la selección de hiperparámetros y de características, y a producir la evaluación final.
- **Espacio de búsqueda:** conjunto de hiperparámetros que se someten a optimización y de los rangos de valores permitidos para cada uno.
- **Hiperparámetro:** parámetro externo del algoritmo que no se aprende de los datos

durante el entrenamiento, sino que se fija antes de él (por ejemplo, profundidad del árbol, tasa de aprendizaje o número de estimadores).

- **Ingeniería de características:** proceso de transformación, combinación y construcción de variables a partir de los datos originales para mejorar su representación en el modelo.
- **Modelo ensemble:** modelo que combina las predicciones de varios aprendices para obtener una estimación agregada.
- **Optimización de hiperparámetros:** procedimiento sistemático de búsqueda de la configuración que maximiza una métrica de validación dentro de un presupuesto definido.
- **Partición temporal:** división de los datos que respeta el orden cronológico del registro, asignando periodos anteriores al entrenamiento y posteriores a la evaluación.
- **Presupuesto de búsqueda:** número máximo de configuraciones evaluadas durante la optimización.
- **Reproducibilidad:** capacidad de repetir un experimento obteniendo resultados equivalentes al documentar datos, particiones, semillas y procedimiento.
- **Sobreajuste de selección:** sesgo producido por elegir la mejor configuración de muchas candidatas atendiendo a su rendimiento en la validación, sin comprobación posterior en un conjunto independiente.
- **Valores SHAP:** contribuciones aditivas de las características a la predicción de un modelo, calculadas a partir de la teoría de juegos cooperativos.
- **Validación cruzada:** técnica que divide los datos de entrenamiento en k partes y repite el ajuste y la evaluación para obtener una estimación estable del rendimiento durante la búsqueda.

Capítulo III

Diseño metodológico y experimental

Este capítulo define el procedimiento mediante el cual se construye la cohorte, se formaliza el desenlace de supervivencia, se auditan los predictores, se realiza la ingeniería de características, se entrena y optimiza un modelo ensemble y se evalúa su rendimiento en un conjunto de datos independiente. Las decisiones aquí descritas se fijan antes de examinar el conjunto de prueba, conforme a las directrices TRIPOD+AI (Collins et al., 2024).

La definición de los campos se apoya en la documentación oficial de la base de datos SEER, administrada por el NCI, y en la clasificación CIE-O-3 (National Cancer Institute, 2025). La versión específica de SEER, la fecha de corte del seguimiento y los diccionarios de variables se registran al ejecutar la extracción. El capítulo no contiene métricas obtenidas, ni tamaños muestrales propios, ni afirmaciones sobre el efecto de la optimización: todo ello depende de la obtención de los microdatos y de la ejecución del procedimiento descrito.

3.1 Tipo, nivel y diseño de investigación

3.1.1 Enfoque de investigación

La investigación tendrá un enfoque cuantitativo, debido a que utilizará datos estructurados de pacientes con cáncer de pulmón registrados en SEER y aplicará procedimientos estadísticos y computacionales para desarrollar, optimizar y evaluar un modelo predictivo de supervivencia. Se utilizarán métricas numéricas para evaluar el desempeño del modelo y técnicas de interpretación para analizar la contribución de las características a las predicciones.

3.1.2 Tipo y nivel de investigación

La investigación será de tipo aplicada, porque busca desarrollar una solución computacional para un problema específico relacionado con la predicción de supervivencia en pacientes con cáncer de pulmón. El nivel será predictivo, debido a que se pretende estimar la supervivencia de los pacientes mediante un modelo ensemble optimizado. Asimismo, se

realizará una fase descriptiva para examinar las características de los registros utilizados.

3.1.3 Diseño de investigación

El diseño será no experimental, retrospectivo y longitudinal, debido a que se analizarán registros secundarios existentes en la base de datos SEER, sin manipular las condiciones clínicas de los pacientes ni intervenir sobre ellos. La dimensión longitudinal se relaciona con el análisis del tiempo hasta el evento de interés y la consideración de observaciones censuradas.

El desarrollo computacional incluirá experimentos controlados para comparar configuraciones del modelo, evaluar estrategias de ingeniería de características y analizar los resultados de SHAP. Estos experimentos se realizarán sobre los datos disponibles, sin constituir una intervención experimental sobre pacientes.

El procedimiento se organizará en tres fases:

1. **Preparación de datos:** selección de registros, definición de criterios de inclusión y exclusión, tratamiento de valores faltantes, codificación de variables, transformación de datos e ingeniería de características.

2. **Construcción y optimización del modelo:** desarrollo del modelo ensemble, selección de características según el procedimiento establecido, ajuste de hiperparámetros y validación.

3. **Evaluación e interpretación:** evaluación del desempeño mediante métricas de supervivencia y análisis de las contribuciones de las características utilizando SHAP.

3.2 Hipótesis de investigación

Las hipótesis se formulan después de haber definido el desenlace, los modelos y las métricas con los que se contrastarán. Como desenlace primario se toma la **supervivencia global**, entendida como el tiempo desde el diagnóstico hasta la muerte por cualquier causa, con censura a la derecha. Como **métrica principal de contraste** se emplea el C-index de Harrell, con el *Brier Score*, el IBS y el área bajo la curva dependiente del tiempo como

métricas complementarias. Estas hipótesis son objeto de contraste según el procedimiento de la sección 3.14; este documento no presenta experimentos ejecutados ni asume su resultado.

3.2.1 Hipótesis general

HG: La **optimización del rendimiento predictivo** de un **modelo ensemble mediante ingeniería de características y selección con SHAP** permite mejorar la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**.

3.2.2 Hipótesis específicas

HE1: El **preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** permiten identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**.

HE2: El **desarrollo y la optimización** de un **modelo ensemble mediante algoritmos de aprendizaje automático** mejoran el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**.

HE3: La aplicación de **SHAP** permite **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**.

3.3 Población, muestra y unidad de análisis

La **población fuente** la constituyen los registros de pacientes con cáncer de pulmón disponibles en la base de datos SEER, administrada por el NCI de Estados Unidos (National Cancer Institute, 2025). La unidad de análisis es el registro oncológico del paciente. La definición de unicidad se establecerá de acuerdo con los identificadores disponibles en la versión de SEER utilizada y se documentará antes de modelar.

Las muestras de trabajo se forman por aplicación secuencial de criterios (Tabla 3.1). Ningún tamaño se declara antes de ejecutar la extracción. El número de registros retirados en cada paso se reportará mediante un diagrama de flujo generado por el propio script de

construcción de la cohorte.

Tabla 3.1
Conjuntos de análisis y regla de formación de cada uno

Conjunto	Regla de formación	Uso
S_0	Registros de SEER con localización primaria de tráquea, bronquios y pulmón, identificados mediante los códigos de topografía de la CIE-O-3, sin más filtros.	Conteo de partida; verificación de la cobertura de la versión de SEER.
S_1	S_0 aplicando los criterios de inclusión y exclusión de la sección 3.5, incluida la verificación de fechas y de la disponibilidad de las variables.	Cohorte elegible; base para todos los análisis posteriores.
S_2	S_1 con tiempo de supervivencia y estado vital construibles a partir de las variables documentadas de SEER.	Análisis de supervivencia (desenlace primario) y análisis SHAP.

