Informe del Proyecto de Análisis de Supervivencia

Javier Méndez Parrilla

18 de diciembre de 2024

Índice

1. Introducción

El presente informe detalla el análisis de supervivencia realizado sobre un conjunto de datos relacionados con el cáncer de mama. Este análisis incluyó varias etapas, como el preprocesamiento de datos, un análisis univariante y bivariante, así como la construcción de curvas de supervivencia mediante el método Kaplan-Meier. Se presentan aquí los resultados más relevantes, acompañados de explicaciones sobre los métodos utilizados y su propósito.

2. Propósito del Código

El objetivo principal del código desarrollado fue explorar la relación entre diversas variables clínicas y la supervivencia global (SG) de los pacientes. Para ello, se realizaron los siguientes pasos:

1. Preprocesamiento de los datos para limpiar y transformar las columnas relevantes.
2. Análisis univariante para entender las distribuciones y frecuencias de cada variable.
3. Análisis bivariante para identificar relaciones significativas entre las variables y la SG.
4. Curvas de supervivencia global y por subgrupos específicos basados en el subtipo IHC (*immunohistochemistry*) del cáncer de mama.

3. Preprocesamiento de Datos

El análisis comenzó con un exhaustivo preprocesamiento de los datos proporcionados, debido a su complejidad y formato inicial. Este paso fue esencial para garantizar la calidad del análisis posterior y abordar las particularidades de las variables contenidas en el conjunto de datos.

3.1. Problemas Iniciales en los Datos

El conjunto de datos presentaba varias dificultades, que requirieron un tratamiento cuidadoso:

- **Fechas representadas como días desde un origen en Excel**: Varias columnas relacionadas con fechas importantes, como la de diagnóstico (`dateof_diagnosis`) o el último control (`fecha_estado_del_ultimo_control_julio`), estaban codificadas como días desde el 30 de diciembre de 1899, lo cual es el formato base de Excel. Para convertirlas al formato estándar de fecha, fue necesario utilizar transformaciones específicas.
- **Variables con valores vacíos o inconsistencias**: Se identificaron columnas con porcentajes significativos de valores faltantes, algunas de las cuales superaban el 20 % de celdas vacías. Estas columnas fueron descartadas del análisis.
- **Variables numéricas agrupadas**: La mayoría de las variables numéricas (como `positive_lymph_nodes`, `ki67`, `mammographic_tumor_size`) se agruparon en rangos clínicamente relevantes para facilitar la interpretación y el análisis estadístico.

- **Variables irrelevantes o redundantes**: Algunas columnas, como identificadores únicos (`patient_id`) o marcadores clínicos duplicados, fueron eliminadas para reducir la complejidad del análisis.

3.2. Columnas Finales del Conjunto de Datos

Tras el preprocesamiento, el conjunto de datos final incluyó las siguientes variables:

- **SG (Estado de Supervivencia Global)**: Variable categórica que indica si el paciente está vivo (`0`) o fallecido (`1`) al final del seguimiento.
- **TSG (Tiempo de Supervivencia Global)**: Variable numérica continua que representa el tiempo transcurrido, en años, entre el diagnóstico y el último control o fallecimiento.
- **recidiva**: Variable categórica (`SI/NO`) que indica si el paciente presentó una recidiva del cáncer.
- **menopausal_status**: Estado menopáusico del paciente (`PRE-PERIMENOPAUSAL` o `POSTMENOPAUSAL`).
- **typeof_chemotherapy**: Tipo de quimioterapia administrada.
- **typeof_hormonal_therapy**: Tipo de terapia hormonal administrada.
- **typeof_surgery**: Tipo de cirugía realizada (`CONSERVATIVE` o `RADICAL`).
- **adjuvant_radiotherapy**: Indica si se administró radioterapia adyuvante (`SI/NO`).
- **residual_cancer_burden**: Burden (carga residual) del cáncer después del tratamiento, evaluada cuantitativamente.
- **pam50subtype**: Subtipo molecular del cáncer según el análisis PAM50 (`Luminal A, Luminal B, HER2-enriched, Basal-like`).
- **lymph_nodes_resected_cat**: Número de ganglios linfáticos extirpados agrupados (`<=5, 6-10, 11-20, >20`).
- **positive_lymph_nodes_cat**: Número de ganglios linfáticos positivos agrupados (`0, 1-3, 4-6, >6`).
- **ihc_subtype**: Subtipo inmunohistoquímico del cáncer (`Luminal, HER2, Triple Negative`).
- **ki67_cat**: Índice de proliferación celular (`<=10, 11-20, 21-50, >50`).
- **mammographic_tumor_size_cat**: Tamaño tumoral agrupado (`<=2 cm, 2-5 cm, 5-10 cm, >10 cm`).
- **final_tumor_size_cat**: Tamaño final del tumor agrupado después del tratamiento (`<=2 cm, 2-5 cm, >5 cm`).
- **ror_cat**: Riesgo de recurrencia agrupado (`Bajo, Intermedio, Alto`).
- **patientage_cat**: Edad del paciente categorizada ($\leq$ 40, `41-50, 51-60, 61-70,` $>$ 70).

