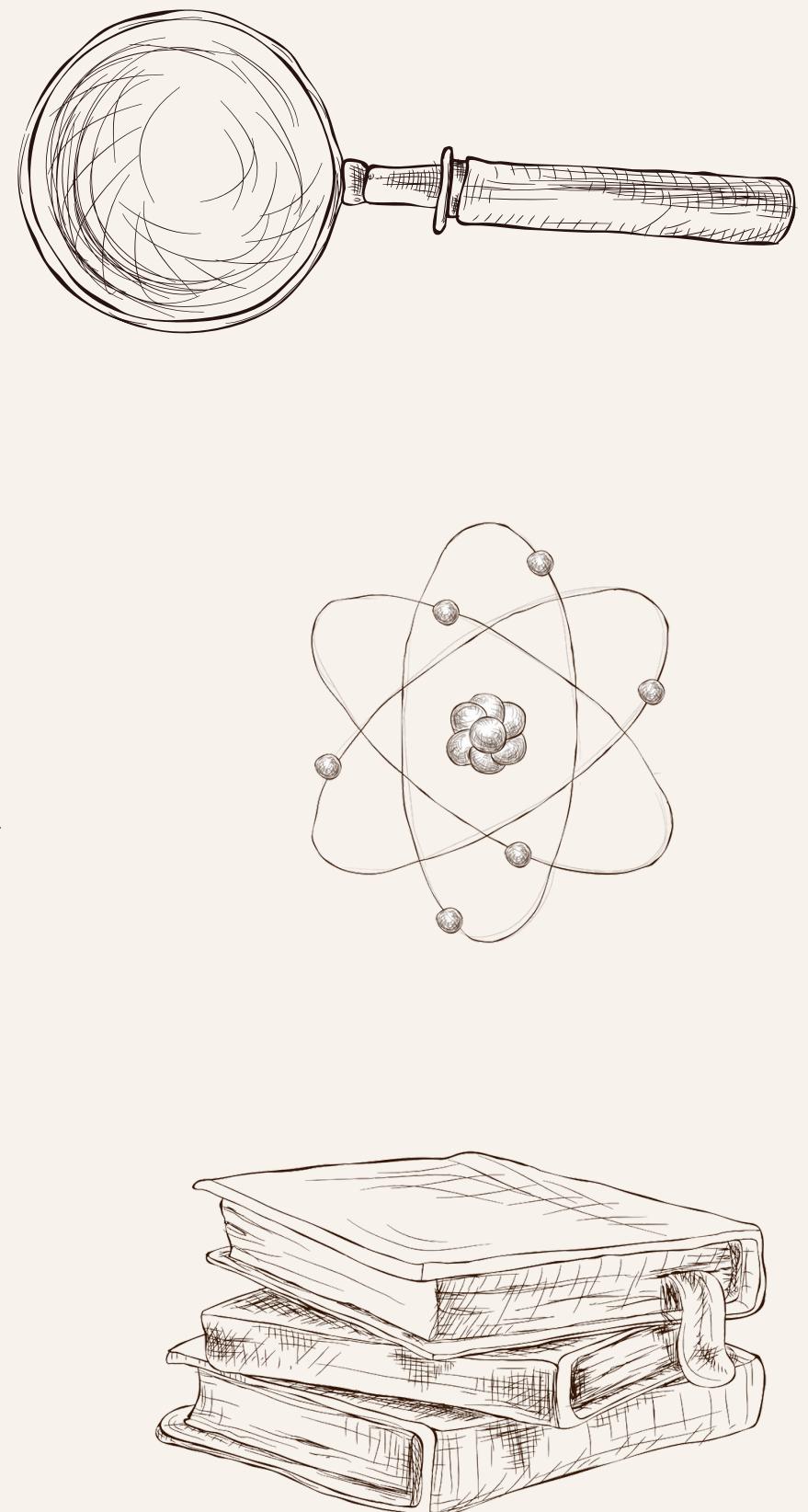


DATA SCIENCE

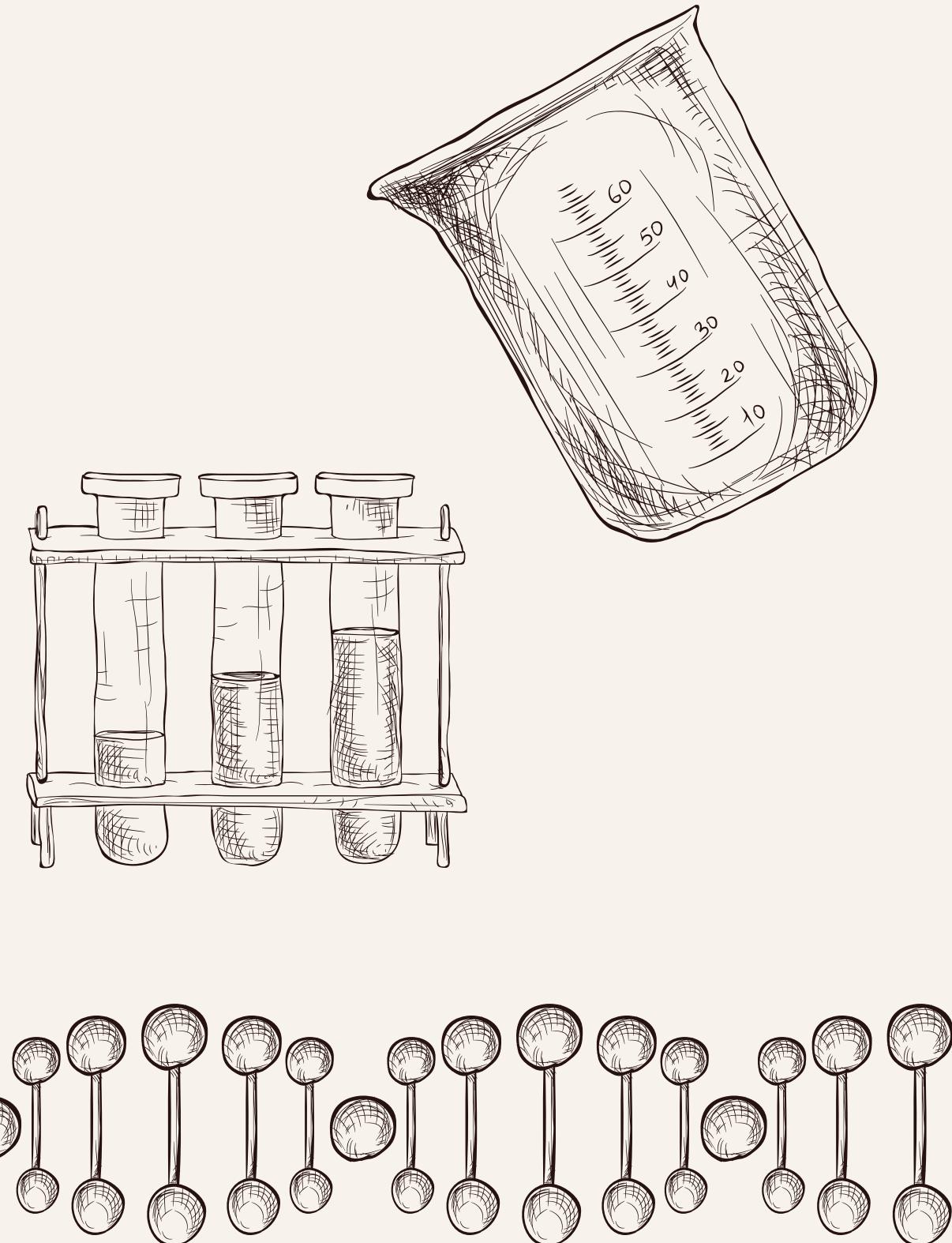
# LUNG CANCER DIAGNOSIS & LIFE EXPECTANCY

Presented By:  
**Betania Medina**  
**Carlos Villa**  
**Elius Trujillo**



# ÍNDICE

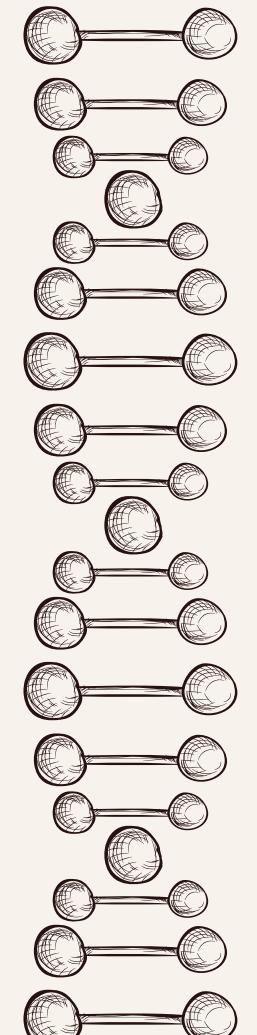
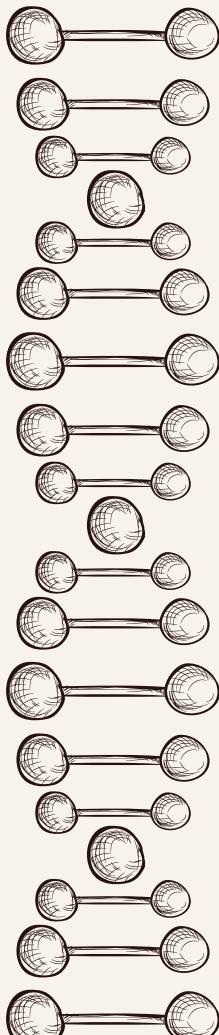
- Introducción
- Recopilación de datos
- Stack tecnológico
- Variables críticas
- Selección de registros
- Matriz de correlación
- El modelo de diagnóstico
- Modelos fallidos
- Procesamiento de datos
- Evaluación del modelo
- Predicción de esperanza de vida
- Streamlit
- Conclusiones



# INTRODUCCIÓN

## **Impacto Epidemiológico**

El cáncer de pulmón es la primera causa de muerte por cáncer a nivel global, lo que exige modelos predictivos de alta precisión.