Nota. Los tamaños $|S_0|$, $|S_1|$ y $|S_2|$, así como el número de registros retirados en cada paso, se reportan con un diagrama de flujo generado por el propio script de construcción de la cohorte. No se declara ningún tamaño antes de ejecutar la extracción.

3.4 Fuente de datos y periodo disponible

La fuente de datos es la base de datos SEER, cuyos registros y diccionarios de variables son publicados por el NCI (National Cancer Institute, 2025). La extracción comprenderá los diagnósticos registrados entre el 1 de enero de 2015 y el 31 de diciembre de 2023, conforme a la delimitación del Capítulo I. Se registrarán la versión de SEER utilizada, la fecha de liberación de los datos y la fecha de corte del seguimiento.

El periodo de diagnóstico no se interpreta como periodo de seguimiento. Un paciente diagnosticado en 2023 no cuenta necesariamente con cinco años de observación; su tiempo de seguimiento depende de la fecha de corte y de su estado vital. Por ello, antes de modelar se reportarán la distribución del tiempo de seguimiento, la proporción de eventos y de censuras, la mediana de seguimiento estimada por Kaplan-Meier y la disponibilidad de cada variable por año de diagnóstico.

La construcción de la variable de tiempo seguirá las definiciones documentadas de SEER, verificando la unidad de tiempo (meses o días) y su conversión. No se asumirá que todos los registros cuentan con el mismo horizonte de seguimiento ni se declarará cobertura

superior a la documentada por la versión utilizada.

3.5 Criterios de inclusión y exclusión

Los criterios se aplican en orden y se documenta el número de registros retirados en cada paso.

Criterios de inclusión:

1. Localización primaria de tráquea, bronquios y pulmón, identificada mediante los códigos de topografía de la CIE-O-3 correspondientes a cáncer de pulmón, de acuerdo con la documentación de SEER.
2. Fecha de diagnóstico dentro de la ventana 2015–2023.
3. Edad al diagnóstico igual o superior a 18 años, salvo que la documentación de SEER o el diseño del estudio justifiquen otro límite.
4. Confirmación microscópica del diagnóstico, cuando la variable correspondiente esté disponible y su uso sea metodológicamente válido.
5. Tiempo de supervivencia y estado vital construibles a partir de las variables documentadas de SEER.
6. Registro con información suficiente para identificar el evento y la censura.

Criterios de exclusión:

1. Registros con tiempo de supervivencia faltante o no válido.
2. Fechas imposibles o incoherentes.
3. Registros correspondientes a tumores secundarios o metástasis de otros órganos, de acuerdo con la topografía y la morfología.
4. Variables identificativas personales, que no se incorporan al extracto de trabajo.
5. Registros con códigos de estado vital desconocido que impidan definir el evento o la censura.

La disponibilidad de cada variable se verifica contra el diccionario de la versión de SEER seleccionada. No se asume que todas las variables estén disponibles para todos los años del periodo 2015–2023. Las variables con cobertura insuficiente se excluyen del modelo principal y se analizan, si corresponde, en un análisis de sensibilidad.

3.6 Definición operacional del desenlace

3.6.1 Origen temporal, tiempo y estado

El desenlace se define con tres elementos documentados en SEER:

- **Origen temporal:** la fecha de diagnóstico. Todas las predicciones se emiten en ese momento.
- **Tiempo hasta el evento o último seguimiento:** el tiempo de supervivencia registrado en SEER, convertido a una unidad común y verificado contra la documentación de la versión utilizada.
- **Indicador del evento:** $\delta = 1$ cuando se registra la muerte por cualquier causa y $\delta = 0$ cuando el paciente permanece vivo al último contacto o se encuentra censurado administrativamente.

3.6.2 Tipo de desenlace y censura

El **desenlace primario** es la **supervivencia global**, entendida como el tiempo desde el diagnóstico hasta la muerte por cualquier causa, con **censura a la derecha** en los pacientes vivos al último contacto. Formalmente se observa el par (T, δ) con $T = \text{mín}(T^*, C)$, donde T^* es el tiempo real hasta la muerte y C el tiempo de censura.

La supervivencia global se diferencia de la supervivencia específica por cáncer. Esta última considera únicamente las muertes atribuibles a la neoplasia y requiere información de causa de muerte cuya calidad debe verificarse. En esta investigación, la supervivencia global constituye el desenlace principal; si se utiliza una definición distinta, deberá justificarse y documentarse.

3.6.3 Horizontes de evaluación y seguimientos incompletos

Los horizontes de evaluación para las métricas dependientes del tiempo se fijarán antes de ejecutar los experimentos, considerando la distribución del seguimiento y la fecha de corte. Se reportarán, por ejemplo, horizontes de 12, 36 y 60 meses cuando la información disponible lo permita, sin asumir que todos los pacientes alcanzan el horizonte máximo.

Las reglas de tratamiento del seguimiento incompleto son las siguientes:

1. Los pacientes vivos al último contacto se censuran en su tiempo de seguimiento y se conservan en el análisis: su información parcial es legítima y su descarte introduciría un sesgo de selección (Denfeld et al., 2023).

2. La censura administrativa se establece en la fecha de corte del extracto; ningún paciente se sigue más allá de esa fecha.

3. Los tiempos nulos o improbables se revisan contra la documentación de SEER y se tratan mediante una regla fijada antes de modelar.

4. Antes de modelar se reportan la distribución del tiempo de seguimiento, la proporción de eventos y de censuras y la mediana de seguimiento estimada por Kaplan-Meier.

3.7 Predictores y control de la fuga de información

3.7.1 Momento de predicción y regla de admisibilidad

El momento de predicción es el **diagnóstico**. Una variable se admite como predictor solo si cumple simultáneamente tres condiciones: (i) está disponible en el registro en ese momento; (ii) no forma parte del desenlace ni de su definición; (iii) no es un descendiente del curso posterior de la enfermedad. Todo lo demás se excluye, se justifica y se documenta (van de Mortel & van Wingen, 2025).

3.7.2 Auditoría de las variables de desenlace

Las variables de estado vital, tiempo de supervivencia, causa de muerte y fecha de último contacto constituyen el desenlace y no pueden utilizarse como predictores. Tampoco

se derivan de ellas variables proxy, como «tiempo de seguimiento» o «años desde el último contacto». Su uso se limita a la construcción del desenlace y a los controles de calidad temporales.

3.7.3 Conjuntos de predictores

Se consideran como posibles variables predictoras la edad al diagnóstico, el sexo, la raza o grupo étnico de acuerdo con la codificación disponible, el año de diagnóstico, la localización del tumor, el tipo histológico, el grado tumoral, el tamaño del tumor si está disponible, el estadio de la enfermedad, la extensión regional, la presencia de metástasis y el tratamiento recibido cuando esté disponible y su uso sea metodológicamente válido.