3.3. Consideraciones sobre la Imputación de Valores Faltantes

En el contexto de datos biológicos, imputar valores faltantes debe hacerse con extrema precaución debido a las siguientes razones:

- **Imputación por moda**: Aunque es adecuada para variables categóricas, puede introducir sesgos significativos si una categoría es dominante. En nuestro caso, la imputación por moda se utilizó sólo para las variables categóricas tras verificar que las distribuciones no se viesen distorsionadas.
- **Imputación por media en variables numéricas**: Este método no fue utilizado porque podría reducir la variabilidad intrínseca de las variables continuas y enmascarar patrones relevantes, especialmente en datos clínicos donde las diferencias individuales son críticas.
- **Contexto biológico**: En estudios de supervivencia, la pérdida de datos puede reflejar diferencias importantes en el seguimiento o el manejo clínico, por lo que se evitó imputar datos clave como `TSG` o `SG`.

3.4. Importancia de `SG` y `TSG` en el Análisis

Estas dos variables son el núcleo del análisis de supervivencia:

- `SG` define el estado del paciente como el evento de interés (fallecimiento).
- `TSG` mide el tiempo hasta dicho evento y permite construir las curvas de supervivencia, que son esenciales para identificar patrones y diferencias entre grupos.

El preprocesamiento, aunque laborioso, aseguró que el conjunto de datos final fuera adecuado para responder las preguntas clínicas planteadas.

4. Análisis Univariante

Se exploraron las distribuciones de las principales variables del conjunto de datos. A continuación, se presentan algunos de gráficos a modo de representación de las variables categóricas analizadas:

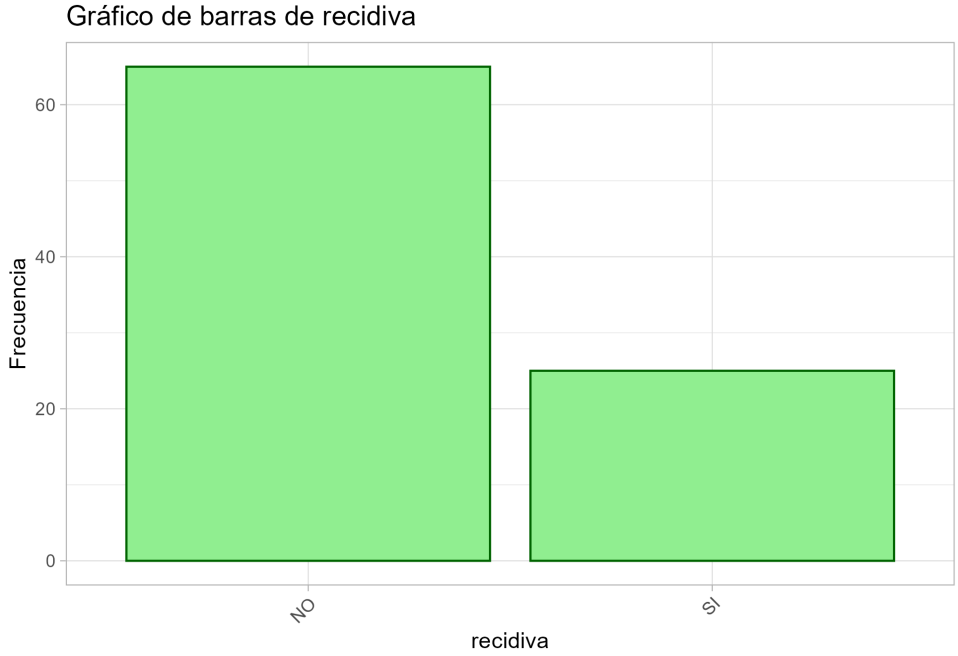

Figura 1: Distribución de *recidiva*

La Figura 11 muestra la distribución de la variable recidiva en la población estudiada. La mayoría de los pacientes no presentó recidiva ("NO"), con más de 60 casos, mientras que aproximadamente 25 pacientes sí experimentaron recidiva ("SÍ").

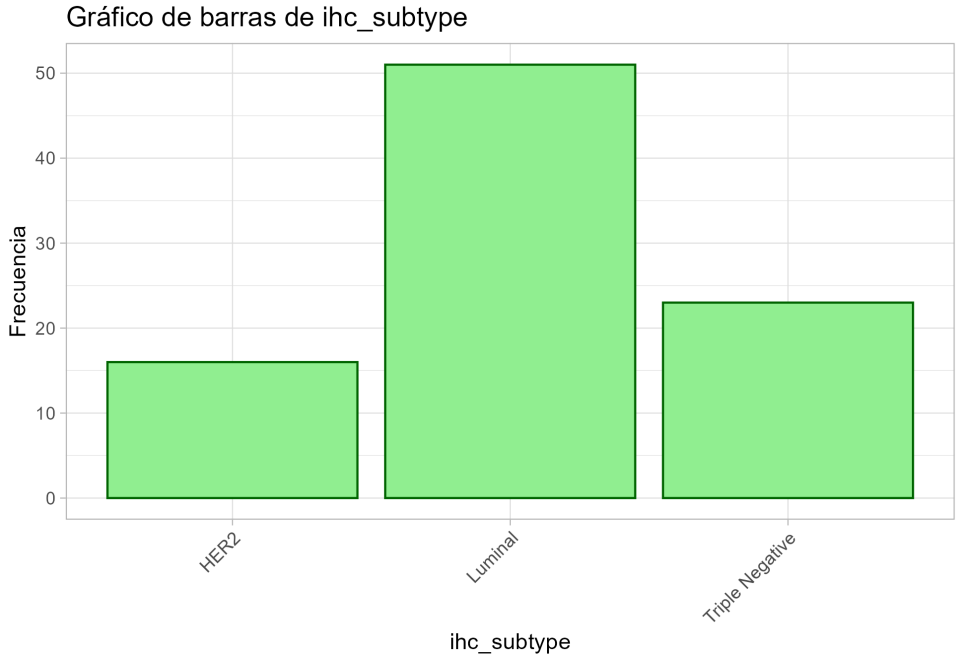

Figura 2: Distribución del *subtipo inmunohistoquímico*

La Figura 2 muestra la distribución de los subtipos de inmunohistoquímica (IHC) en la población estudiada. Se observa que el subtipo "Luminal" es el más frecuente, con más de 50 casos, seguido por el subtipo "Triple Negative" y finalmente "HER2", que presenta la menor frecuencia.