## **Multifactoriedad del Pronóstico**

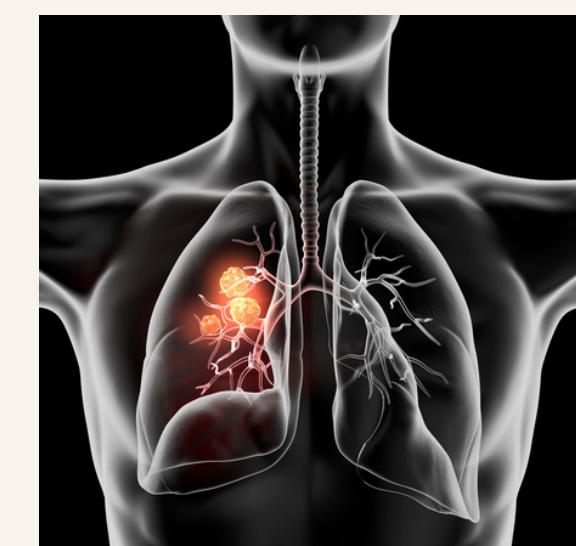
La supervivencia no depende de un solo factor, sino de la interacción entre biología tumoral y respuesta al tratamiento

## **Objetivo del Análisis**

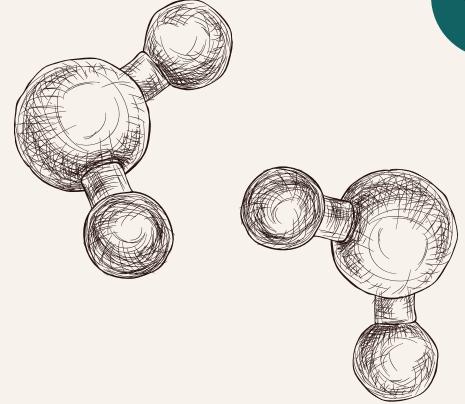
Transformar datos clínicos complejos en patrones comprensibles para apoyar la medicina personalizada

## **Valor de la Predicción**

Pasar de la estadística descriptiva al soporte de decisiones clínicas en tiempo real.



# RECOPILACIÓN DE DATOS



The screenshot shows the SEER\*Stat software interface. At the top, there's a banner with the NIH logo and the text "NATIONAL CANCER INSTITUTE Surveillance, Epidemiology, and End Results Program". Below the banner, the main window has a blue header bar with "SEER\*Stat - sesion consulta.sl" and standard menu options: File, New Session, Actions, View, Help, and a profile dropdown. The "Actions" tab is currently selected. Below the header are four buttons: Execute, Execute Remotely, Dictionary, and Export. The main area is divided into two main sections: "Display Variables" and "Available Variables". The "Display Variables" section contains a "Column" list with various oncological variables like Age recode with <1 year olds and 90+, Race recode (White, Black, Other), Primary Site - labeled, Histologic Type ICD-O-3, Grade Clinical (2018+), Grade Recode (thru 2017), Combined Summary Stage with Expanded Regional Codes (2004+), Derived EOD 2018 Stage Group Recode (2018+), 7th Edition Stage Group Recode (2016-2017), Derived AJCC Stage Group, 7th ed (2010-2015), CS tumor size (2004-2015), Tumor Size Summary (2016+), Survival months, Survival months flag, and Vital status recode (study cutoff used). The "Available Variables" section shows a hierarchical tree of data items under categories such as Age at Diagnosis, Race, Sex, Year Dx, Site and Morphology, Stage, Therapy, Site-Specific Data Items, Extent of Disease, Cause of Death (COD) and Follow-up, and Multiple Primary Fields. At the bottom of the interface, there are buttons for Column, Sort, Both, Copy Variables..., and Find... . A "Ready" message is displayed at the very bottom.

01

Trabajamos con registros entre el año 2012 al 2022.

02

Se seleccionan los registros donde el individuo desarrolló cáncer de pulmón.

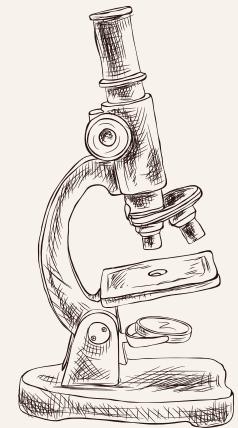
03

Las edades de los registros van desde recién nacidos a personas mayores de 90 años.

04

No disponemos de características que traten los hábitos de los pacientes, las variables son oncológicas, exceptuando la variable de ingresos económicos.

# STACK TECNOLÓGICO



- **Lenguaje de Programación**

- Python
- Librerías clave: Pandas, NumPy y Scikit-Learn



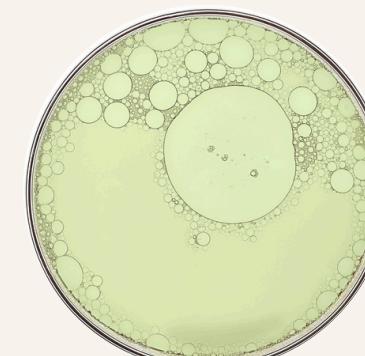
- **Visualización**

Matplotlib, Seaborn y Streamlit



- **Entorno de Desarrollo**

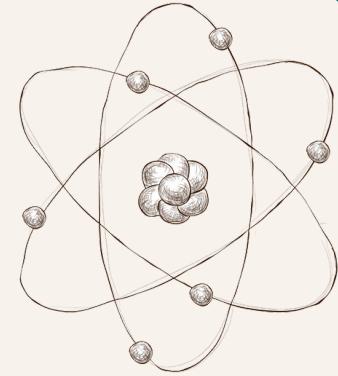
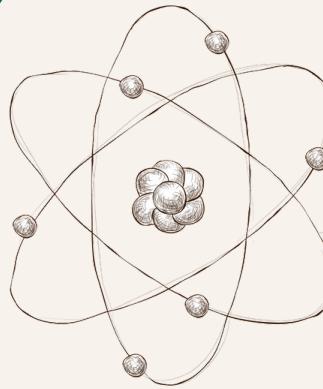
VS Code



- **Control de Versiones**

Git & GitHub

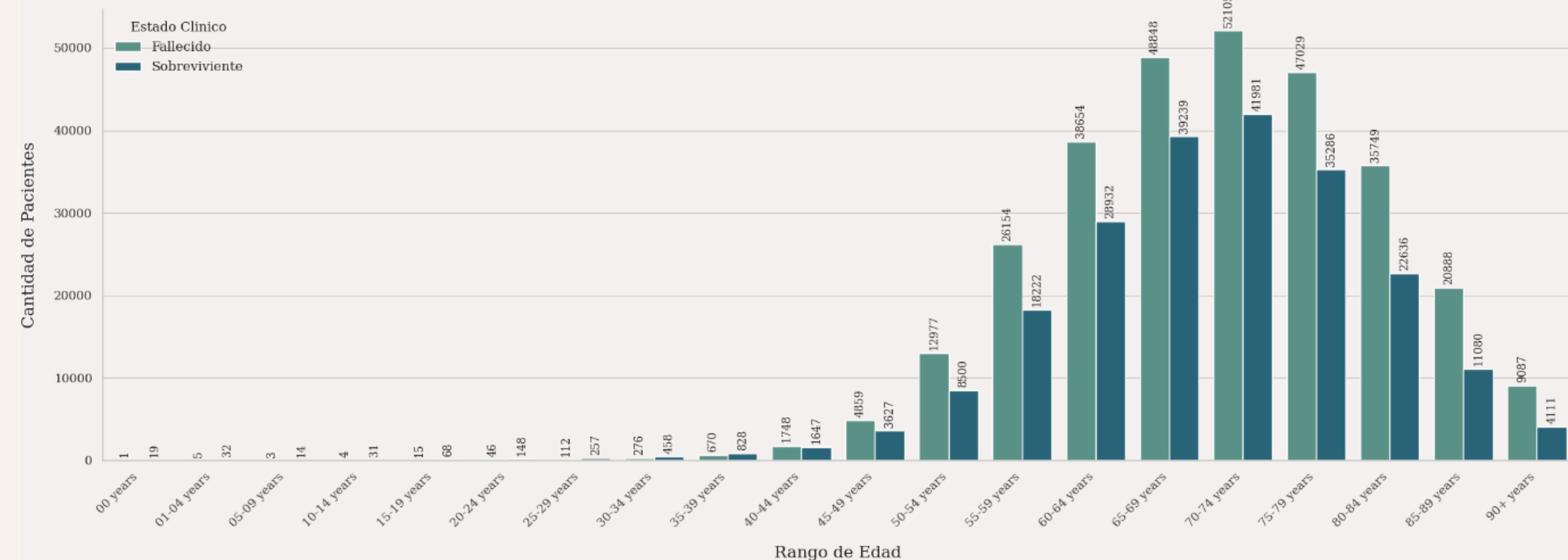
# VARIABLES CRÍTICAS

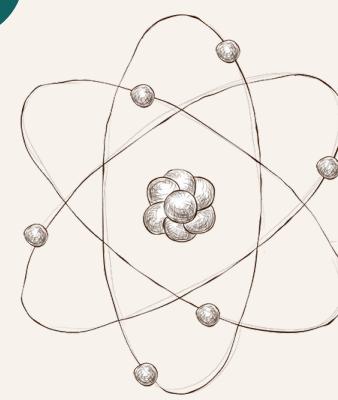


**Age recode with <1 year olds and 90+**

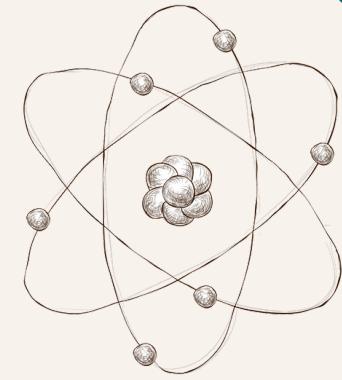
- Correlación Edad-Mortalidad.
- Pico de Incidencia.
- Limitación Terapéutica.
- Calidad de los Datos.