No se asume que todas las variables estén disponibles para todos los años del periodo 2015–2023; su disponibilidad se verifica en el diccionario de datos y en la documentación de la versión SEER seleccionada, tal como se resume en la Tabla 3.2.

Por su parte, las variables de tratamiento se incluyen únicamente si su registro corresponde al momento basal o si pueden construirse sin incorporar información posterior al diagnóstico. Cuando la fecha de tratamiento no sea verificable, la variable se excluye del modelo principal y se analiza en un análisis de sensibilidad. Asimismo, en la Tabla 3.3 se detallan las variables excluidas del modelo por producir fuga de información o constituir derivaciones directas del desenlace.

3.8 Preprocesamiento e ingeniería de características

El preprocesamiento se ejecuta como una secuencia versionada de pasos, cada uno de los cuales deja registro de los registros afectados:

1. **Lectura y tipado:** conversión explícita de fechas y variables numéricas, y verificación de dominios categóricos contra el diccionario de SEER (National Cancer Institute, 2025).

2. **Controles de coherencia:** fecha de diagnóstico dentro de la ventana declarada; tiempo de supervivencia no negativo; edad dentro de un rango admisible; frecuencia de valores

Tabla 3.2

Predictores admisibles al momento basal y su tratamiento

Bloque	Variables de SEER	Tratamiento
Demográficos	Edad al diagnóstico, sexo, raza o grupo étnico y año de diagnóstico, de acuerdo con la codificación disponible.	Edad continua y categorías clínicas; sexo y raza como categóricas; año de diagnóstico como variable numérica y, si se justifica, por periodos.
Tumor y localización	Localización primaria, morfología y grado tumoral codificados según la CIE-O-3.	Topografía restringida a cáncer de pulmón; histología agrupada en categorías clínicas amplias; grado armonizado y categorías poco frecuentes agrupadas.
Estadio y extensión	Estadio derivado, extensión regional, tamaño tumoral y presencia de metástasis, cuando estén disponibles para el año de diagnóstico.	Estadio ordinal; variables derivadas de extensión y metástasis; tamaño tumoral numérico y por categorías si su cobertura es suficiente.
Tratamiento	Indicadores de cirugía, radioterapia y quimioterapia, cuando su registro corresponda al momento basal y su codificación sea verificable.	Indicadores binarios o categóricos; se excluyen del modelo principal si la fecha de tratamiento no permite verificar su disponibilidad basal.

Nota. La disponibilidad y la codificación de cada variable se verifican en el diccionario de la versión de SEER seleccionada (National Cancer Institute, 2025). No se asume que todas las variables estén disponibles para todos los años del periodo 2015–2023.

Tabla 3.3
Variables excluidas del modelo por fuga de información o por derivación del desenlace

Variables	Naturaleza temporal	Motivo de exclusión
Tiempo de supervivencia en meses y estado vital final	Definen el desenlace	Constituyen el evento y el tiempo; usarlos como predictor anularía el problema predictivo.
Clasificación de muerte específica por cáncer y causa de muerte	Posteriores al diagnóstico	Revelan el desenlace y solo están disponibles después del seguimiento.
Fecha de último contacto y estado en el último contacto	Definen la censura	Su uso como predictor introduce fuga de información y desplaza el momento de predicción.
Variables de recidiva o progresión, si están disponibles	Posteriores al diagnóstico	La recidiva ocurre después de la línea base y condiciona la supervivencia.
Identificadores personales o geográficos excesivamente detallados	Identificativos	No aportan valor pronóstico y su inclusión puede comprometer la confidencialidad.

Nota. Toda variable se admite únicamente si su disponibilidad es anterior o igual al momento basal y si no es un descendiente del desenlace (van de Mortel & van Wingen, 2025). La lista se verifica contra el diccionario de la versión de SEER utilizada.

de código reservado por variable y por año de diagnóstico.

3. **Duplicados y unicidad:** aplicación de la regla de unicidad definida para la versión de SEER utilizada, con reporte del número de duplicados detectados.

4. **Armonización de la morfología:** recodificación de la histología en categorías clínicas amplias (carcinoma de células pequeñas frente a no pequeñas, con sus subtipos principales), en lugar de códigos finos, y verificación de la edición de la CIE-O-3 aplicable al periodo.

5. **Codificación y escalado:** la codificación de categóricas y la estandarización se aprenden exclusivamente con la partición de entrenamiento de cada pliegue y se aplican después; para los modelos basados en árboles se emplean códigos enteros sin escalar.

6. **Valores faltantes:** las categorías «sin información» y «no se aplica» se conservan como categorías informativas cuando el diccionario lo permita. Si fuera necesaria una imputación numérica, se realiza con mediana o moda del entrenamiento, dentro del pliegue (Nijman et al., 2022).

7. **Control de fuga por construcción:** ninguna transformación, selección de categorías, umbral ni escala se calcula con información de validación o de prueba (van de Mortel & van Wingen, 2025); el código de preprocesamiento es un transformador entrenable que se ajusta dentro del pliegue correspondiente.

3.8.1 Ingeniería de características

La ingeniería de características se define antes de modelar y se aplica de forma consistente en todas las particiones. Las transformaciones previstas son:

- **Edad:** edad continua al diagnóstico y categorías clínicas (por ejemplo, grupos etarios) para capturar relaciones no lineales.
- **Año de diagnóstico:** variable numérica y, si se justifica, categorías por periodo para representar cambios en la práctica diagnóstica y en la codificación.
- **Estadio y extensión:** codificación ordinal del estadio y variables derivadas de extensión regional y metástasis, cuando estén disponibles.
- **Histología y localización:** agrupación de categorías poco frecuentes y construcción de indicadores para los subtipos principales.
- **Tamaño tumoral:** transformación numérica y categorías clínicas, si la variable está disponible y su cobertura es suficiente.
- **Interacciones:** términos seleccionados entre edad, estadio, histología y localización, definidos a priori y evaluados en la ablación.
- **Tratamiento:** indicadores de tratamiento recibido, cuando su disponibilidad basal y su codificación sean verificables.

El número de variables derivadas se mantiene acotado para evitar un aumento innecesario de la dimensionalidad. Cada transformación se documenta con su fórmula, su justificación clínica y su tratamiento de valores faltantes. La comparación entre el conjunto original y el conjunto con ingeniería de características permite aislar el efecto de esta fase.

3.8.2 Selección de características y ablación

Para contrastar HE1 se define una **ablación pre-registrada** con tres configuraciones del mismo modelo y la misma partición:

1. **R0 (referencia de preprocesamiento):** predictores básicos, codificación categórica directa y sin ingeniería de características ni selección explícita.

2. **R1 (ingeniería de características):** predictores básicos más las variables derivadas de la sección anterior, con la secuencia de preprocesamiento, pero sin selección de características.

3. **R2 (ingeniería de características y selección):** R1 más selección de características ajustada dentro de cada pliegue de entrenamiento.