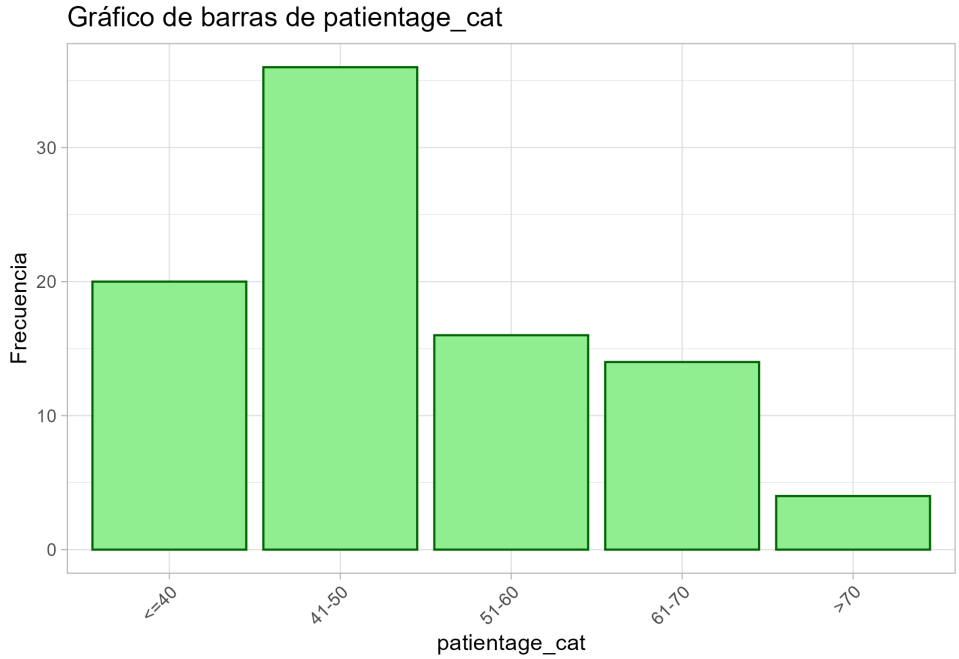

Figura 3: Distribución de la *edad* de los pacientes

La Figura 3 presenta la distribución de los pacientes según categorías de edad. El grupo de edad "41-50 años" es el más representado, con más de 35 casos. Le siguen los grupos "$\leq$ 40 años", "51-60 años" y "61-70 años", con frecuencias similares. Finalmente, el grupo de "$>$ 70 años" tiene la menor representación.

Resumen

En resumen, los gráficos de barras evidencian la predominancia del subtipo "Luminal", una mayor representación del grupo de edad "41-50 años" y una tendencia clara hacia la ausencia de recidiva en la mayoría de los pacientes estudiados. Estos hallazgos proporcionan una visión general de las características clínicas de la muestra analizada.

5. Análisis Bivariante

Se realizaron pruebas estadísticas como Chi-cuadrado y el coeficiente V de Cramer para medir la asociación entre las variables categóricas y la Supervivencia Global (SG). Las variables con valores de p significativos y mayor asociación se destacan a continuación, junto con representaciones gráficas.

5.1. Resultados del test Chi-cuadrado y V de Cramer

	variable	chi_p_value	cramer_value
1	recidiva	0.0004997501	0.609
2	lymph_nodes_resected_cat	0.0244877561	0.325
3	positive_lymph_nodes_cat	0.0579710145	0.288
4	final_tumor_size_cat	0.0749625187	0.260
5	typeof_hormonal_therapy	0.0844577711	0.208
6	ki67_cat	0.1034482759	0.262
7	residual_cancer_burden	0.1439280360	0.250
8	pam50subtype	0.1504247876	0.238
9	typeof_surgery	0.1869065467	0.161
10	typeof_chemotherapy	0.2808595702	0.233
11	ihc_subtype	0.5242378811	0.126
12	ror_cat	0.6671664168	0.180
13	adjuvant_radiotherapy	0.7441279360	0.068
14	patientage_cat	0.8185907046	0.129
15	menopausal_status	1.0000000000	0.018
16	mammographic_tumor_size_cat	NaN	NaN

Figura 4: Resultados del test Chi-cuadrado y V de Cramer por variable.

La Figura 4 muestra los valores del test Chi-cuadrado y del coeficiente V de Cramer para todas las variables categóricas. Las variables `recidiva` ($p = 0.0005$, V $= 0.609$) y `lymph_nodes_resected_cat` ($p = 0.024$, V $= 0.325$) muestran las asociaciones más significativas con la Supervivencia Global (SG), seguidas por `positive_lymph_nodes_cat` y `final_tumor_size_cat`.

