UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Comparación de modelos de supervivencia interpretables para datos multi-ómicos de alta dimensionalidad

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Alós Maldonado, Raúl

Tutor/a: Tarazona Campos, Sonia

Director/a Experimental: SALGUERO GARCIA, PEDRO

CURSO ACADÉMICO: 2022/2023

Agradecimientos

Este trabajo no hubiera sido posible sin el apoyo de mis tutores, Sonia y Pedro, a los que muestro mi agradecimiento por la oportunidad ofrecida.

A mis familiares, amigos y a Elena, por su apoyo incondicional.

Resumen

Los datos multiómicos, esto es, bases de datos de alta dimensionalidad formadas por bloques de variables, comúnmente atribuidos a información biológica y clínica en pacientes, se emplean cada vez más en la investigación de numerosas patologías para la búsqueda de biomarcadores. Las variables de interés de dichos ensayos clínicos contienen información sobre el tiempo hasta un evento de interés y su censura, y deben ser tratados mediante técnicas estadísticas basadas en el análisis de supervivencia. Para hacer frente a datos multiómicos, y sus características, se seleccionaron numerosos modelos de predicción para su comparación en rendimiento, selección de variables y coste computacional, empleando dos bases de datos procedentes de *The Cancer Genome Atlas*. Algunos de los métodos comparados están basados en regresión penalizada, otros son algoritmos *boosting* o *random forest*. A su vez, se pueden categorizar en función de si implementa una estrategia multi – bloque al analizar la base de datos. Estos métodos se compararon con el modelo Cox usando únicamente variables clínicas como predictores. Los parámetros de rendimiento empleados fueron el *cindex* y el Brier *score*. Los resultados indican que los modelos que implementan la estrategia multi – bloque tienen, por lo general, un mejor rendimiento, al facilitar la inclusión de variables con más información predictiva pertenecientes a bloques de menor dimensión. La selección de modelos y el análisis de las variables seleccionadas ha permitido validar su utilidad en la detección de factores pronóstico, al identificar biomarcadores consolidados para ambas patologías de estudio. El análisis del tiempo de ejecución de los modelos limita la aplicación de algunos con un rendimiento adecuado, al ser significativamente superior a modelos con un rendimiento similar y una mejor interpretabilidad.

Palabras clave:

análisis de supervivencia; datos multiómicos; cáncer; *machine learning*.

Abstract

Multiomic data, that is, high-dimensional databases formed by blocks of variables commonly attributed to biological and clinical information in patients, are increasingly used in the research of numerous pathologies to search for biomarkers. The variables of interest in these clinical trials contain information about the time until an event of interest and its censorship and must be treated using statistical techniques based on survival analysis. To address multiomic data and its characteristics, numerous prediction models were selected for comparison in terms of performance, variable selection, and computational cost, using two databases from The Cancer Genome Atlas. Some of the compared methods are based on penalized regression, while others are boosting or random forest algorithms. They can also be categorized based on whether they implement a multi-block strategy when analyzing the database. These methods were compared with the Cox model using only clinical variables as predictors. The performance parameters used were the cindex and the Brier score. The results indicate that models implementing the multi-block strategy generally perform better by allowing the inclusion of variables with more predictive information from smaller-dimensional blocks. The selection of models and the analysis of the selected variables have validated their utility in detecting prognostic factors by identifying established biomarkers for both study pathologies. The analysis of model execution time limits the application of some models with adequate performance, as it is significantly higher than models with similar performance and better interpretability.