La selección de R2 puede combinar un filtro univariado y un criterio basado en la importancia de las variables. Si se emplea SHAP para apoyar la selección, el procedimiento se define explícitamente: se ajusta el modelo ensemble con los datos de entrenamiento del pliegue, se calculan los valores SHAP sobre la partición de validación interna y se retienen las variables cuya contribución acumulada alcance un umbral pre-registrado.

En ningún caso el conjunto de prueba participa en esta selección. De este modo, la comparación R0–R1 aísla el efecto de la ingeniería de características y R1–R2 el de la selección, estimándose ambas diferencias con la métrica principal y con bootstrap emparejado.

3.9 Pipeline experimental

El procedimiento completo se ejecuta en diez etapas, en este orden, y cada una deja un artefacto verificable:

Algoritmo III.1: Pipeline experimental de diez etapas

1. **Construcción documentada de la cohorte:** script que aplica los criterios de la sección 3.5 y produce los conjuntos S_0–S_2 con su diagrama de flujo.

2. **Auditoría del desenlace y de la fuga:** verificación de la construcción de (T, δ), de la lista de variables excluidas y de la disponibilidad temporal de cada predictor.

3. **Preprocesamiento e ingeniería de características por pliegue:** transformadores ajustados únicamente con el entrenamiento (sección 3.8).

4. **Partición en entrenamiento, validación y prueba:** congelada y persistida como archivo, con su semilla y su regla de formación (sección 3.12).

5. **Entrenamiento de las configuraciones de referencia:** modelo ensemble con valores por defecto o documentados, sin búsqueda alguna.

6. **Fijación de los espacios de búsqueda y del presupuesto:** Tabla 3.4, registrada antes de la primera búsqueda.

7. **Optimización:** búsqueda sobre entrenamiento y validación, con registro de cada ensayo (sección 3.11).

8. **Selección y congelamiento:** se elige una configuración, se fija el número de iteraciones con la validación y se guarda el modelo final.

9. **Evaluación independiente:** una única pasada sobre el conjunto de prueba, que hasta este momento permanece inaccesible al procedimiento de búsqueda.

10. **Contraste, interpretabilidad y registro:** comparación emparejada de configuraciones (sección 3.14), análisis SHAP (sección 3.15) y registro de experimentos (sección 3.16).

La separación entre las etapas de optimización y la evaluación es el mecanismo central de protección del conjunto de prueba: la búsqueda nunca consulta ese conjunto, y la evaluación se ejecuta una sola vez por configuración congelada (Barreñada et al., 2024).

3.10 Modelo ensemble y configuración de referencia

No se persigue una colección exhaustiva de algoritmos, sino un conjunto mínimo que permita construir un modelo ensemble para el desenlace de supervivencia y contrastar el efecto de la ingeniería de características y de la optimización. Se definen los siguientes componentes:

- **Modelo de riesgos proporcionales de Cox** (Wiegrebe et al., 2024), en su versión penalizada (*ridge* o red elástica neta) (Dai & Breheny, 2024). Actúa como modelo de referencia interpretable y como línea base explicativa.
- ***Random Survival Forests*** (Wiegrebe et al., 2024), que particiona los nodos con un criterio de log-rank adaptado a la censura y captura interacciones sin asumir linealidad.
- ***XGBoost* con objetivo de supervivencia**, formulado preferentemente como modelo de tiempo de fallo acelerado (AFT) (Barnwal et al., 2022), que no exige proporcionalidad de riesgos y admite observaciones censuradas.
- **Modelo ensemble** construido a partir de los aprendices anteriores mediante promediado de predicciones o apilamiento con un metamodelo regularizado. La estrategia de combinación, los modelos base y el procedimiento de validación interna se fijan antes de la optimización.

La **configuración de referencia** del ensemble se define con valores por defecto o documentados de los modelos base y una regla de combinación simple, sin búsqueda de hiperparámetros. La configuración optimizada se obtiene mediante la búsqueda descrita en la sección 3.11. No se atribuye superioridad al ensemble frente a los modelos individuales antes de obtener los resultados; la comparación se reporta como parte del contraste experimental.

3.11 Optimización de hiperparámetros

3.11.1 Función objetivo y selección

La función objetivo es la **media del *C*-index de Harrell** (Alabdallah et al., 2024) en la partición de validación interna. La calibración no se incorpora a la función objetivo: se reserva para la evaluación, de modo que la optimización responda a una sola dimensión y la degradación de la calibración pueda detectarse y reportarse como resultado secundario.

3.11.2 Estrategia y presupuesto

La estrategia principal es la **búsqueda bayesiana por proceso de estimadores de**

tamaños de muestras (TPE) implementada en *Optuna* (Optuna Developers, 2025), basada en la evidencia de que la búsqueda aleatoria bien acotada supera a la búsqueda en cuadrícula y que los métodos adaptativos aprovechan mejor el presupuesto (Bischl et al., 2023). Como estrategia de referencia se ejecuta, con el mismo presupuesto, una **búsqueda aleatoria**.

El presupuesto pre-registrado es de **50 ensayos por configuración**, con poda mediana a partir de 5 evaluaciones intermedias, límite de tiempo por ensayo y detención temprana dentro de cada ensayo. Estos valores son decisiones de diseño: si el cómputo disponible o la estructura de los datos obligan a modificarlos, el cambio se decide y se registra antes de examinar el conjunto de prueba, con su justificación.

3.11.3 Selección del modelo final

Se selecciona la configuración con mejor media de la función objetivo en la validación interna; en caso de empate se prefiere la menor IBS y, después, la configuración de menor complejidad. Se fija entonces el número de iteraciones mediante detención temprana sobre la validación, se reentrena el modelo con esa configuración y ese número de iteraciones, y se congela antes de la evaluación.

El número total de configuraciones exploradas se reporta junto con el rendimiento obtenido, dado que el sobreajuste de selección crece con el presupuesto empleado (Barreñada et al., 2024). En la Tabla 3.4 se especifican los espacios de búsqueda y los tipos de muestreo pre-registrados para cada componente del modelo ensemble.

3.12 Partición y validación

3.12.1 Partición principal

La partición primaria respeta el orden cronológico del registro: los pacientes se ordenan por año de diagnóstico y se asignan aproximadamente los primeros 60 % al entrenamiento, los 20 % siguientes a validación y el 20 % final a prueba, con los cortes fijados en años completos y registrados antes de modelar. Este esquema evita que información de periodos posteriores entre en el entrenamiento y constituye una validación temporal (Riley et al., 2024).