5.2. Relación entre SG y ganglios linfáticos extirpados

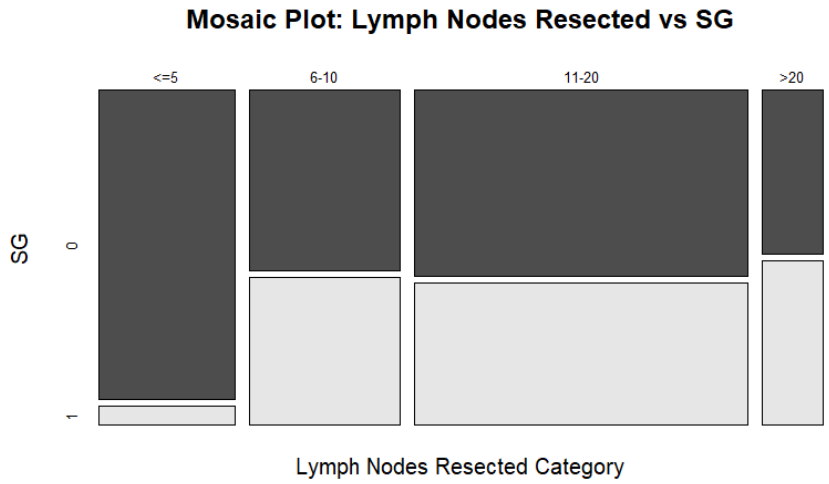

Figura 5: Gráfico de mosaico: SG vs ganglios linfáticos extirpados.

La Figura 5 presenta un gráfico de mosaico que ilustra la relación entre la Supervivencia Global (SG) y el número de ganglios linfáticos extirpados, agrupados en categorías. Se observa lo siguiente:

- La categoría ≤ 5 `ganglios` muestra una alta proporción de pacientes vivos (`SG = 0`).
- A medida que aumenta el número de ganglios linfáticos extirpados (categorías `11-20` y `>20`), la proporción de pacientes fallecidos aumenta considerablemente.

Esto sugiere que la cantidad de ganglios extirpados puede influir en la supervivencia global, posiblemente debido a que cuantos más ganglios han sido extirpados, más podrían quedar en el paciente por extirpar o tal vez los daños causados en el tejido pese a haber sido extirpados afecta a la supervivencia de dichos pacientes.

5.3. Distribución de SG según recidiva

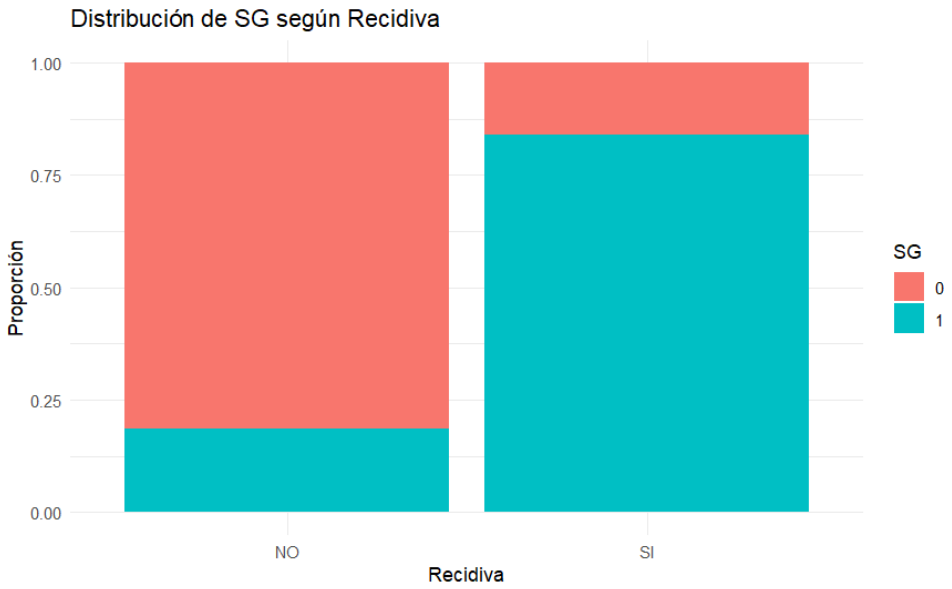

Figura 6: Distribución de SG según la presencia de recidiva.

La Figura 6 muestra la distribución de la Supervivencia Global (SG) según la presencia de recidiva. Se pueden observar las siguientes tendencias:

- En los pacientes sin recidiva (`NO`), la mayoría permanecen vivos (`SG = 0`), con una proporción considerablemente baja de fallecidos.
- Por el contrario, en los pacientes con recidiva (`SI`), la proporción de fallecidos (`SG = 1`) aumenta significativamente.

Estos resultados sugieren una fuerte asociación entre la recidiva del cáncer y una menor Supervivencia Global, como lo indica también el p-valor significativo y el alto coeficiente V de Cramer ($p = 0.0005$, V $= 0.609$).

5.4. Resumen

En el análisis bivariante, las variables con mayor asociación estadística con la Supervivencia Global (SG) fueron:

- **Recidiva**: La presencia de recidiva está fuertemente asociada con un peor pronóstico.
- **Lymph_nodes_resected_cat**: La cantidad de ganglios linfáticos extirpados muestra una tendencia donde un mayor número se asocia con una peor Supervivencia Global.

El resto de las variables no mostraron asociaciones estadísticamente significativas en este análisis.

6. Curvas de Supervivencia

Las curvas de Kaplan-Meier se generaron para evaluar la probabilidad de supervivencia global (SG) en función del tiempo, considerando distintos factores clínicos y subgrupos. A continuación, se presentan los resultados más relevantes.