**Distribución de Pacientes por Edad y Supervivencia**



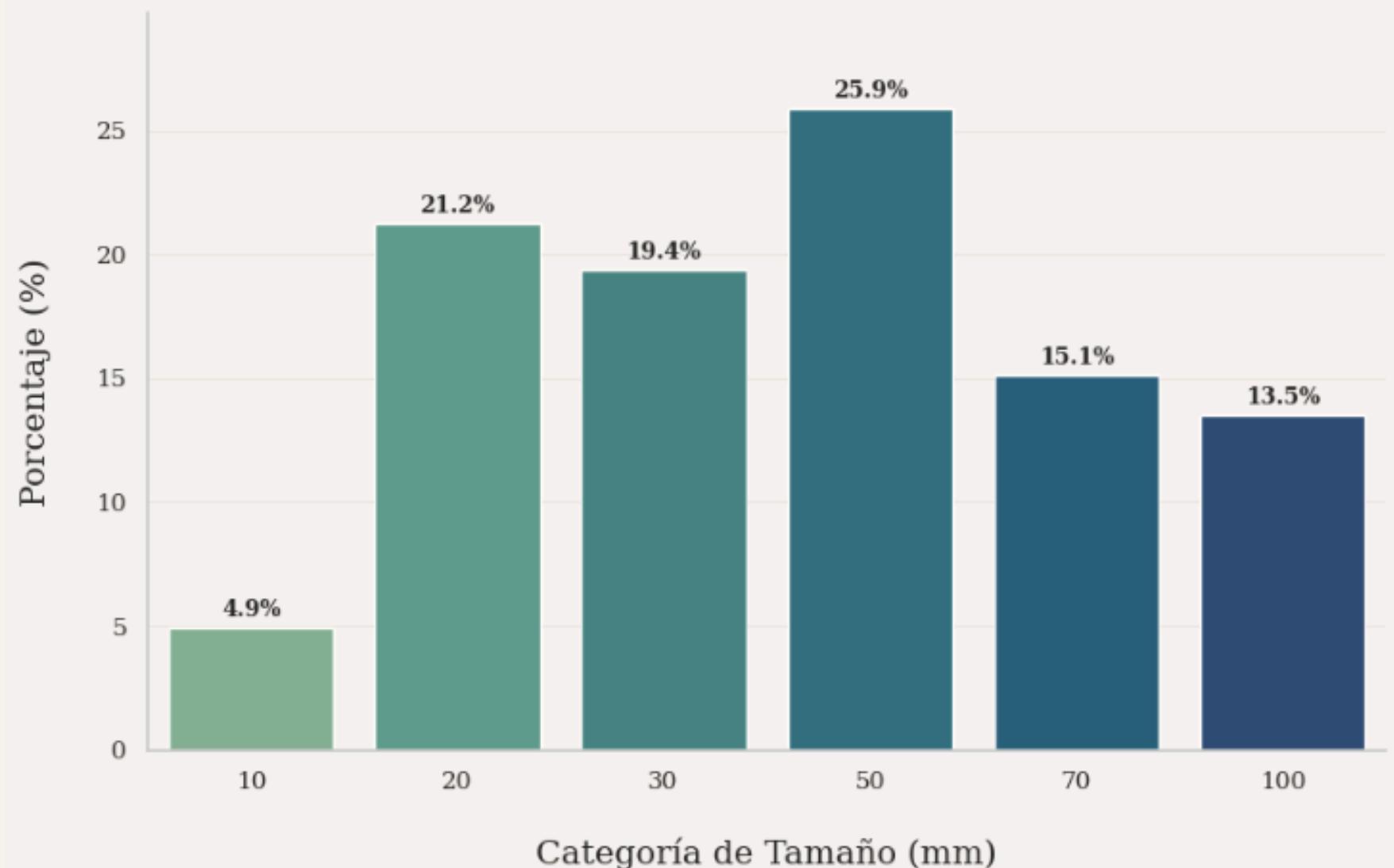


# VARIABLES CRÍTICAS



## CS tumor size (2004-2015) y Tumor Size Summary (2016+)

Distribución Porcentual por Categoría de Tamaño de Tumor



### Factor Pronóstico Primario

El tamaño del tumor es uno de los indicadores fundamentales para determinar el estadio clínico y la agresividad de la enfermedad.

### Unificación de Datos

Combinamos ambas variables en una única columna para clasificar el tamaño del cáncer en mm.

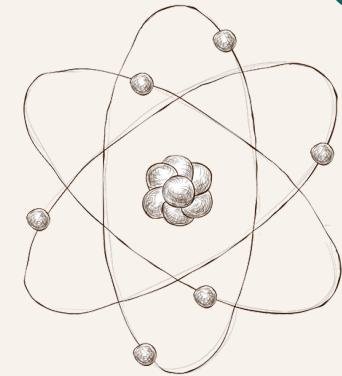
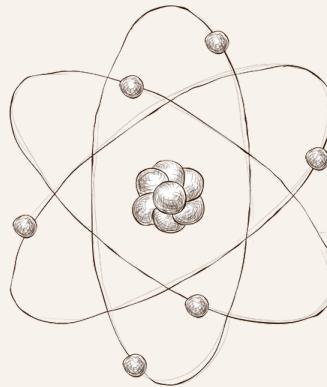
### Distribución de Riesgo

La mayor concentración de casos se encuentra en el rango de 20 a 50mm, zonas donde la precisión del modelo es crítica para decidir entre intervención quirúrgica o tratamientos sistémicos.

### Relación Logística

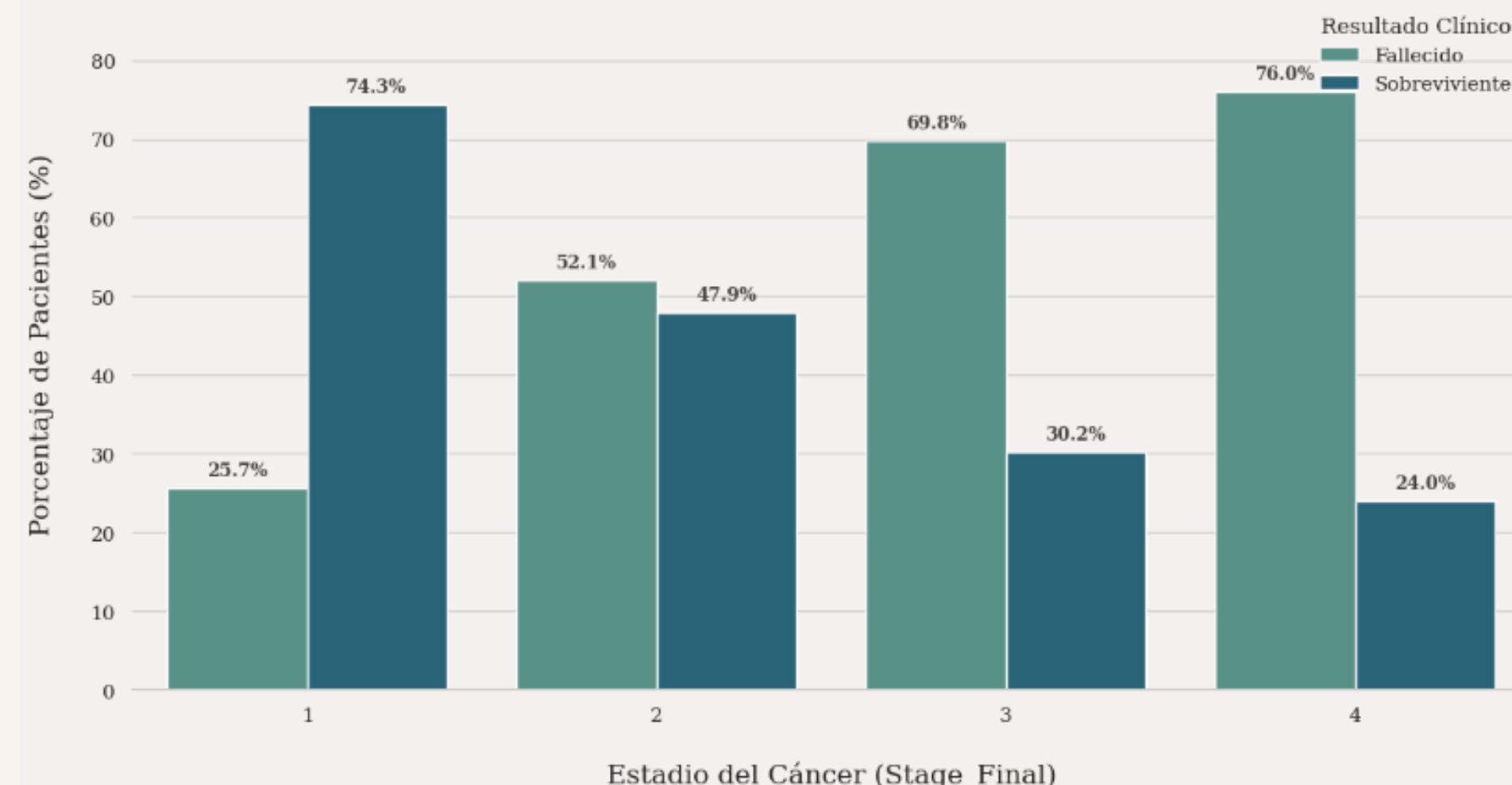
Existe una correlación directa entre el incremento del diámetro tumoral y la reducción en las probabilidades de supervivencia a largo plazo registradas por el algoritmo.