Tabla 3.4
Espacios de búsqueda pre-registrados para la optimización del modelo ensemble

Componente	**Hiperparámetros y rangos permitidos**	**Tipo de muestreo**
Cox penalizado	$\lambda \in [10^{-4}, 10^{2}]$ (logarítmico) y $\alpha \in [0,1]$ (mezcla elástica neta); enlace con la penalización de *ridge* cuando $\alpha = 0$.	Log-uniforme / uniforme
Random Survival Forest	`n_estimators` $\in [200,1000]$; `max_depth` $\in [3,12]$; `min_samples_split` $\in [2,20]$; `min_node_size` $\in [1,30]$; `max_features` $\in \{\sqrt{p}, p/3, 0.5p\}$.	Entera, uniforme y categórica
XGBoost (supervivencia)	`n_estimators` ≤ 2000 con detención temprana; tasa de aprendizaje $\in [10^{-3}, 3\times10^{-1}]$ (logarítmico); `max_depth` $\in [2,12]$; `min_child_weight` $\in [1,20]$; `subsample` $\in [0.5,1]$; `colsample_bytree` $\in [0.5,1]$; ℓ_1 y $\ell_2 \in [10^{-8}, 10^{2}]$ (logarítmicos); escala de la pérdida AFT $\in [0.5,2]$.	Mixta
Metamodelo del ensemble	Tipo de combinación (promedio o apilamiento); regularización del metamodelo $\in [10^{-6}, 10^{1}]$ (logarítmica); número de pliegues internos de apilamiento $\in \{3,5\}$.	Categórica y log-uniforme

Nota. La estrategia de combinación y los modelos base se fijan antes de la búsqueda. Cualquier modificación de estos rangos debe registrarse antes de examinar el conjunto de prueba y justificarse por restricciones de cómputo o por la estructura observada de los datos.

Como alternativa documentada, si la distribución de años de diagnóstico o el número de eventos no permite sostener una partición temporal, se recurre a una **división aleatoria estratificada** por estado del evento, con semilla fija. En ambos casos se registra el esquema utilizado, sus proporciones y su semilla.

3.12.2 Validación cruzada interna

Dentro de la partición de entrenamiento se emplea validación cruzada de 5 pliegues, estratificados por estado del evento y bloqueados temporalmente cuando la partición principal es temporal. Los pliegues guían la búsqueda de hiperparámetros y la estimación de la variabilidad; el preprocesamiento y la ingeniería de características se ajustan de nuevo en cada pliegue.

3.12.3 Uso del conjunto de prueba y robustez

El conjunto de prueba no interviene en la selección de hiperparámetros, de variables ni de número de iteraciones; se evalúa una sola vez por configuración congelada. Adicionalmente,

la configuración ganadora y su configuración de referencia se repiten con **cinco semillas distintas** para describir la variabilidad debida al azar del entrenamiento. Esta repetición es un análisis de sensibilidad descriptivo: lo que se reporta es el rango y la desviación de las métricas entre semillas.

3.13 Métricas de evaluación

La **métrica principal de contraste se fija antes de los experimentos**: el ***C*-index de Harrell** (Alabdallah et al., 2024), calculado sobre el conjunto de prueba. No se utiliza la exactitud como métrica principal, porque no representa adecuadamente el tiempo hasta el evento ni la censura.

Por su parte, las métricas secundarias se agrupan por función en la Tabla 3.5: discriminación (*C*-index con ponderación IPCW (Hartman et al., 2023) y AUC dependiente del tiempo a 12 y 36 meses (Beyene & Chen, 2024)), calibración (*Brier Score* a 12 y 36 meses (Zhu et al., 2025), IBS en el intervalo evaluado, pendiente e intersección de calibración) y costo computacional.

Dos principios rigen el cálculo: primero, la **discriminación y la calibración se reportan por separado**, porque un modelo puede ordenar bien a los pacientes y producir al mismo tiempo probabilidades mal calibradas; segundo, **todos los modelos se comparan sobre los registros equivalentes** para cada métrica, con ponderación por censura cuando la métrica lo requiere, de modo que las diferencias no sean artefactos de muestras distintas.

3.14 Comparación e incertidumbre estadística

Las comparaciones son **emparejadas**: cada configuración optimizada se compara con su propia configuración de referencia sobre las mismas predicciones del conjunto de prueba, y las configuraciones del ensemble se comparan entre sí sobre esos mismos registros.

- **Intervalos de confianza:** para las diferencias de *C*-index, AUC dependiente del tiempo y *Brier Score* se emplea **bootstrap por paciente** ($B = 1000$ réplicas), con intervalos percentiles del 95 %, respetando el carácter emparejado de las predicciones (Zhu et al.,

Tabla 3.5
Métricas de evaluación pre-definidas por función

Métrica	Función	Tratamiento de la censura y uso
C-index de Harrell	Discriminación	**Métrica principal de contraste** en el análisis de supervivencia; se calcula sobre el conjunto de prueba con las predicciones de todos los modelos comparados.
C-index con ponderación IPCW (Uno)	Discriminación	Análisis de sensibilidad de la métrica principal, orientado a reducir la dependencia de la función de censura (Hartman et al., 2023).
AUC dependiente del tiempo, $t = 12$ y 36 meses	Discriminación	Parejas acumulativas/dinámicas con ponderación por censura (Beyene & Chen, 2024); evalúa el ordenamiento de riesgos en horizontes clínicamente interpretables.
Brier Score a $t = 12$ y 36 meses	Calibración	Error cuadrático medio de la probabilidad predicha con ponderación por la función de censura (Zhu et al., 2025); métrica de calibración.
Integrated Brier Score en el intervalo evaluado	Calibración	Resumen integrado de la calibración en todo el intervalo evaluado; se reporta siempre junto con el intervalo de tiempo de integración.
Pendiente e intersección de calibración a 36 meses	Calibración	Contraste formal entre riesgo predicho y riesgo observado en el horizonte de tres años.
Tiempo total de búsqueda, número de configuraciones y tiempo de inferencia por caso	Costo computacional	Se registran para cada configuración y estrategia de búsqueda, con el mismo hardware y el mismo límite de tiempo.

Nota. La exactitud (*accuracy*) no se utiliza como métrica principal porque no representa adecuadamente el tiempo hasta el evento ni la censura. La discriminación y la calibración se reportan por separado. Todas las métricas se calculan sobre el mismo conjunto de registros de prueba para cada comparación.

2025).

- **Contraste de HE1:** se estiman dos diferencias emparejadas de C-index, $\Delta C_{\text{ingeniería}} = C_{\text{R1}} - C_{\text{R0}}$ y $\Delta C_{\text{selección}} = C_{\text{R2}} - C_{\text{R1}}$, cada una con su intervalo bootstrap. La hipótesis se considera apoyada si ambas diferencias son positivas y sus intervalos excluyen el cero; si solo una lo hace, se reporta apoyo parcial y se discute. Los dos componentes se reportan por separado para atribuir el efecto.

- **Contraste de HE2:** se estima $\Delta C = C_{\text{optimizado}} - C_{\text{referencia}}$ con su intervalo bootstrap; se considera apoyada si el intervalo excluye el cero a favor de la configuración optimizada, con corrección de Holm por el número de comparaciones (Barnett et al., 2022).

- **Contraste de HE3:** el análisis SHAP se evalúa mediante la estabilidad del ranking global de importancia entre pliegues o semillas, la identificación de las variables con mayor contribución y la coherencia de los patrones de dependencia con el conocimiento clínico. No se formula una hipótesis de mejora de rendimiento atribuida a SHAP; su función es interpretativa.

- **Métrica de sensibilidad:** el C-index con ponderación IPCW (Hartman et al., 2023) se repite como contraste de robustez de la métrica principal frente a la dependencia de la función de censura.