6.1. Supervivencia Global

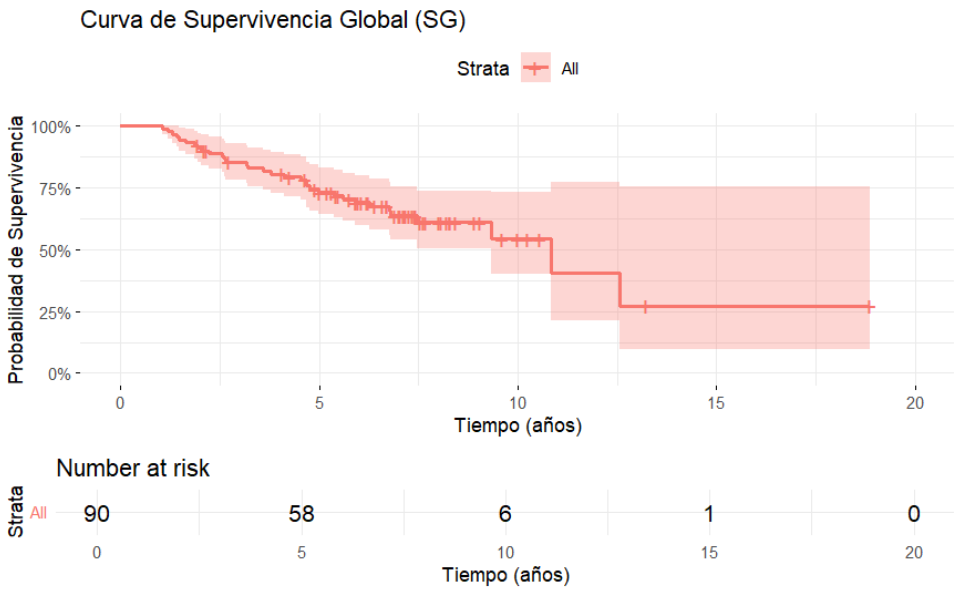

Figura 7: Curva de Supervivencia Global (SG) para todos los pacientes.

La Figura 7 muestra la curva de supervivencia global para toda la cohorte. Se observa que la probabilidad de supervivencia disminuye progresivamente con el tiempo, alcanzando aproximadamente el 25 % a los 20 años de seguimiento.

6.2. Comparación según Tipo de Cirugía

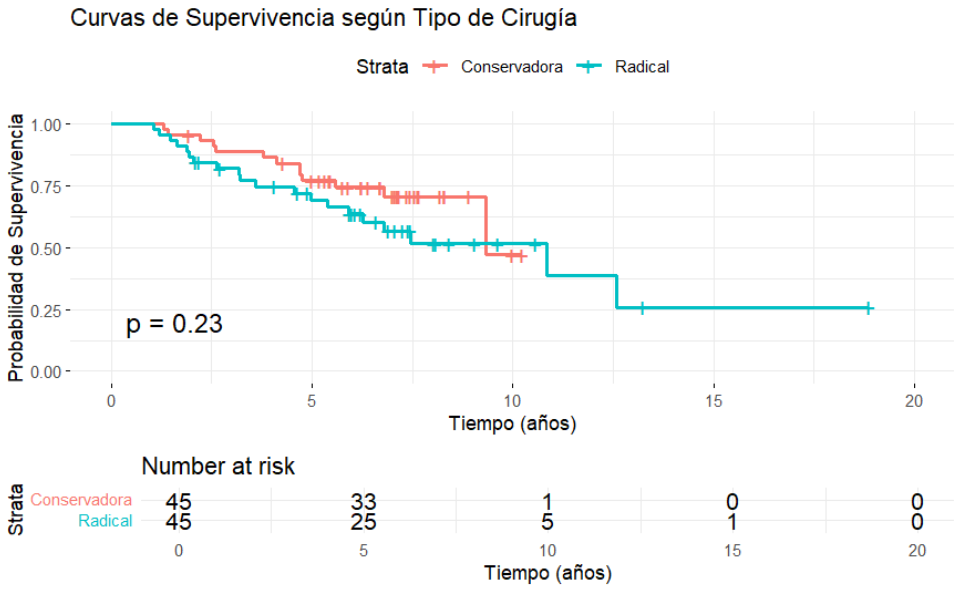

Figura 8: Curvas de Supervivencia según Tipo de Cirugía (Conservadora vs Radical).

En la Figura 8, se comparan las curvas de supervivencia entre los pacientes sometidos a cirugía conservadora y radical. Aunque se observa una ligera tendencia a mejores resultados en la cirugía conservadora, la diferencia no es estadísticamente significativa ($p = 0.23$).