# VARIABLES CRÍTICAS



- Combined Summary Stage with Expanded Regional Codes (2004+)
- Derived AJCC Stage Group, 7th ed (2010-2015),
- 7th Edition Stage Group Recode (2016-2017)
- Derived EOD 2018 Stage Group Recode (2018+)

## Validación de Pronóstico: Estadio Final y Desenlace



## Armonización de Estándares Médicos

Se unificaron cuatro sistemas de codificación distintos en una variable maestra llamada Stage\_Final para permitir un análisis coherente desde 2004 hasta la actualidad.

## Validación de Gravedad Clínica

Los datos confirman la progresión de la enfermedad: mientras que en el Estadio 1 la supervivencia es del 74.3%, en el Estadio 4 (metastásico) la probabilidad de supervivencia en este grado cae drásticamente al 24%.

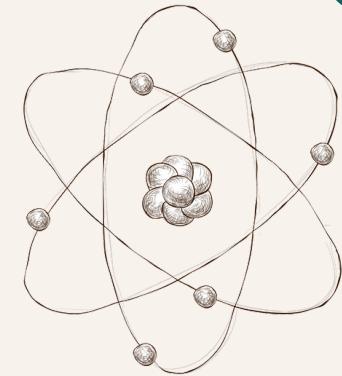
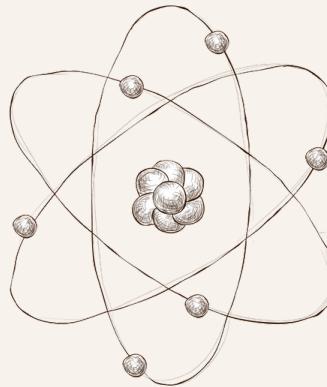
## El Salto del Estadio 2 al 3

Disminuye la tasa de supervivencia, son una frontera vital para el éxito del tratamiento.

## Fundamento del Modelo

La variable Stage\_Final tiene una gran influencia en el modelo, ya que conecta directamente la gravedad clínica con la probabilidad de supervivencia.

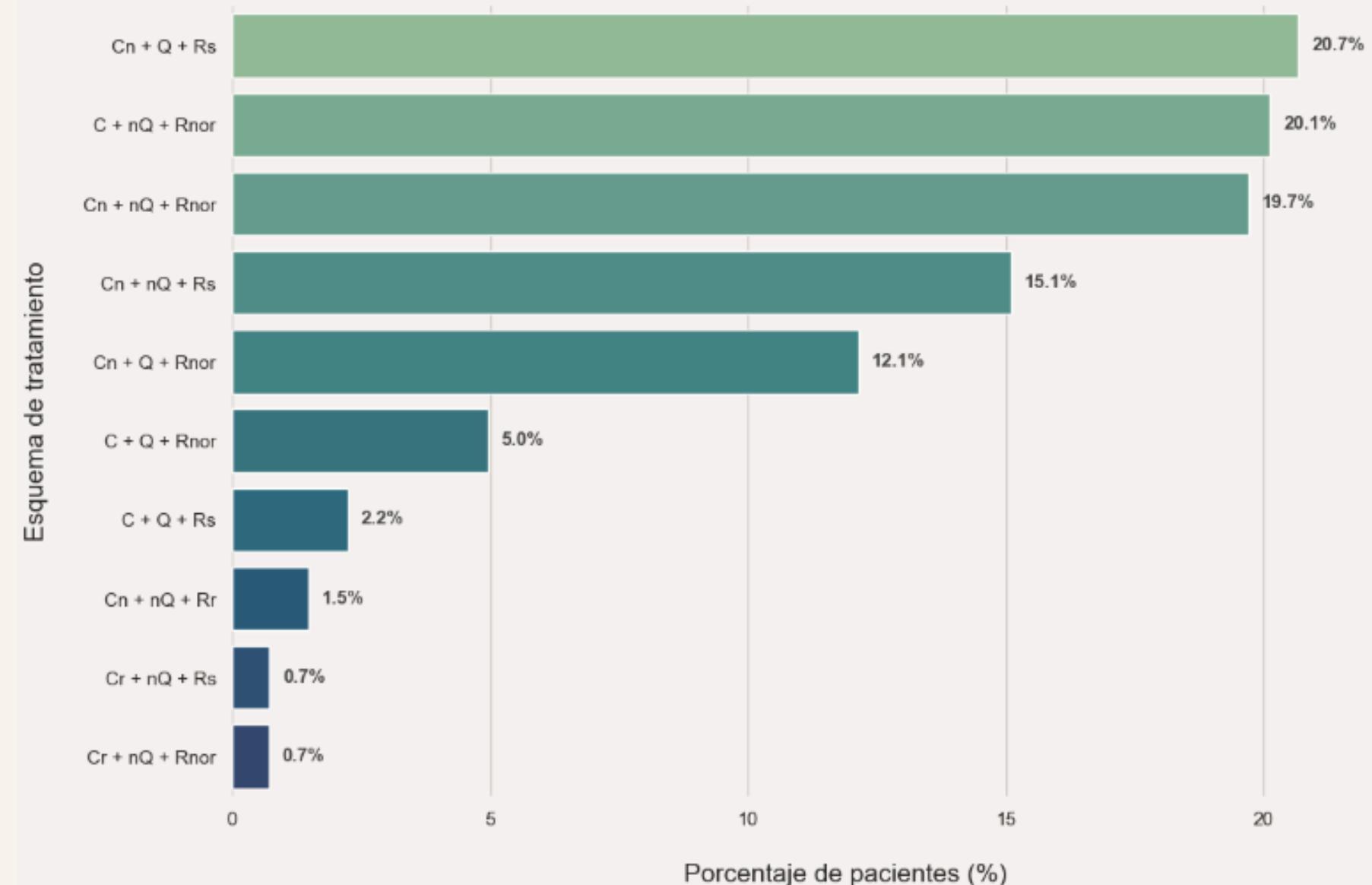
# VARIABLES CRÍTICAS



- Radiation recodeGrade Recode (thru 2017)
- Chemotherapy recode (yes, no/unk)
- Reason no cancer-directed surgery

## TRATAMIENTO

Top 10 Combinaciones de tratamiento más frecuentes

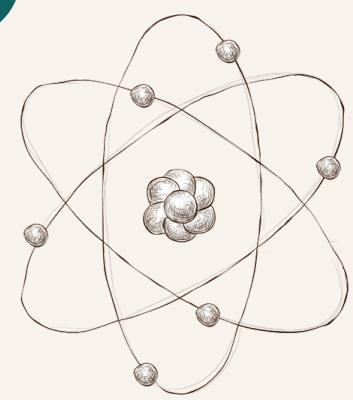


## Ingeniería de Características

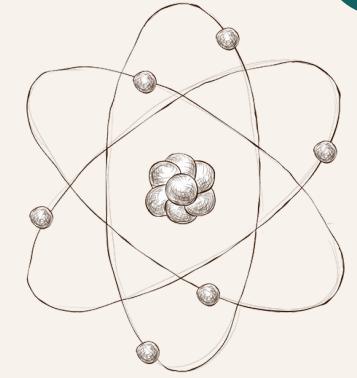
Se unificaron variables de radiación y cirugía en un esquema de combinaciones frecuentes para identificar los protocolos con mayor tasa de éxito.

## Jerarquía de Intervención

La combinación de cirugía y quimioterapia aparece como uno de los esquemas más frecuentes, reflejando los estándares clínicos actuales para estadios controlables.