Se reportan siempre tamaños de efecto e intervalos de confianza, no solo valores p. Si un intervalo contiene el cero, la hipótesis respectiva no se rechaza y se declara así explícitamente: **no se presupone que la optimización mejore el rendimiento**, y ningún resultado de este tipo se afirma en este documento antes de ejecutar el procedimiento.

3.15 Interpretabilidad mediante SHAP

La interpretabilidad se evalúa con valores SHAP (Lundberg & Lee, 2017; Salih et al., 2024), aplicados al conjunto de prueba o a una muestra aleatoria del mismo, con tamaño y semilla documentados cuando el costo de cálculo lo exija. Se reportan: el ranking global de importancia, los gráficos de dependencia de las variables principales, la distribución de

contribuciones por observación y el contraste descriptivo de atribución entre la configuración de referencia y la optimizada.

Se presta atención específica a tres cuestiones metodológicas. Primera, la **consistencia temporal**: si alguna variable posterior al diagnóstico alcanza una atribución alta, indica un fallo del procedimiento de exclusión y se investiga su origen. Segunda, la **correlación entre predictores**: las variables clínicas correlacionadas reparten la atribución de forma inestable, por lo que los rankings se leen como descriptivos del comportamiento del modelo y no como estimaciones de efecto causal.

Tercera, se verifica la **estabilidad** comparando el ranking de importancia entre pliegues o semillas para evaluar si las variables principales se mantienen. Asimismo, en el modelo de Cox se reportan los coeficientes penalizados, los cuales ofrecen una lectura aditiva directa.

Si se contempla una selección de características basada en SHAP, el procedimiento se define explícitamente y se realiza únicamente con datos de entrenamiento o validación, sin utilizar el conjunto de prueba para seleccionar variables. Cualquier mejora del modelo atribuida a esta selección deberá demostrarse mediante la comparación experimental de la sección 3.14.

3.16 Herramientas y reproducibilidad

Se declara el conjunto de **herramientas previstas**; las versiones exactas se fijan y se registran en el manifiesto del experimento al momento de ejecutarse, y no se anticipan aquí.

- **Lenguaje y bibliotecas:** Python con `pandas` y `numpy` para manipulación; `scikit-learn` para particiones y transformadores; *Optuna* para la búsqueda (Optuna Developers, 2025); *XGBoost* y *LightGBM* para los modelos de árboles (Boldini et al., 2023; Wiens et al., 2025); una biblioteca de análisis de supervivencia en Python, como `lifelines` o `scikit-survival`, para Cox, *C*-index y *Brier Score*; SHAP para la atribución (Salih et al., 2024).

- **Entorno:** definición del entorno mediante archivo de bloqueo de dependencias (*lockfile*) o contenedor, para que la recreación del entorno no dependa de la máquina.

- **Registro de experimentos:** semilla maestra y lista explícita de semillas; hash criptográfico del archivo de datos extraídos; particiones persistidas como archivos; parámetros, métricas y tiempos de cada ensayo en un registro tabular; **commit** del repositorio de código correspondiente a cada ejecución; especificaciones de hardware y tiempo total de cómputo.

- **Reproducibilidad verificable:** un único punto de entrada reconstruye la cohorte, entrena las configuraciones de referencia, ejecuta la búsqueda, congela los modelos y emite las tablas de resultados; otro investigador con el mismo archivo de datos debe poder regenerar las mismas tablas con variaciones acotadas al azar controlado por semillas.

- **Reporte:** el cumplimiento de las directrices TRIPOD+AI (Collins et al., 2024) se verifica con la lista de verificación oficial y se incluye como anexo de la tesis.

3.17 Consideraciones éticas

El estudio es un **análisis secundario** de datos oncológicos ya registrados, sin contacto con los pacientes y sin modificación de su atención. La base de datos SEER es administrada por el NCI y su uso se sujeta a las condiciones de acceso y a la documentación oficial del organismo titular (National Cancer Institute, 2025).

Se observan las siguientes previsiones:

- **Confidencialidad y minimización:** el extracto de trabajo se limita estrictamente a las variables necesarias para el análisis y no incluye campos identificativos personales. No se realizan intentos de reidentificación ni se combinan los datos con fuentes externas.

- **Marco normativo:** los datos proceden de una fuente poblacional administrada por el NCI y se analizan conforme a sus condiciones de uso; en el país de la investigación se respetan las disposiciones de la Ley N.º 29733 de protección de datos personales (Autoridad Nacional de Protección de Datos Personales, 2025).

- **Autorizaciones:** antes de iniciar el análisis se verificarán las condiciones de acceso a

SEER y, si corresponde, el pronunciamiento del comité de ética de la institución de adscripción. Este documento no afirma que esas aprobaciones existan todavía.

- **Uso de los resultados:** los modelos aquí descritos tienen fines de investigación metodológica; no constituyen un dispositivo clínico ni una recomendación terapéutica, y su despliegue exigiría una validación externa y una evaluación de impacto que este estudio no contempla.

3.18 Correspondencia entre objetivos, procedimientos e hipótesis

La Tabla 3.6 resume cómo cada uno de los tres objetivos específicos del Capítulo I se implementa en este diseño y qué hipótesis contrasta. OE1 se vincula con HE1, OE2 con HE2 y OE3 con HE3; las tres hipótesis encuentran en la sección 3.14 su definición operativa de contraste.

Tabla 3.6
Correspondencia entre objetivos, procedimientos del Capítulo III e hipótesis

Objetivo	Procedimiento que lo implementa	Producto verificable	Hipótesis
OE1	Construcción documentada de la cohorte (S_0–S_2), definición del desenlace, auditoría de fuga, preprocesamiento ajustado por partición e ingeniería de características con ablación R0–R2. Secciones 3.3 a 3.8.2.	Diagrama de flujo con conteos, tabla de variables admisibles/excluidas y tabla de ablación R0–R2.	HE1
OE2	Definición de la configuración de referencia y búsqueda sobre espacios pre-registrados, exclusivamente con entrenamiento y validación, con selección y congelamiento de la configuración ganadora. Secciones 3.10, 3.11 y 3.12.	Registro de ensayos, configuración seleccionada y costo de búsqueda.	HE2
OE3	Análisis de valores SHAP sobre el modelo optimizado, con importancia global, contribuciones locales, dependencia y estabilidad entre pliegues o semillas. Sección 3.15.	Ranking de importancia, gráficos de dependencia y reporte de estabilidad de las contribuciones.	HE3

Nota. La correspondencia se declara antes de ejecutar los experimentos; ningún resultado de rendimiento se anticipa en este capítulo.

Todo lo descrito en este capítulo constituye un **protocolo**: su validez se pone a prueba

al ejecutarse. Las cuestiones que dependen de la obtención de los microdatos —tamaño real de la cohorte, proporción de censura, número de eventos, cobertura efectiva de las variables y fecha de corte definitiva— se resuelven en ese momento y se reportan antes de interpretar cualquier métrica de rendimiento.