Key words:

survival analysis; multiomic data; cancer; *machine learning.*

Resum

Les dades multiòmiques, això és, bases de dades d'alta dimensionalitat formades per blocs de variables, comunament atribuïts a informació biològica i clínica en pacients, s'empren cada vegada més en la investigació de nombroses patologies per a la cerca de biomarcadors. Les variables d'interés d'aquests assajos clínics contenen informació sobre el temps fins a un esdeveniment d'interés i la seua censura, i han de ser tractats mitjançant tècniques estadístiques basades en l'anàlisi de supervivència. Per a fer front a dades multiòmiques, i les seues característiques, es van seleccionar nombrosos models de predicció per a la seua comparació en rendiment, selecció de variables i cost computacional, emprant dues bases de dades procedents de *The Cancer Genome Atlas*. Alguns dels mètodes comparats estan basats en regressió penalitzada, uns altres són algorismes *boosting* o *random forest*. Al seu torn, es poden categoritzar en funció de si implementa una estratègia multi – bloc en analitzar la base de dades. Aquests mètodes es van comparar amb el model Cox usant únicament variables clíniques com a predictors. Els paràmetres de rendiment emprats van ser el *cindex* i el Brier *score*. Els resultats indiquen que els models que implementen l'estratègia multi – bloc tenen, en general, un millor rendiment, en facilitar la inclusió de variables amb més informació predictiva pertanyents a blocs de menor dimensió. La selecció de models i l'anàlisi de les variables seleccionades ha permés validar la seua utilitat en la detecció de factors pronòstic, en identificar biomarcadors consolidats per a totes dues patologies d'estudi. L'anàlisi del temps d'execució dels models limita l'aplicació d'alguns amb un rendiment adequat, en ser significativament superior a models amb un rendiment similar i una millor interpretabilitat.

Paraules clau:

anàlisi de supervivència; dades multi - òmiques; càncer; *machine learning.*

Índice

Índice de figuras

Índice de tablas

Lista de acrónimos y símbolos

ADN	Ácido desoxirribonucleico
ARN	Ácido ribonucleico
MPLE	*Maximum partial pikelihood – estimator*
TCGA	The Cancer Genome Atlas
NIH	Institutos Nacionales de la Salud
BRCA	*Breast Invasive Carcinoma*
HNSC	*Head and Neck Squamous Cell Carcinoma*
TMM	*Trimmed Mean of M – Values*
CQN	*Conditional Quantile Normalization*
HPV	Virus del papiloma humano
LASSO	*Least Absolute Shrinkage and Selection Operator*
IPF	*Integrative LASSO with Penalty Factor*
PH	*Proportional Hazards*
RFSC	*Random Survival Forest*
OOB	*Out of bag*
CHF	*Cumulative Hazard Function*
CV	Validación cruzada
EPV	*Events per Variable*

1. Introducción

1.1 Datos multiómicos

Los datos ómicos se refieren a un tipo de información que describe interacciones y funciones de diversos sistemas biológicos, como organismos, células o tejidos. Este campo ha sido impulsado, en gran medida, por avances tecnológicos, como la técnica de hibridación de microarrays o la modelización de sistemas biológicos, que han hecho posible el análisis de alto rendimiento de moléculas biológicas (Hasin et al., 2017).

Estos avances han permitido, junto a una secuenciación del genoma humano de alta calidad y herramientas estadísticas con grandes cohortes de miles de pacientes, el descubrimiento de miles de variantes genéticas, que han contribuido a la caracterización de numerosas patologías (Hasin et al., 2017).

Este tipo de datos puede dividirse, según su origen, en distintas categorías:

- Genómica: es la rama de la biología molecular que estudia el genoma de los organismos, empleada para la identificación de variantes genéticas asociadas a patologías o con respuestas a tratamientos específicos dentro de la investigación médica. Esto conlleva la genotipificación de miles de individuos utilizando múltiples marcadores genéticos para la identificación de diferencias estadísticamente significativas entre grupos de casos y controles, lo que puede proporcionar evidencia de asociaciones genéticas.
- Epigenómica: se centra en la caracterización de modificaciones reversibles del ADN, como su metilación. Estas modificaciones pueden estar originadas por factores genéticos o ambientales, y se ha demostrado que poseen una gran importancia en procesos biológicos y desarrollo del cáncer o patologías cardiacas, entre otras (Hasin et al., 2017).
- Transcriptómica: analiza los niveles de ARN de todo el genoma, tanto cualitativamente (qué transcritos están presentes, etc.) cómo cuantitativamente (expresión de cada transcrito). Este marcador es un intermediario molecular entre el ADN y las proteínas que se da bajo ciertas circunstancias. Proporciona información sobre cómo un conjunto de genes trabaja para la formación de rutas biológicas metabólicas y regulatorias dentro de la célula (Xiong, 2006).
- Proteómica: es el estudio a gran escala de las proteínas, más concretamente, de su función y estructura. Cuantifica la abundancia, modificación e interacción de péptidos, y, por tanto, de la función de los genes. A diferencia de la genómica, proporciona un enfoque más directo para el entendimiento de funciones celulares, dado que la mayoría de las funciones de los genes se realizan por medio de las proteínas (Xiong, 2006). Por tanto, a través de la proteómica, se puede obtener una comprensión más profunda de las funciones celulares y los procesos biológicos.
- Metabolómica: cuantifica múltiples tipos de moléculas, como aminoácidos, ácidos grasos, carbohidratos y otras moléculas producto de las funciones metabólicas de las células. Su medición ayuda a la detección de ciertas patologías, cuando sus niveles se encuentran fuera de rangos normales. En conjunto con la modelización, se ha utilizado ampliamente para el estudio del flujo de metabolitos (Hasin et al., 2017).
- Microbiómica: consiste en el estudio de diversas poblaciones de microorganismos encontrados en el cuerpo humano: piel, superficies mucosas y el intestino, colonizados por bacterias, virus y hongos. Es decir, se centra en la microbiota y los genes que la constituyen. Su variación puede atribuirse no solo a una amplia variedad de patologías, sino que también se asocia a factores ambientales, el uso de drogas o la edad, siendo por tanto enormemente compleja.

La integración de estos tipos de datos ómicos da lugar al término de datos multiómicos, que proporcionan un mejor entendimiento de las relaciones entre la causa original de la patología con las

consecuencias funcionales e interacciones relevantes (Xiong, 2006). Dicha integración es necesaria para la detección de los cambios causales, y no únicamente reactivos (Hasin et al., 2017).

Una síntesis de este concepto se presenta en el siguiente gráfico (Figura 1) (Hasin et al., 2017), donde se pueden comprobar las distintas capas o bloques de diversos datos ómicos que conforman estas bases de datos. Considerando cada círculo como una molécula, se observan las interacciones o correlaciones potenciales, en forma de flecha, detectadas entre moléculas en distintas capas: por ejemplo, el transcrito rojo puede estar correlacionado con múltiples proteínas. Aunque no estén representadas, cabe tener en cuenta la posibilidad de que existan interacciones dentro de una misma capa.

Además, el gráfico permite ver cómo la consolidación de los datos ómicos proviene tanto de un enfoque genómico como de un enfoque de fenotipo, que actúa en todos los niveles exceptuando el propio genoma. Así, exceptuando el genoma, todas las capas de datos se ven afectadas tanto por una regulación genética (eje horizontal) como por una regulación del entorno (eje vertical), lo cual influye de distintas formas a cada molécula individual (Hasin et al., 2017).