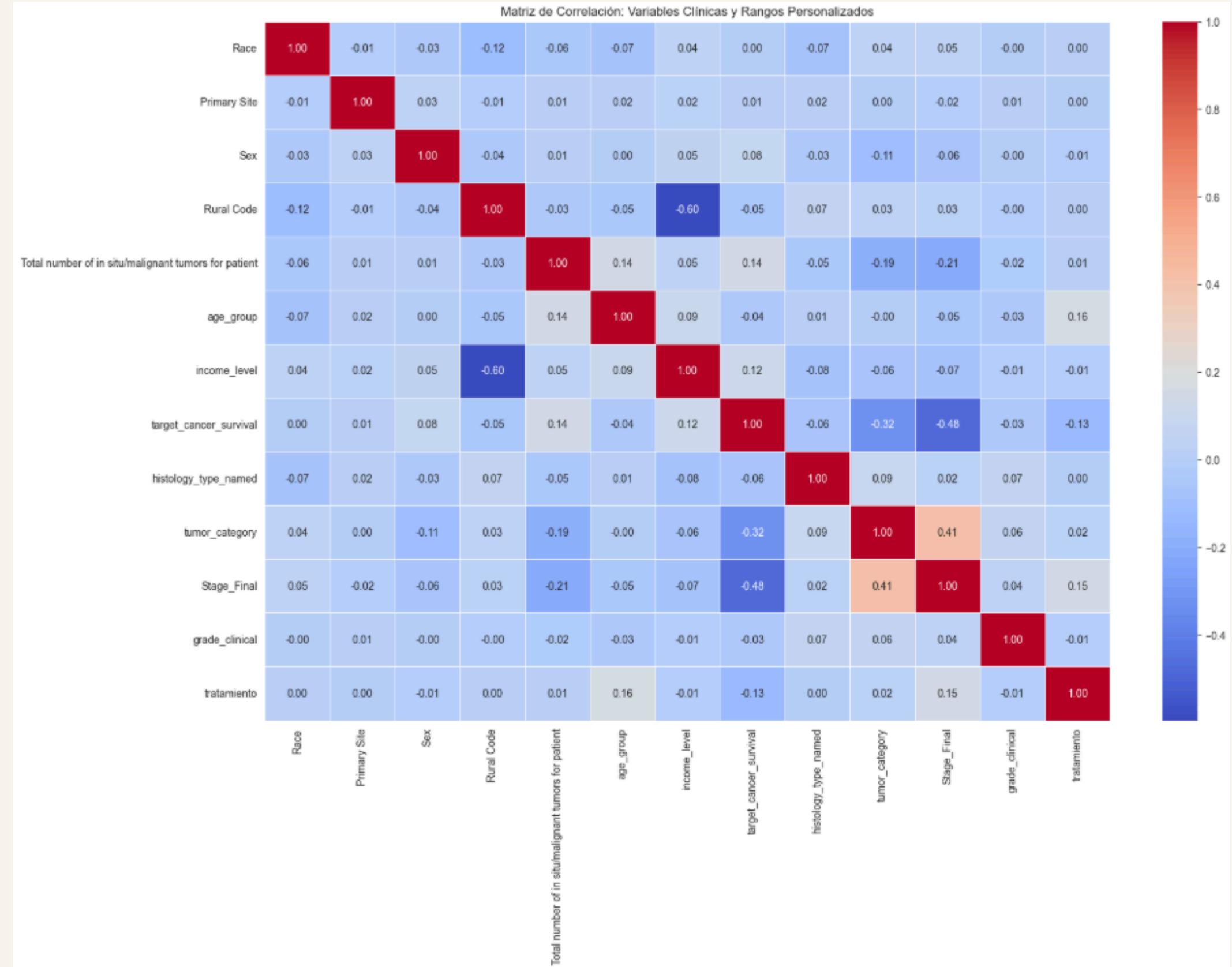
# SELECCIÓN DE REGISTROS



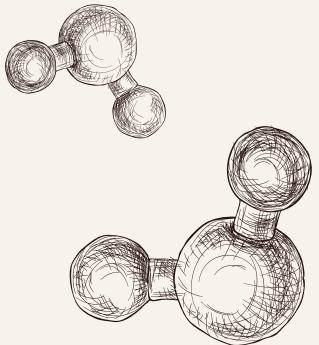
## Registros con características distintas a:

- Diagnóstico del tumor por medio de autopsias.
- Registros provenientes por acta de defunción.
- Tumores de grado desconocido.
- Persona cuyo fallecimiento no se debe al cáncer.
- Datos incompletos debido a que tiene cero días de seguimiento.
- Lugar de crecimiento del tumor desconocido.
- Tipo de célula cancerígena desconocida.

# MATRIZ DE CORRELACIÓN



# EL MODELO DE DIAGNOSTICO



## Variables predictoras

- age\_group
- tumor\_category
- income\_level
- Total number of in situ/malignant tumors  
for patient
- tratamiento
- Stage\_Final
- Sex
- grade\_clinical
- histology\_type\_named

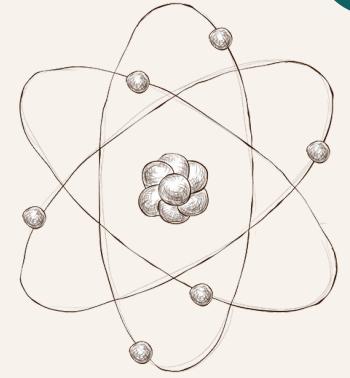
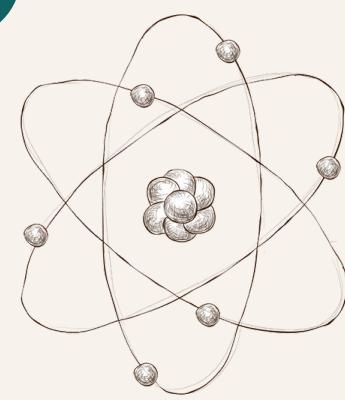
## Variable a predecir

target\_cancer\_survival

## Problema de selección de datos El Sesgo de los Datos "Censurados"

## La Solución

Filtro Dinámico por Hitos (Cortes)



# MODELOS FALLIDOS

## Idea principal

### Predicción de meses de supervivencia

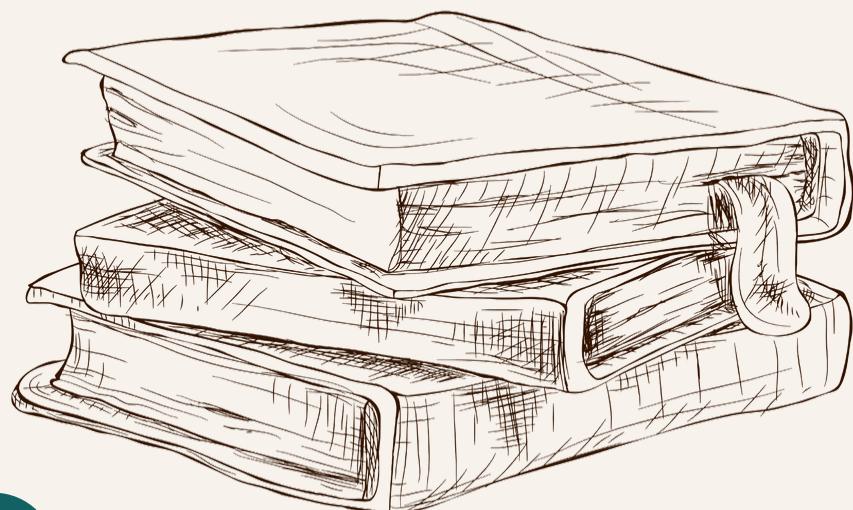
Modelo XGBoost

--- Métricas del Modelo ---

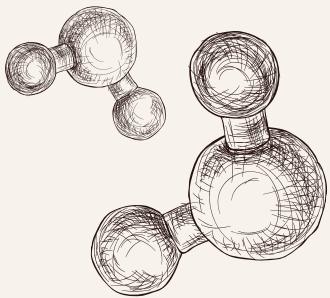
R<sup>2</sup> (Coeficiente de determinación): 0.5422

MAE (Error Absoluto Medio): 9.3885 meses

RMSE (Raíz del Error Cuadrático Medio): 13.8223 meses

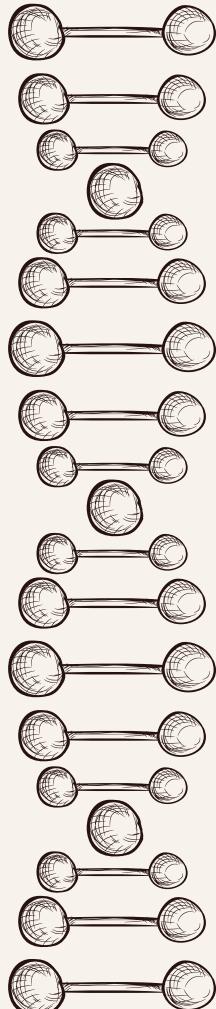


# PROCESAMIENTO DE DATOS



## Cortes temporales

12, 24, 36, 48 y 60 meses.



## Lógica de Selección

### Inclusión

Solo incluimos pacientes que ya han "completado" el tiempo del hito o que fallecieron antes.

### Exclusión

Eliminamos a los pacientes con seguimiento menor al mes del corte que aún están vivos (casos de incertidumbre).

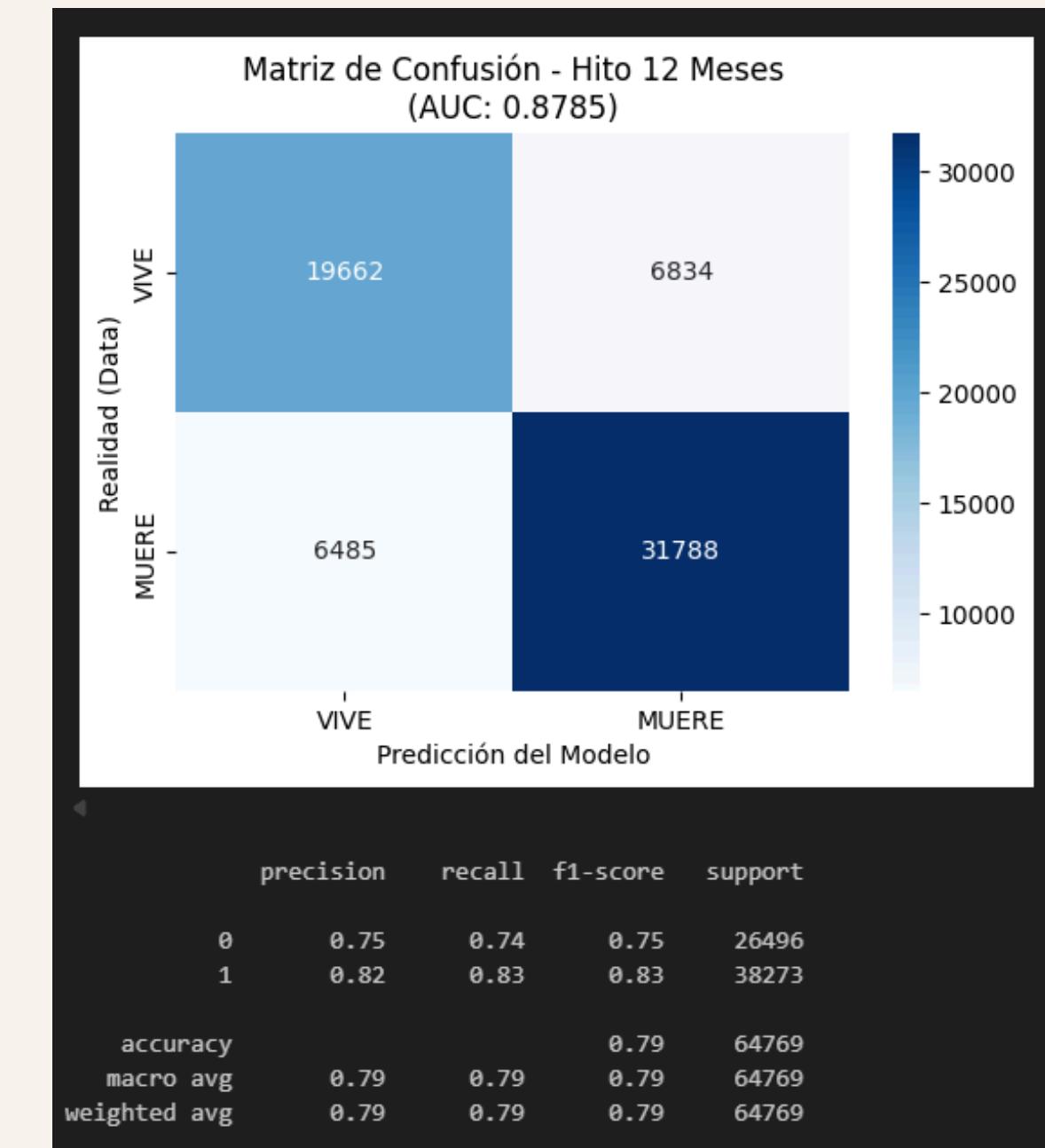
### Clase 0 (Evento)