Capítulo IV

Experimentación

Referencias bibliográficas

Alabdallah, A., Ohlsson, M., Pashami, S., & Rögnvaldsson, T. (2024). The concordance index decomposition: A measure for a deeper understanding of survival prediction models. *Artificial Intelligence in Medicine*, *148*, 102781. https://doi.org/10.1016/j.artmed.2024.102781

Astley, J. R., Reilly, J. M., Robinson, S., Wild, J. M., Hatton, M. Q., & Tahir, B. A. (2024). Explainable deep learning-based survival prediction for non-small cell lung cancer patients undergoing radical radiotherapy. *Radiotherapy and Oncology*, *193*, 110084. https://doi.org/10.1016/j.radonc.2024.110084

Autoridad Nacional de Protección de Datos Personales. (2025). Protección de datos personales en el Perú. https://www.gob.pe/anpd

Barnett, M. J., Doroudgar, S., Khosraviani, V., & Ip, E. J. (2022). Multiple comparisons: To compare or not to compare, that is the question. *Research in Social and Administrative Pharmacy*, *18*(2), 2331-2334. https://doi.org/10.1016/j.sapharm.2021.07.006

Barnwal, A., Cho, H., & Hocking, T. (2022). Survival regression with accelerated failure time model in XGBoost. *Journal of Computational and Graphical Statistics*, *31*(4), 1292-1302. https://doi.org/10.1080/10618600.2022.2067548

Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A.-L., & Van Calster, B. (2024). Understanding overfitting in random forest for probability estimation: A visualization and simulation study. *Diagnostic and Prognostic Research*, *8*(1), 14. https://doi.org/10.1186/s41512-024-00177-1

Beyene, K. M., & Chen, D.-G. (2024). Time-dependent receiver operating characteristic curve estimator for correlated right-censored time-to-event data. *Statistical Methods in Medical Research*, *33*(1), 162-181. https://doi.org/10.1177/09622802231220496

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs*

Data Mining and Knowledge Discovery, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, *15*(1), 73. https://doi.org/10.1186/s13321-023-00743-7

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229-263. https://doi.org/10.3322/caac.21834

Büyükkeçeci, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. *Gazi University Journal of Science*, *36*(4), 1506-1520. https://doi.org/10.35378/gujs.993763

Cardoso, L. B., Egydio, J. E., Toporcov, T. N., Utida, N. Y., Curado, M. P., Fernandes, G. A., Ribeiro, A. G., Chin, B. G., & Parro, V. C. (2026). Cross-cancer survival prediction using machine learning models. *Scientific Reports*, *16*(1), 9623. https://doi.org/10.1038/s41598-025-34133-w

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Dai, B., & Breheny, P. (2024). Cross-validation approaches for penalized Cox regression. *Statistical Methods in Medical Research*, *33*(4), 702-715. https://doi.org/10.1177/09622802241233770

Denfeld, Q. E., Burger, D., & Lee, C. S. (2023). Survival analysis 101: An easy start guide to analysing time-to-event data. *European Journal of Cardiovascular Nursing*, *22*(3), 332-337. https://doi.org/10.1093/eurjcn/zvad023

Didier, A. J., Nigro, A., Noori, Z., Omballi, M. A., Pappada, S. M., & Hamouda, D. M. (2024). Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis. *Frontiers in Artificial Intelligence*, *7*, 1365777. https://doi.org/10.3389/frai.2024.1365777

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ*, *386*, e078276. https://doi.org/10.1136/bmj-2023-078276

Gago, A., Aguirre, J. M., & Wong, L. (2024). Machine learning system for the effective diagnosis and survival prediction of breast cancer patients. *International Journal of Online and Biomedical Engineering (iJOE)*, *20*(2), 95-113. https://doi.org/10.3991/ijoe.v20i02.42883

Gálvez-Nino, M., Ruiz, R., Roque, K., Coanqui, O., Valdivieso, N., Olivera, M., Ganti, A. K., & Mas, L. (2023). Real-world outcomes of anti-EGFR therapy in advanced non-small cell lung cancer EGFR mutated in Peru. *Thoracic Cancer*, *14*(1), 61-67. https://doi.org/10.1111/1759-7714.14714

Gerencia Regional de Salud Cusco. (2026). Boletín epidemiológico regional 2026 [Notificaciones de tumores malignos de tráquea, bronquios y pulmón, periodo 2021–2025].

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems 35*, 507-520. https://doi.org/10.52202/068431-0037

Hartman, N., Kim, S., He, K., & Kalbfleisch, J. D. (2023). Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, *42*(13), 2179-2190. https://doi.org/10.1002/sim.9717

Hospital Regional del Cusco. (2025). Reporte de vigilancia de cáncer, enero–agosto de 2025.

International Agency for Research on Cancer. (2026). Peru fact sheet: GLOBOCAN 2024 [Estimaciones de incidencia, mortalidad y prevalencia para el Perú]. https://gco.iarc.who.int/media/globocan/factsheets/populations/604-peru-fact-sheet.pdf

Jin, L., Zhao, Q., Fu, S., Cao, F., Hou, B., & Ma, J. (2023). Development and validation of machine learning models to predict survival of patients with resected stage-III NSCLC. *Frontiers in Oncology*, *13*, 1092478. https://doi.org/10.3389/fonc.2023.1092478

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, 4765-4774. https://arxiv.org/abs/1705.07874

Mamani Barrios, E. (2023). *Algoritmos de machine learning para predecir la anemia en niños de 6 a 35 meses de edad en Cusco* [Tesis de Maestría en Estadística]. Universidad Nacional de San Antonio Abad del Cusco. https://hdl.handle.net/20.500.12918/8469

Microsoft Corporation. (2025). LightGBM documentation. https://lightgbm.readthedocs.io/en/latest/

Monterrubio-Gómez, K., Constantine-Cooke, N., & Vallejos, C. A. (2024). A review on statistical and machine learning competing risks methods. *Biometrical Journal*, *66*(2), 2300060. https://doi.org/10.1002/bimj.202300060

National Cancer Institute. (2025). Surveillance, Epidemiology, and End Results (SEER) Program [Documentación oficial, diccionarios de variables y datos de incidencia y supervivencia]. https://seer.cancer.gov/

Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., Asselbergs, F. W., Moons, K. G. M., & Debray, T. P. A. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, *142*, 218-229. https://doi.org/10.1016/j.jclinepi.2021.11.023

Optuna Developers. (2025). Optuna documentation. https://optuna.readthedocs.io/en/stable/

Quispe Rodríguez. (2022). *Cáncer de pulmón: características clínico epidemiológicas y sociodemográficas en el Hospital Antonio Lorena del Cusco, 2015–2021* [Tesis]. Universidad Nacional de San Antonio Abad del Cusco. http://hdl.handle.net/20.500.12918/6703

Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ*, *384*, e074820. https://doi.org/10.1136/bmj-2023-074820

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. https://doi.org/10.1002/aisy.202400304

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84-90. https://doi.org/10.1016/j.inffus.2021.11.011