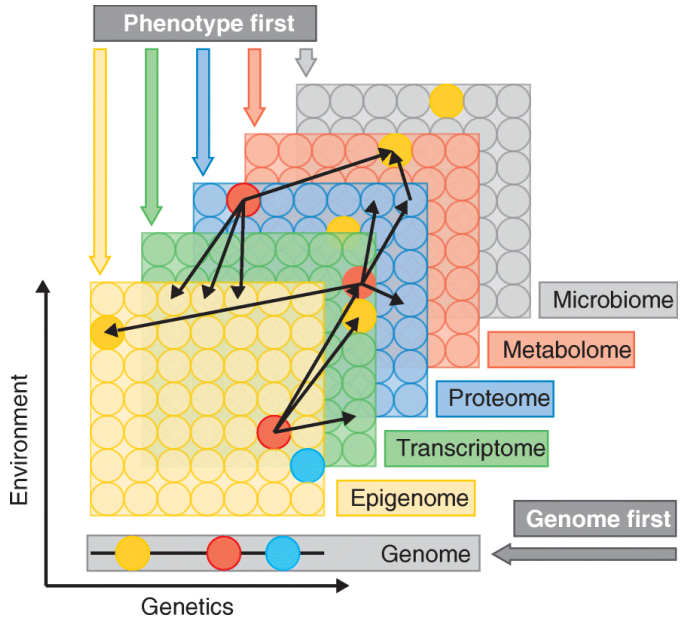

Figura 1: Tipologías de datos ómicos. Cada capa o color describe un tipo distinto. Cada capa contiene el conjunto completo de moléculas, representadas con círculos. Las flechas negras finas representan interacciones potenciales o correlaciones entre moléculas de distintas capas. Las flechas más gruesas indican marcos conceptuales para la consolidación de cada capa. Gráfico de Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. In Genome Biology (Vol. 18, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-017-1215-1.

La integración de estos marcos lleva a la generación de bases de datos de grandes dimensiones. La capacidad de detectar relaciones de interés en éstas depende en gran medida de su heterogeneidad y del tamaño de la muestra. Además, al ser estudios humanos, se debe considerar la presencia de factores de confusión, como su dieta o su estilo de vida (Hasin et al., 2017).

Consecuentemente, dicha dimensionalidad conlleva un replanteamiento de las herramientas estadísticas dedicadas al análisis de datos, y se la suele denominar "maldición de la dimensionalidad" (Donoho, 2000). Si se plantea un problema de regresión tradicional entre un conjunto de variables explicativas contenidas en una matriz $\boldsymbol{X}$ para la predicción de una variable respuesta contenida en el vector $\boldsymbol{y}$, los

estimadores de máxima verosimilitud $\widehat{\boldsymbol{\beta}}$ de sus parámetros se obtienen mediante la expresión (Johnstone & Titterington, 2009):

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad (1)$$

Donde $\boldsymbol{X}$ es una matriz de dimensiones $n \times p$, siendo *n* el número de muestras y *p* el número de variables explicativas. Para dicho ajuste, debe cumplirse que $p \leq n$, dado que, de lo contrario, la expresión ($\boldsymbol{X^T X}$) es singular y la estimación de los parámetros de regresión no es exacta (Johnstone & Titterington, 2009).

Por tanto, en el contexto de bases de datos de alta dimensionalidad, donde $p > n$ o $p \gg n$ deben plantearse nuevas estrategias para asegurar la no singularidad de ($\boldsymbol{X^T X}$), además de que la probabilidad de que haya multicolinealidad es mucho mayor (Johnstone & Titterington, 2009), más en el contexto de bases de datos multiómicos.

1.2 Análisis de supervivencia

El análisis de supervivencia es un campo de la estadística que trata el análisis y modelización de datos donde la variable respuesta es el tiempo hasta que sucede un evento de interés. Una de sus principales aplicaciones es de determinar los factores pronósticos que influyen sobre el tiempo de supervivencia a una determinada enfermedad (Wang et al., 2019).

Hay ocasiones en las que dichos eventos no se observan para ciertos individuos. Dicho fenómeno se denomina censura, y puede darse por distintas razones, definiendo tres posibles tipos de censura:

- Censura derecha: cuando el tiempo de supervivencia observado es menor o igual que el tiempo de supervivencia real.
- Censura izquierda: cuando el tiempo de supervivencia observado es mayor o igual que el tiempo de supervivencia real.
- Censura en un intervalo: cuando solo se conoce un tiempo de intervalo dado en el que el evento de interés ha ocurrido.