El paciente falleció antes de cumplirse el mes del corte

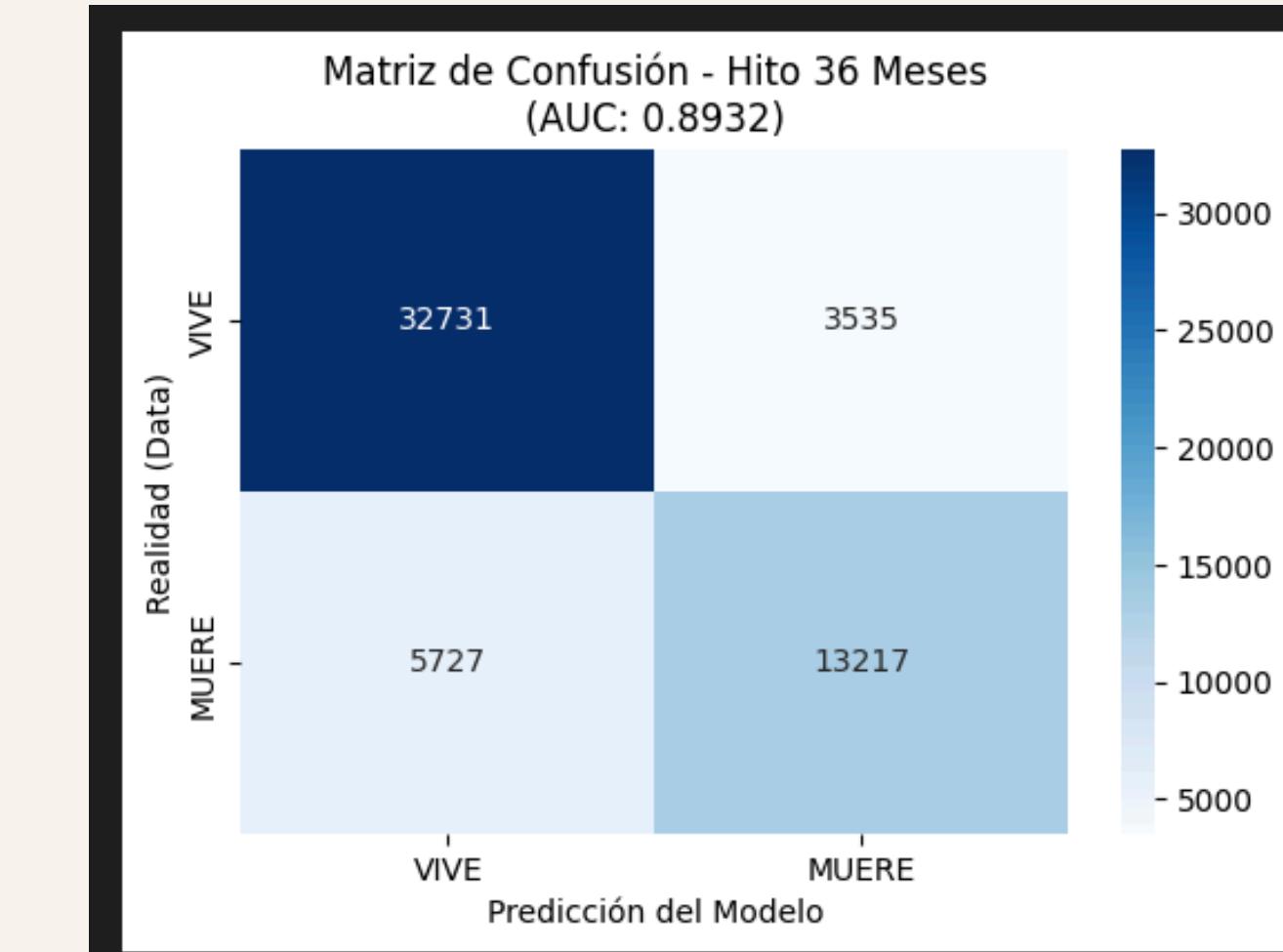
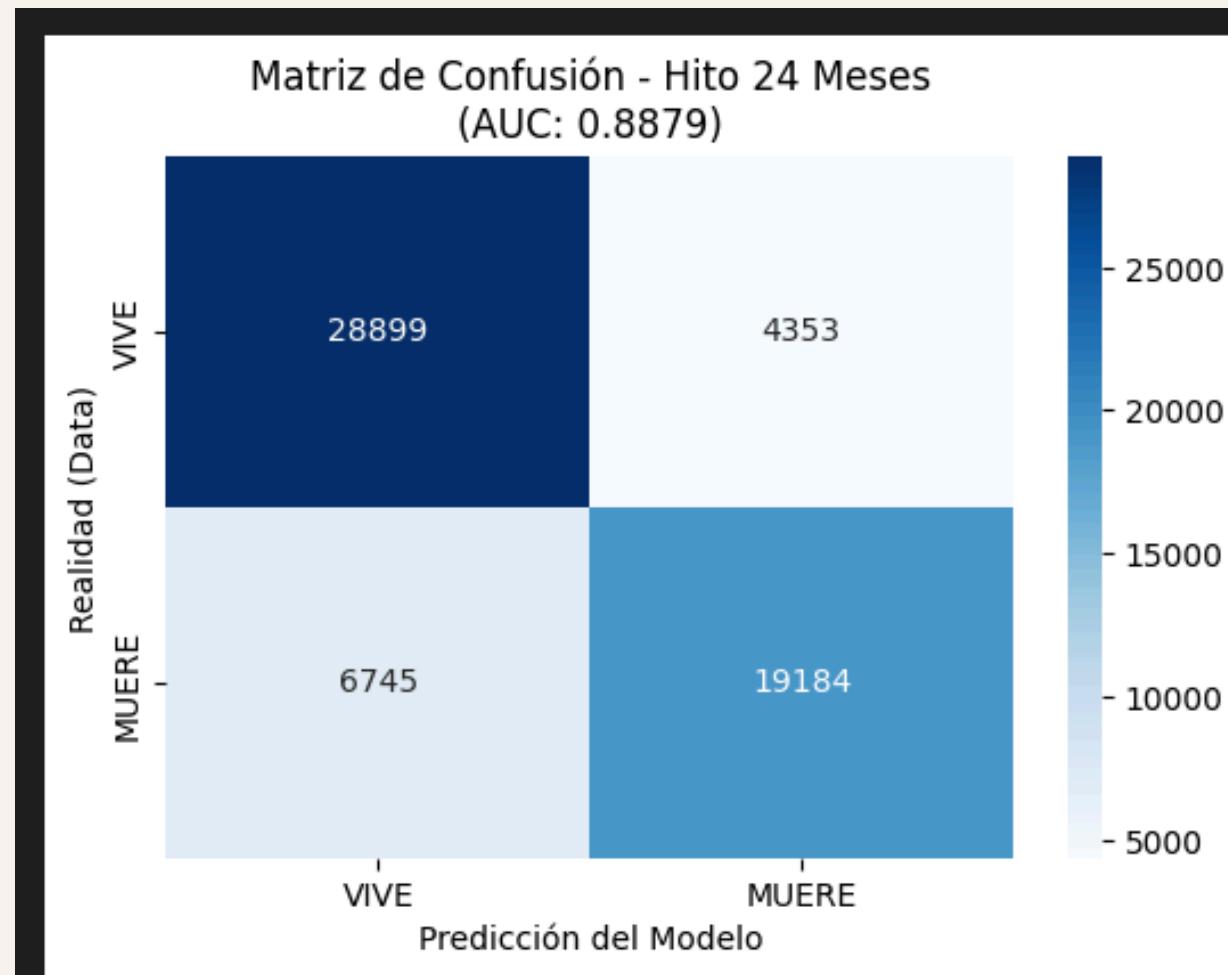
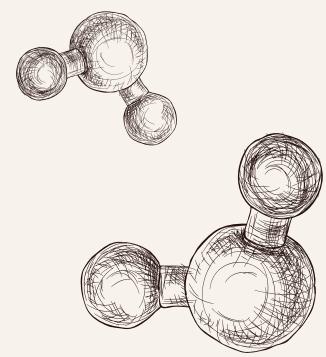
### Clase 1 (Supervivencia)