Sonabend, R., Bender, A., & Vollmer, S. (2022). Avoiding C-hacking when evaluating survival distribution predictions with discrimination measures. *Bioinformatics*, *38*(17), 4178-4184. https://doi.org/10.1093/bioinformatics/btac451

Takefuji, Y. (2025). Beyond SHAP: Reliable feature selection methods for clinical prediction models. *Archives of Gerontology and Geriatrics*, *135*, 105873. https://doi.org/10.1016/j.archger.2025.105873

Tetko, I. V., van Deursen, R., & Godin, G. (2024). Be aware of overfitting by hyperparameter optimization! *Journal of Cheminformatics*, *16*(1), 139. https://doi.org/10.1186/s13321-024-00934-w

van Geloven, N., Giardiello, D., Bonneville, E. F., Teece, L., Ramspek, C. L., van Smeden, M., Snell, K. I. E., van Calster, B., Pohar-Perme, M., Riley, R. D., Putter, H., & Steyerberg, E. (2022). Validation of prediction models in the presence of competing risks: A guide through modern methods. *BMJ*, *377*, e069249. https://doi.org/10.1136/bmj-2021-069249

van de Mortel, L. A., & van Wingen, G. A. (2025). Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance. *Molecular Psychiatry*, *30*(12), 6070-6071. https://doi.org/10.1038/s41380-025-03336-y

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review*, *57*(3), 65. https://doi.org/10.1007/s10462-023-10681-3

Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A tutorial and use case example of the eXtreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clinical and Translational Science*, *18*(3), e70172. https://doi.org/10.1111/cts.70172

World Health Organization. (2025). Cancer [Hoja informativa sobre el cáncer]. https://www.who.int/news-room/fact-sheets/detail/cancer

Yuan, Y., Zhang, G., Gu, Y., Hao, S., Huang, C., Xie, H., Mi, W., & Zeng, Y. (2025). Artificial intelligence-assisted machine learning models for predicting lung cancer survival. *Asia-Pacific Journal of Oncology Nursing*, *12*, 100680. https://doi.org/10.1016/j.apjon.2025.100680

Zhu, K., Zheng, Y., & Chan, K. C. G. (2025). Weighted Brier score—An overall summary measure for risk prediction models with clinical utility consideration. *Statistics in Biosciences*. https://doi.org/10.1007/s12561-025-09505-5

Anexos

Anexo 1. Matriz de consistencia de la investigación

En la Tabla A.1 se presenta la matriz de consistencia del proyecto de tesis en formato horizontal, con el propósito de visualizar de manera integrada la coherencia lógica, vertical y horizontal entre los componentes fundamentales del estudio. La estructura articula en cinco columnas el planteamiento del problema general y específico, los objetivos de investigación, el sistema de hipótesis, la operacionalización de las dos variables con sus dimensiones y el diseño metodológico adoptado.

Cada problema específico (PE1, PE2 y PE3) guarda correspondencia directa con un objetivo específico (OE1, OE2 y OE3), una hipótesis específica (HE1, HE2 y HE3) y las dimensiones establecidas para la optimización del modelo ensemble y la predicción de supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015–2023.

Tabla A.1

Matriz de consistencia: Optimización de un modelo ensemble mediante ingeniería de características y SHAP para la predicción de supervivencia en pacientes con cáncer de pulmón utilizando la base de datos SEER, 2015-2023

Problema	Objetivos	Hipótesis	Variables y dimensiones	Metodología
Problema general: ¿De qué manera **optimizar el rendimiento predictivo** de un **modelo ensemble mediante ingeniería de características y selección con SHAP** para la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**? **Problemas específicos:** **PE1:** ¿Cómo realizar el **preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** para identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**? **PE2:** ¿De qué manera **desarrollar y optimizar** un **modelo ensemble mediante algoritmos de aprendizaje automático** para mejorar el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**? **PE3:** ¿En qué medida la aplicación de **SHAP** permite **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**?	**Objetivo general:** **Optimizar el rendimiento predictivo** de un **modelo ensemble mediante ingeniería de características y selección con SHAP** para la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**. **Objetivos específicos:** **OE1: Realizar el preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** para identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**. **OE2: Desarrollar y optimizar** un **modelo ensemble mediante algoritmos de aprendizaje automático** para mejorar el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**. **OE3: Aplicar SHAP** para **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**.	**Hipótesis general:** **HG:** La **optimización del rendimiento predictivo** de un **modelo ensemble mediante ingeniería de características y selección con SHAP** permite mejorar la predicción de la **supervivencia en pacientes con cáncer de pulmón registrados en la base de datos SEER durante el periodo 2015-2023**. **Hipótesis específicas:** **HE1:** El **preprocesamiento y la ingeniería de características** de los datos clínicos y demográficos de **SEER** permiten identificar y construir **variables predictivas relevantes** asociadas con la **supervivencia de pacientes con cáncer de pulmón**. **HE2:** El **desarrollo y la optimización** de un **modelo ensemble mediante algoritmos de aprendizaje automático** mejoran el **rendimiento predictivo** de la **supervivencia en pacientes con cáncer de pulmón**. **HE3:** La aplicación de **SHAP** permite **identificar e interpretar las características** que más contribuyen a las **predicciones del modelo ensemble optimizado para la supervivencia de pacientes con cáncer de pulmón**.	**Variable 1 (Independiente):** Optimización de un modelo ensemble mediante ingeniería de características y SHAP. *Dimensiones:* • Preprocesamiento e ingeniería de características clínicas y demográficas (ablación R0, R1 y R2). • Desarrollo y optimización de hiperparámetros del modelo ensemble (Cox penalizado, *Random Survival Forest*, *XGBoost* AFT y búsqueda bayesiana TPE). • Selección e interpretación de características mediante valores SHAP (importancia global, contribución local y estabilidad). **Variable 2 (Dependiente):** Predicción de supervivencia en pacientes con cáncer de pulmón. *Dimensiones:* • Tiempo de supervivencia global con censura a la derecha (T, δ) desde el diagnóstico. • Capacidad discriminativa del pronóstico (*C*-index de Harrell, *C*-index IPCW y AUC dependiente del tiempo a 12 y 36 meses). • Calibración probabilística (*Brier Score* a 12 y 36 meses, *Integrated Brier Score* y pendiente de calibración).	**Tipo de investigación:** Aplicada. **Enfoque de investigación:** Cuantitativo. **Nivel de investigación:** Predictivo (con fase descriptiva). **Diseño de la investigación:** No experimental, retrospectivo, longitudinal y computacional. **Población y fuente:** Registros de pacientes con cáncer de pulmón en la base de datos SEER (2015–2023), codificados según la CIE-O-3. **Muestra y validación:** Cohorte secuencial (S_0 a S_2); partición 60 % entrenamiento, 20 % validación y 20 % prueba independiente; validación cruzada de 5 pliegues. **Técnicas e instrumentos:** Extracción documental SEER, optimización con *Optuna* (50 ensayos), análisis SHAP, intervalos bootstrap ($B = 1000$) y protocolo TRIPOD+AI.

Nota. Elaboración propia en concordancia con los capítulos I, II y III del documento de tesis.