6.3. Comparación según Ganglios Linfáticos Extirpados

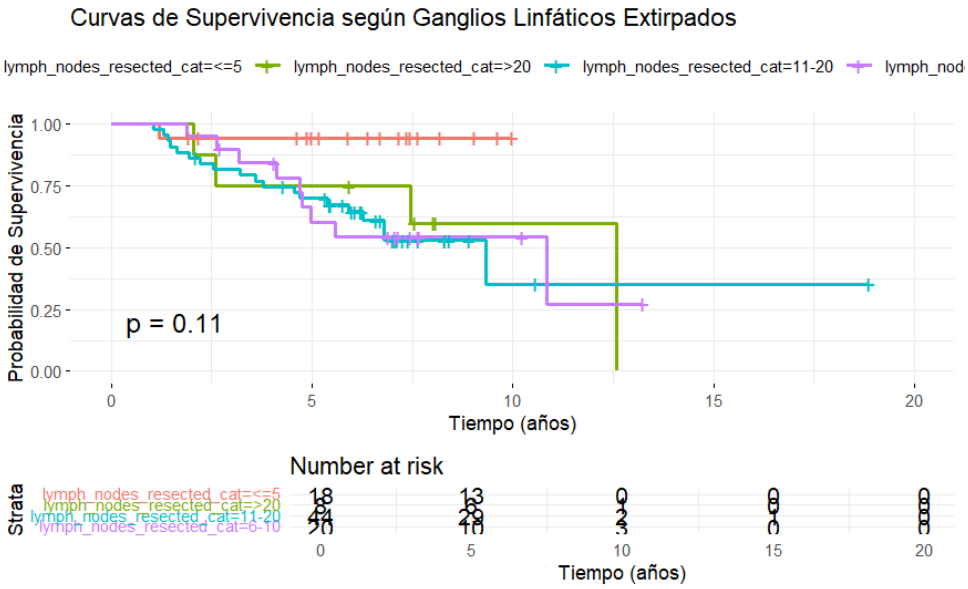

Figura 9: Curvas de Supervivencia según Ganglios Linfáticos Extirpados.

La Figura 9 presenta la comparación de supervivencia según el número de ganglios linfáticos extirpados. Los pacientes con `>20 ganglios` presentan una tendencia a una

mejor supervivencia, aunque la diferencia no es estadísticamente significativa ($p = 0.11$).

6.4. Comparación según Ganglios Linfáticos Positivos

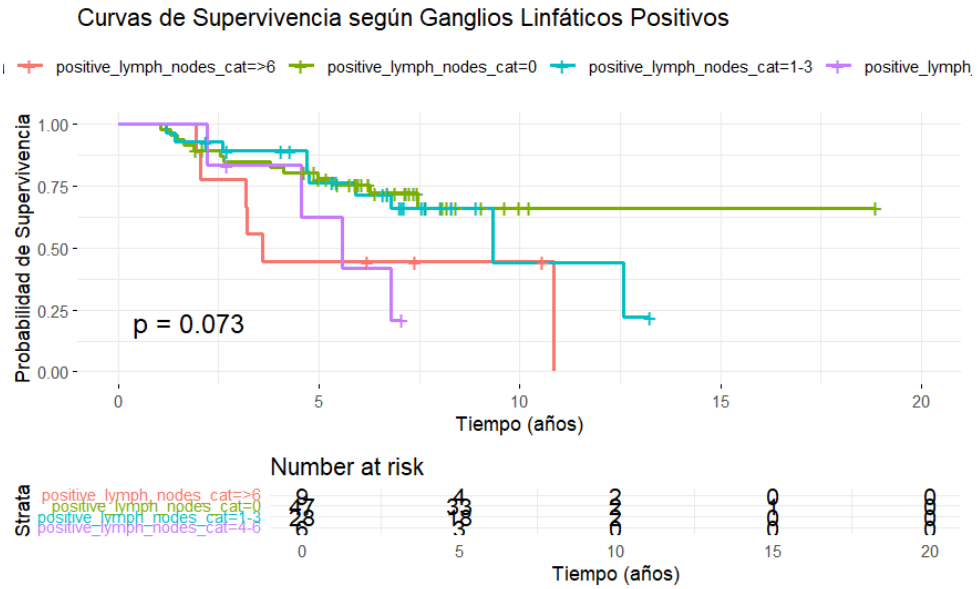

Figura 10: Curvas de Supervivencia según Ganglios Linfáticos Positivos.

En la Figura 10, se comparan las curvas de supervivencia en función del número de ganglios linfáticos positivos. Los pacientes sin ganglios positivos (`cat = 0`) muestran la mejor probabilidad de supervivencia, mientras que aquellos con `>=6 ganglios` tienen un pronóstico significativamente peor ($p = 0.073$).

6.5. Comparación según Recidiva

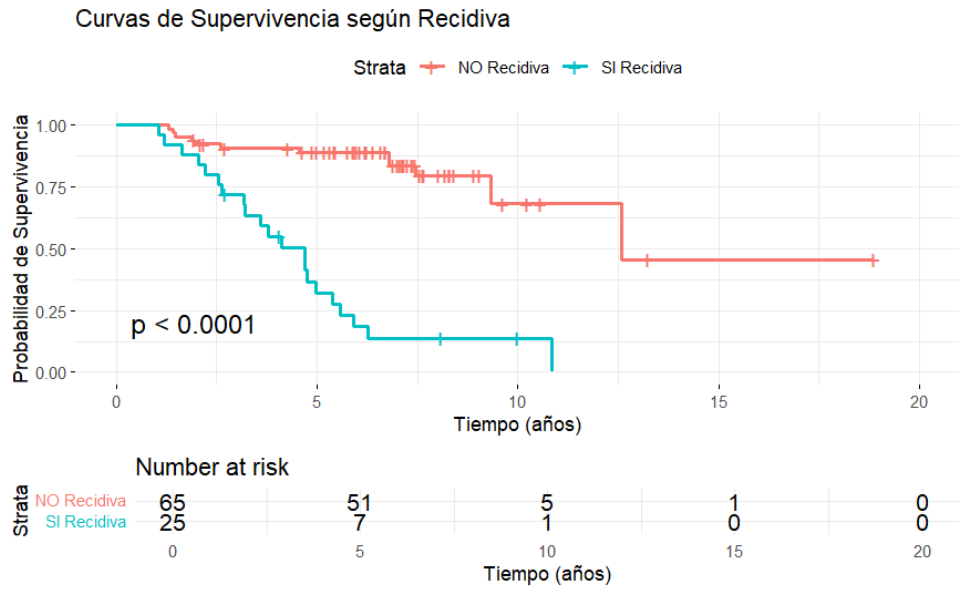

Figura 11: Curvas de Supervivencia según la presencia de Recidiva.