El paciente superó con vida el mes del corte

## Modelo XGBoost Classifier



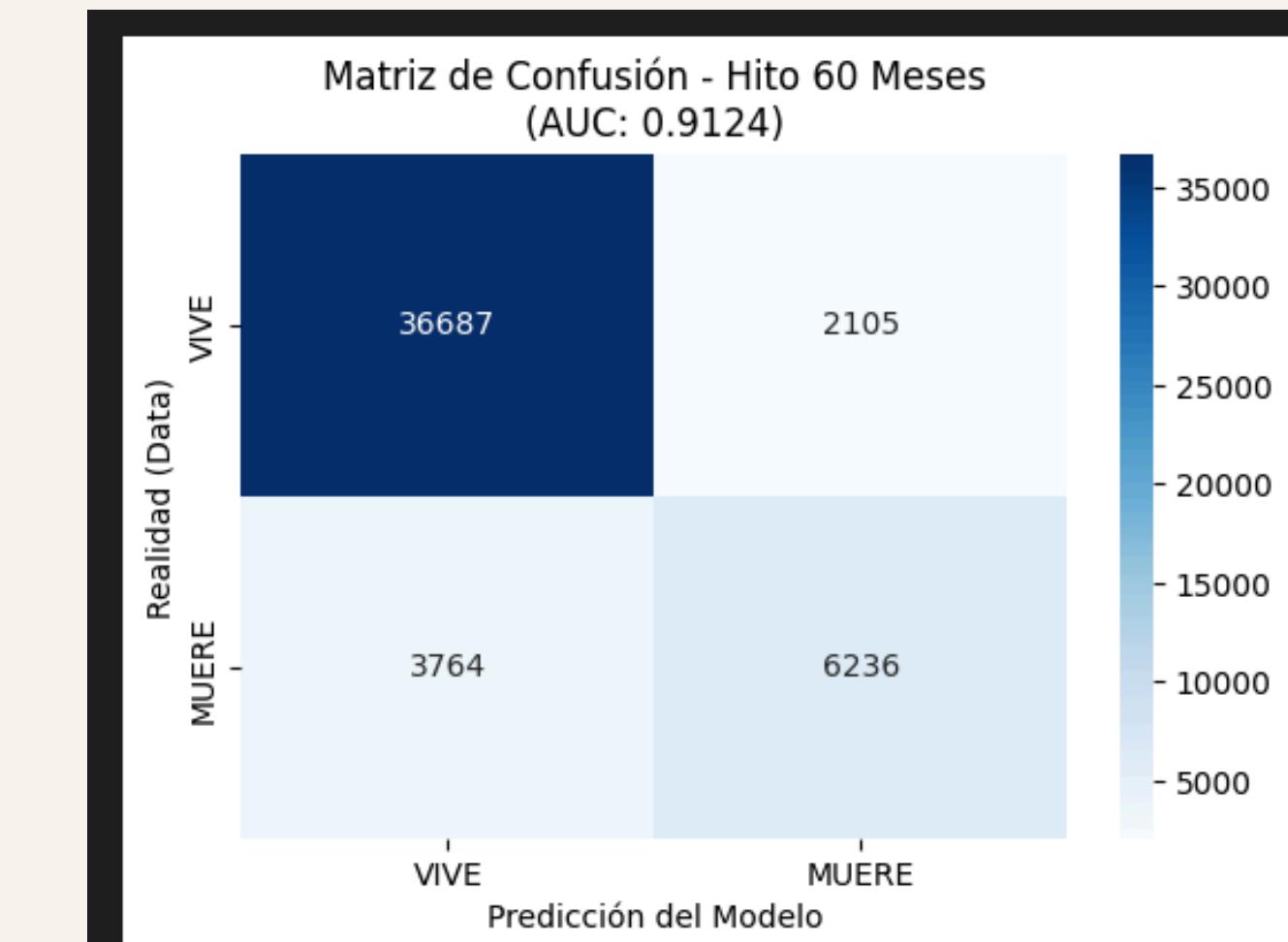
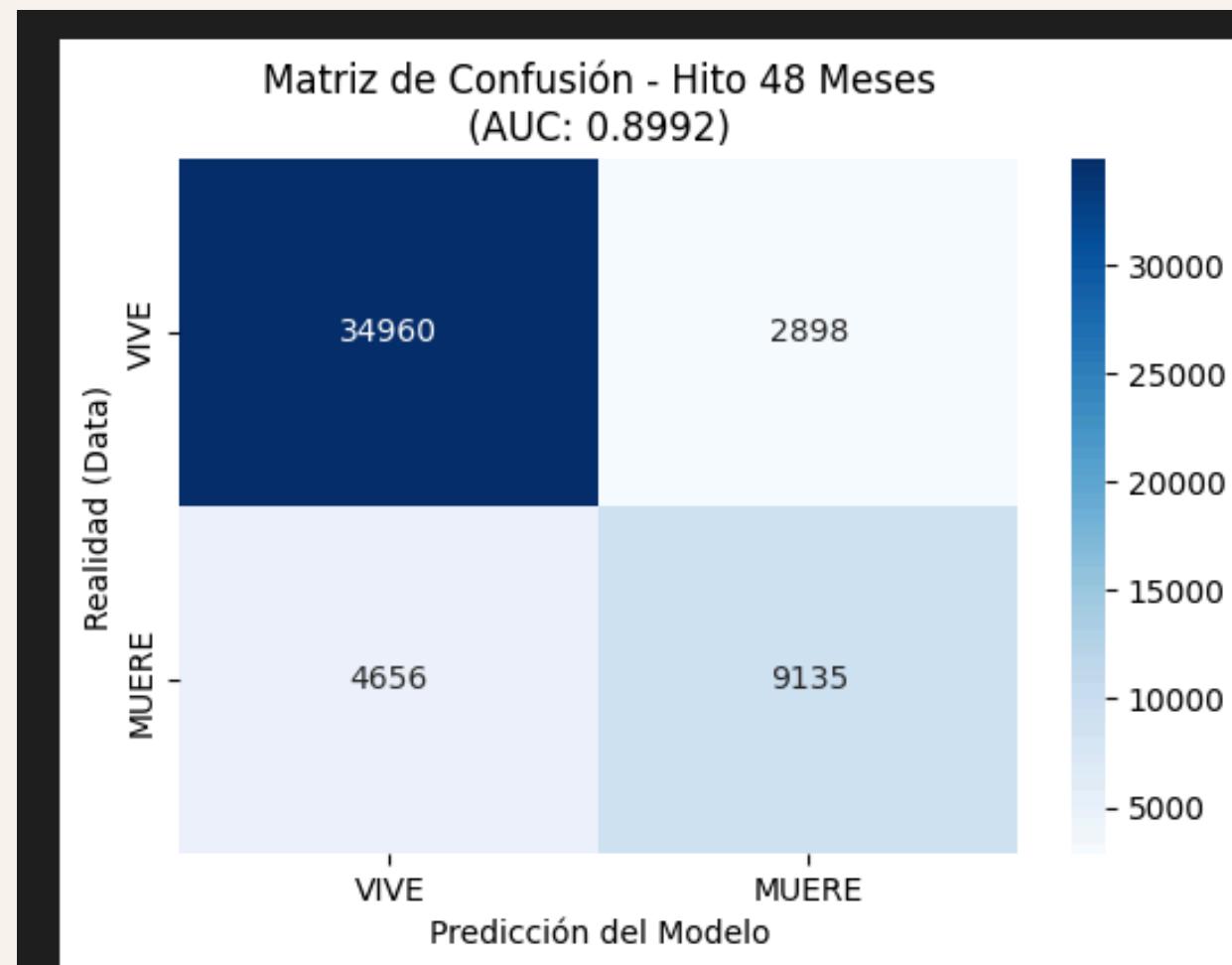
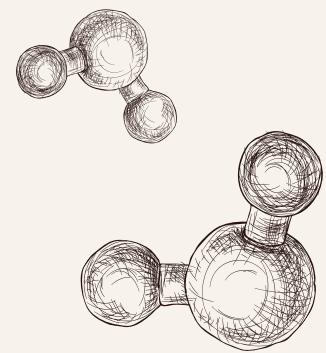
# EVALUACIÓN DEL MODELO



	precision	recall	f1-score	support
0	0.81	0.87	0.84	33252
1	0.82	0.74	0.78	25929
accuracy			0.81	59181
macro avg	0.81	0.80	0.81	59181
weighted avg	0.81	0.81	0.81	59181

	precision	recall	f1-score	support
0	0.85	0.90	0.88	36266
1	0.79	0.70	0.74	18944
accuracy			0.83	55210
macro avg	0.82	0.80	0.81	55210
weighted avg	0.83	0.83	0.83	55210

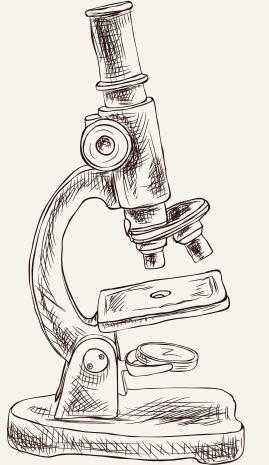
# EVALUACIÓN DEL MODELO



	precision	recall	f1-score	support
0	0.88	0.92	0.90	37858
1	0.76	0.66	0.71	13791
accuracy			0.85	51649
macro avg	0.82	0.79	0.80	51649
weighted avg	0.85	0.85	0.85	51649

	precision	recall	f1-score	support
0	0.91	0.95	0.93	38792
1	0.75	0.62	0.68	10000
accuracy			0.88	48792
macro avg	0.83	0.78	0.80	48792
weighted avg	0.87	0.88	0.88	48792

# PREDICCIÓN DE ESPERANZA DE VIDA

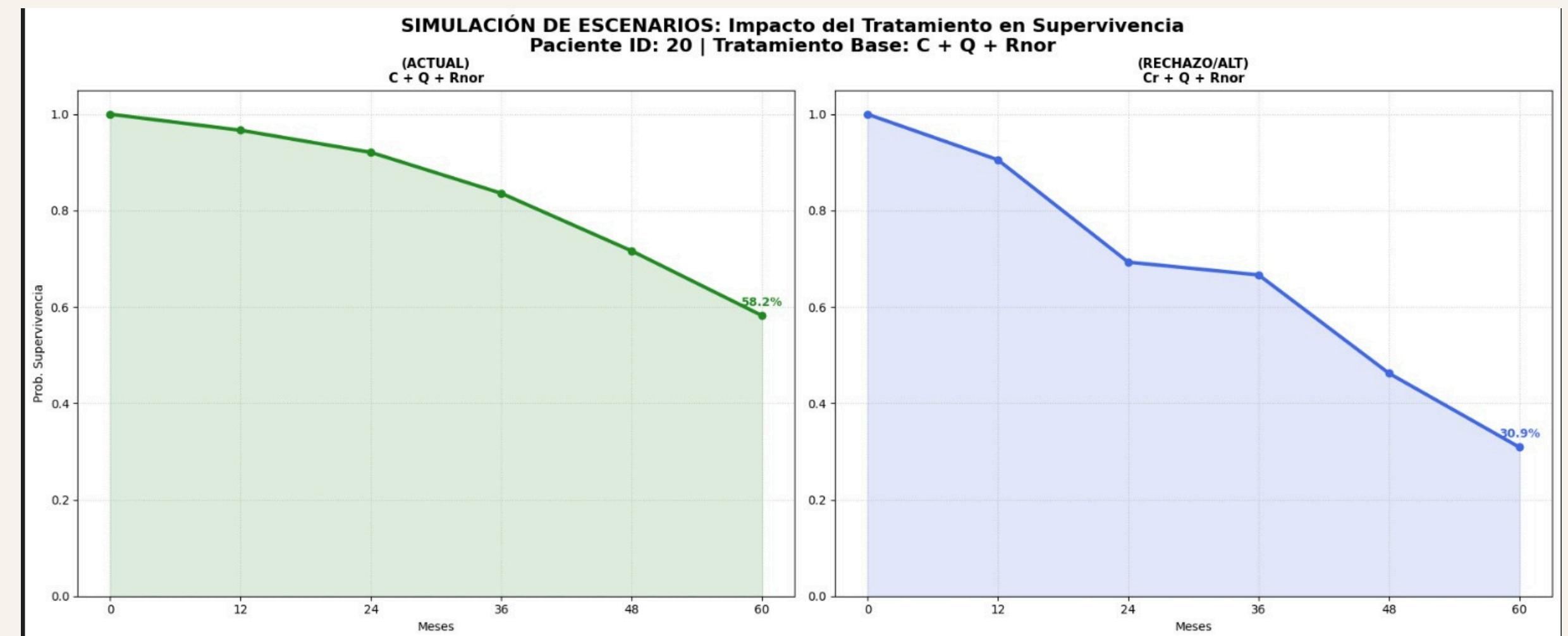
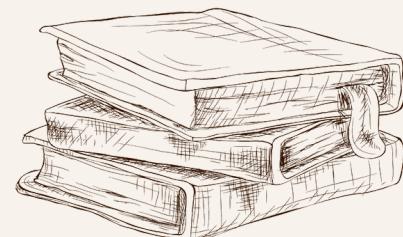


## Interpretación

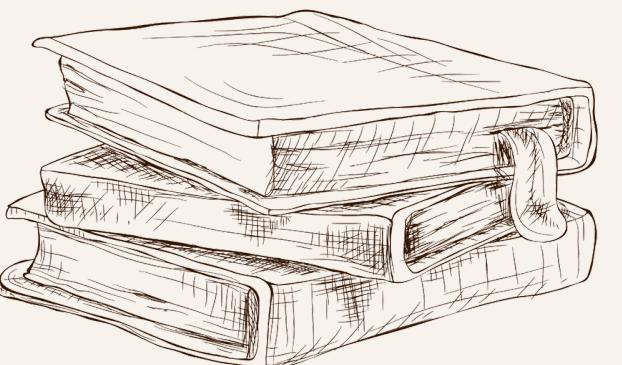
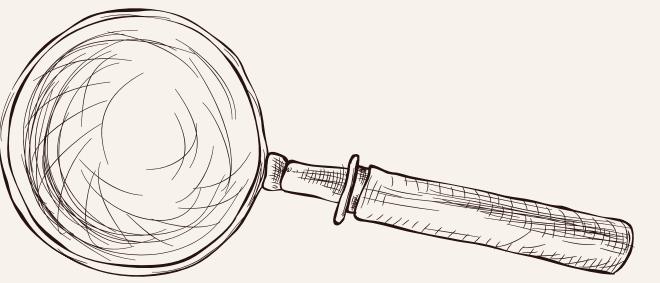
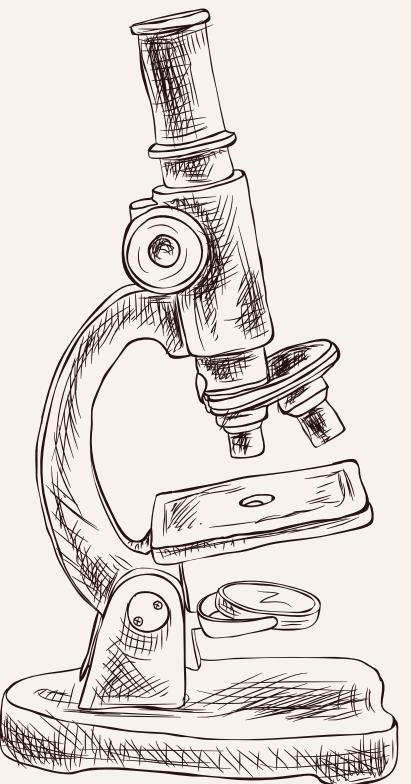
El modelo no solo predice la probabilidad de que un paciente sobreviva o fallezca, también estima la probabilidad de que siga vivo en cada paso del tiempo.

## Análisis de Sensibilidad

Un mapeo de tratamientos permite ver cómo cambia esa probabilidad si se altera el protocolo médico (ej. pasar de un tratamiento estándar a uno de rechazo).



# Streamlit



# Conclusiones

01

Los tumores detectados y tratados en un grado primario aumentan las probabilidades de supervivencia.

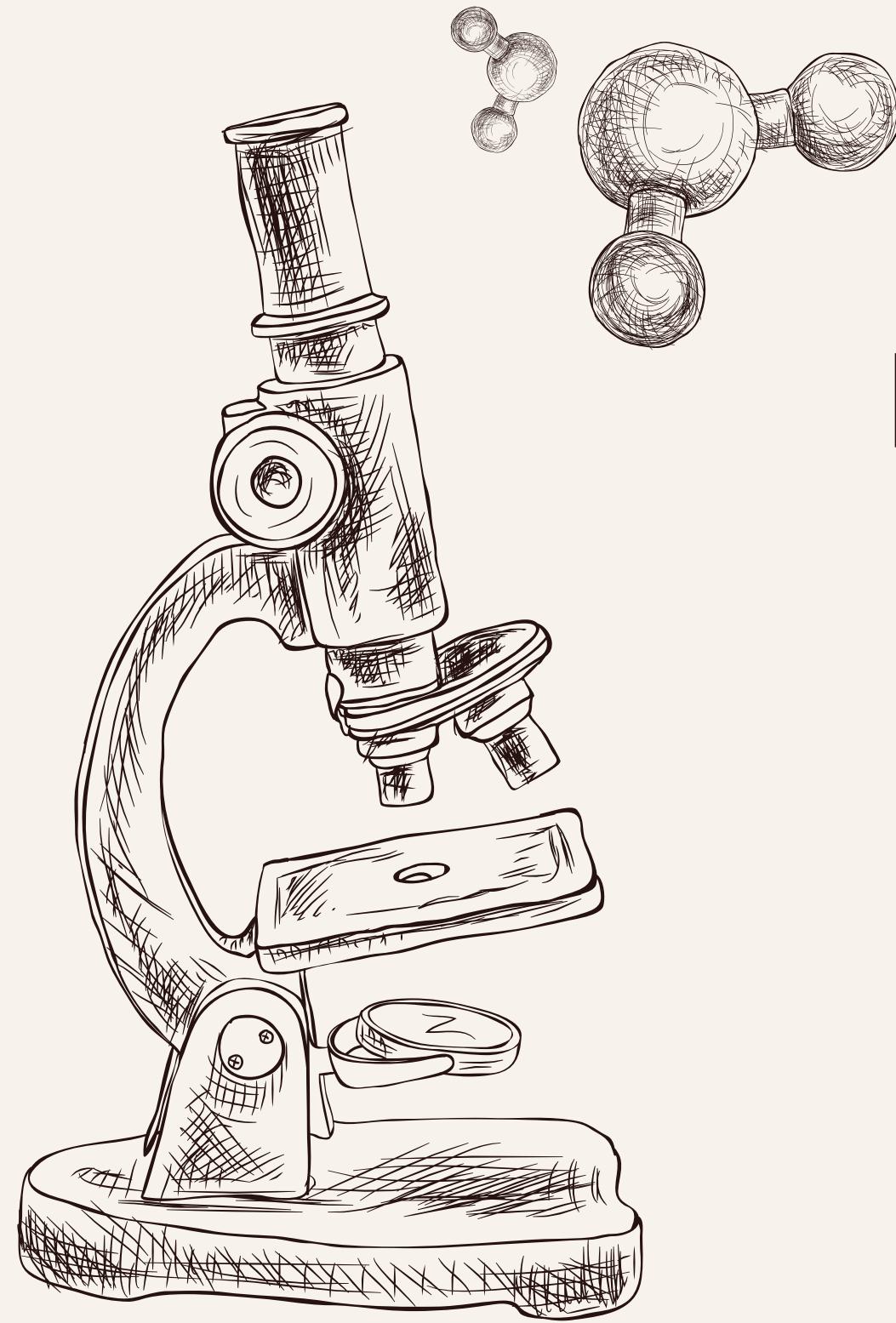
02

Las personas de edad avanzada tienen mayor riesgo de desarrollar cancer además se incrementa la complejidad de la intervención terapéutica.

03

Es vital que el tratamiento busque que el paciente viva más, pero también que viva bien, cuidando su estado de ánimo y usando métodos lo menos agresivos posible.





# Thank you!

**Do you have any questions?**