La Figura 11 muestra un efecto significativo de la recidiva en la supervivencia global ($p < 0.0001$). Los pacientes con recidiva (`SI`) tienen una disminución considerable en la probabilidad de supervivencia comparado con aquellos sin recidiva (`NO`).

6.6. Curvas por Subtipo IHC

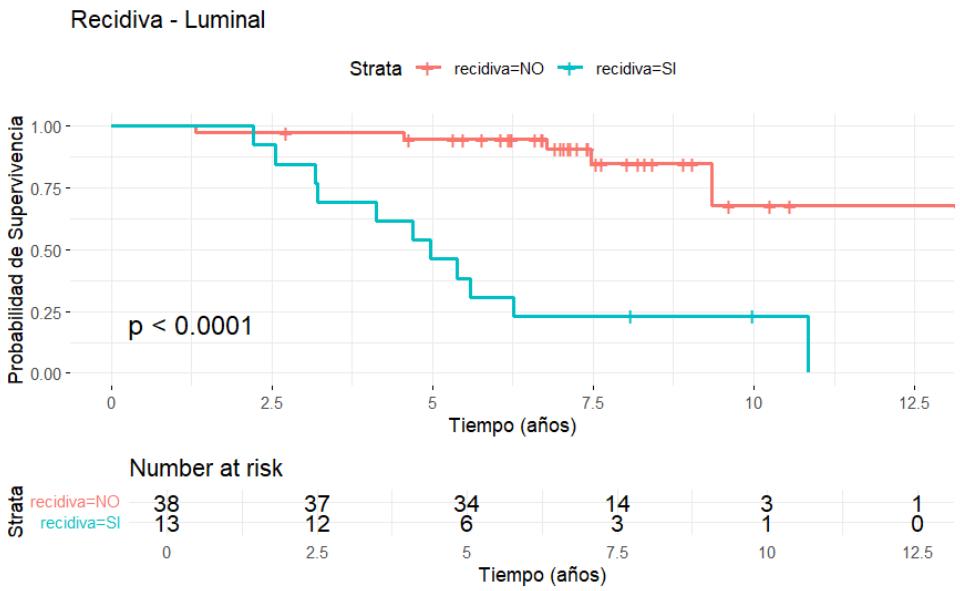

Figura 12: Curva de Supervivencia - Subtipo Luminal según Recidiva.

Luminal La Figura 12 muestra las curvas de supervivencia para el subtipo Luminal. Se observa que la presencia de recidiva (`SI`) tiene un efecto negativo significativo en la supervivencia ($p < 0.0001$).

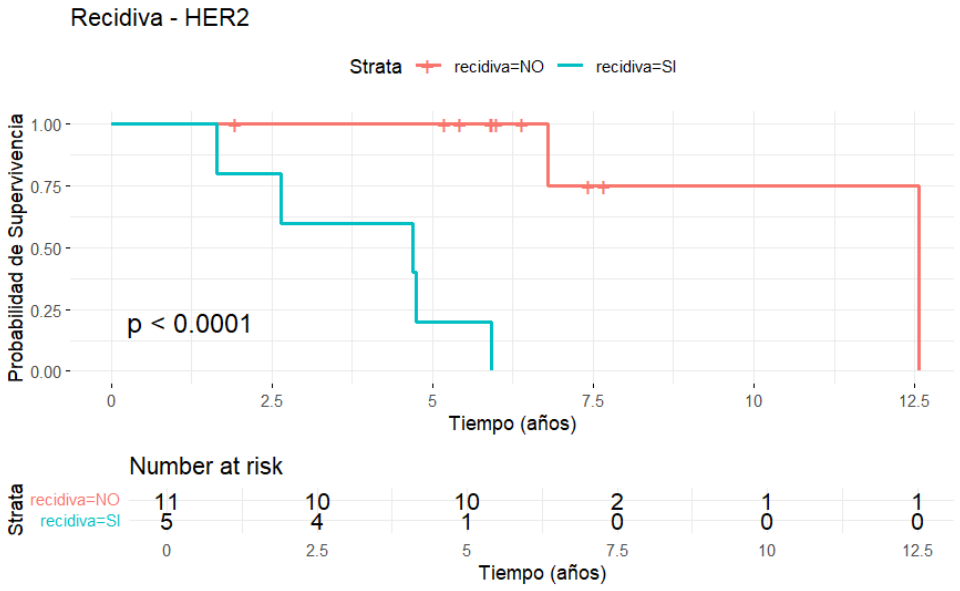

Figura 13: Curva de Supervivencia - Subtipo HER2 según Recidiva.

HER2 En la Figura 13, los pacientes con subtipo HER2 y recidiva presentan una peor supervivencia global en comparación con aquellos sin recidiva (p ¡0.0001).

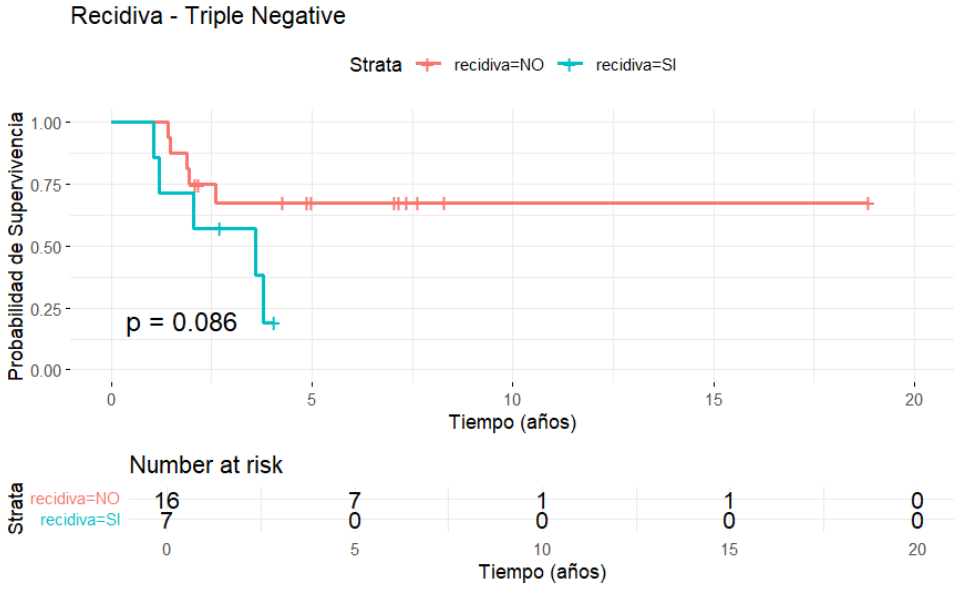

Figura 14: Curva de Supervivencia - Subtipo Triple Negative según Recidiva.

Triple Negative Finalmente, la Figura 14 muestra las curvas de supervivencia para el subtipo Triple Negative. Aunque existe una tendencia a una menor supervivencia en pacientes con recidiva ($p = 0.086$), la diferencia no alcanza significancia estadística.

7. Conclusiones

El análisis de supervivencia realizado en este estudio permitió identificar los siguientes hallazgos relevantes:

- La recidiva del cáncer tuvo un impacto significativo en la supervivencia global, con una diferencia notable entre pacientes con y sin recidiva (p ¡0.0001).
- El número de ganglios linfáticos extirpados y el número de ganglios linfáticos positivos mostraron asociaciones relevantes con la supervivencia, aunque no alcanzaron significancia estadística en algunos casos.
- Los subtipos IHC presentaron diferencias en la supervivencia, especialmente en el subtipo HER2 y Luminal, donde la recidiva tuvo un efecto particularmente negativo.
- La cirugía conservadora mostró una ligera ventaja en comparación con la cirugía radical, aunque esta diferencia no fue estadísticamente significativa.

En conclusión, la presencia de recidiva y los factores relacionados con los ganglios linfáticos son determinantes clave en la supervivencia global de los pacientes. Estos resultados pueden servir como base para futuros estudios y estrategias clínicas orientadas a mejorar el pronóstico de los pacientes con cáncer.

Referencias

A continuación se listan las librerías utilizadas en el análisis y visualización de datos en R:

- **DescTools** [7]: Utilizada para el cálculo del coeficiente V de Cramer, que mide la asociación entre variables categóricas.
- **vcd** [2]: Empleada para la creación de gráficos de mosaico, facilitando la visualización de relaciones entre variables categóricas.
- **survival** [8]: Implementa métodos para análisis de supervivencia, incluyendo modelos de Kaplan-Meier.
- **survminer** [1]: Complementa a `survival` al permitir la visualización de curvas de Kaplan-Meier de manera personalizada y profesional.
- **ggplot2** [9]: Librería fundamental para la creación de gráficos y visualización de datos en R.
- **readxl** [5]: Utilizada para la lectura de archivos Excel como fuente de datos.
- **dplyr** [6]: Proporciona herramientas para la manipulación y limpieza eficiente de datos.
- **lubridate** [4]: Facilita el manejo y la manipulación de fechas en análisis temporales.
- **janitor** [3]: Utilizada para limpiar y ordenar nombres de variables en dataframes.

Estas librerías permitieron realizar análisis estadísticos, visualizaciones avanzadas y manipulación eficiente de datos, contribuyendo al desarrollo completo del presente estudio.

Referencias

[1] Marcin Kosinski Alboukadel Kassambara. *Survminer: Drawing Survival Curves using 'ggplot2'*, 2023. R package version 0.4.9.

[2] Achim Zeileis David Meyer. *Visualizing Categorical Data*, 2023. R package version 1.4-11.

[3] Sam Firke. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2023. R package version 2.2.0.

[4] Hadley Wickham Garrett Grolemund. *Lubridate: Make Dealing with Dates a Little Easier*, 2023. R package version 1.9.2.

[5] Jennifer Bryan Hadley Wickham. *Readxl: Import Excel Files into R*, 2023. R package version 1.4.0.

[6] Romain François Hadley Wickham. *dplyr: A Grammar of Data Manipulation*, 2023. R package version 1.1.2.

[7] Andri Signorell. *DescTools: Tools for Descriptive Statistics*, 2023. R package version 0.99.48.

[8] Terry M. Therneau. *Survival Analysis*, 2023. R package version 3.5-5.

[9] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016.