El caso que se presenta más comúnmente en análisis de supervivencia es la censura derecha, en la que, dentro del área de salud, eventos de interés como la muerte, recuperación o recaída en una patología no llegan a registrarse (Jenkins, 2005). Existen diversas razones para esto, como el no experimentar dicho evento antes de la finalización del estudio, la pérdida del seguimiento del paciente o el origen de otro evento que lleva a su retiro del estudio (Kleinbaum & Klein, n.d.).

Un ejemplo de esta tipología de datos se resume en el siguiente gráfico (Figura 2) (Wang et al., 2019), en el que se simula la participación de seis sujetos en un estudio de supervivencia. Dos de ellos, los sujetos cinco y dos (S5 – S2) llegan al final del estudio sin experimentar el evento, mientras que los sujetos tres y uno (S3 – S1) no lo experimentan por otros factores, como su retiro de dicho estudio, siendo estos cuatro los sujetos que presentan censura. Por otra parte, los sujetos seis y cuatro (S6 – S4) son aquellos que realmente han experimentado el evento durante su monitorización.

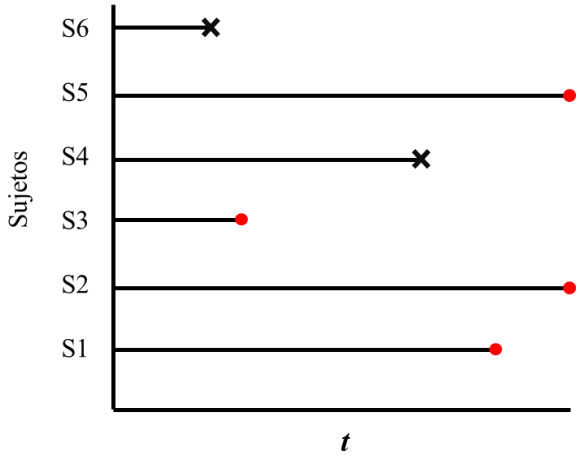

Figura 2: Gráfico que demuestra el problema en análisis de supervivencia. Se representa el avance de distintos sujetos (eje Y) en función del tiempo (eje X). Un punto rojo indica censura, una cruz indica el evento de interés. Gráfico adaptado de Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. ACM Computing Surveys, 51(6). https://doi.org/10.1145/3214306.

Existen multitud de estudios en los que el registro de datos de supervivencia, junto a las correspondientes técnicas estadísticas, es necesario, como ensayos clínicos para el estudio del tiempo de remisión en pacientes con leucemia, análisis del tiempo de infección en pacientes con diálisis renal o estudio del tiempo hasta el fallecimiento de pacientes con cáncer de lengua (Klein & Moeschberger, 2003).

Otro de los objetivos de este tipo de análisis es la obtención de las funciones de supervivencia y de riesgo (Wang et al., 2019). Por una parte, la función de supervivencia ($S(t)$) se usa para la representación de la probabilidad de que el tiempo hasta el evento de interés no es menor que un tiempo *t*. Esta función tiene la siguiente expresión:

$$S(t) = P(T \geq t) \qquad (2)$$

Donde *T* es el tiempo de supervivencia. Esta función decrece a medida que aumenta *t*, tomando un valor inicial de 1 para *t* igual a 0, asumiendo que al comienzo de la monitorización de los individuos todos tienen un 100% de probabilidad de supervivencia.

Por otra parte, la función de riesgo ($h(t)$) se refiere como la probabilidad de que se produzca el evento de interés instantáneamente. Dicha función no negativa y sin límite superior sigue la expresión (Wang et al., 2019):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \qquad (3)$$

Por lo que describe el límite de la tasa instantánea de fallo cuando el incremento de *t* tiende a cero. A diferencia de la función de supervivencia, no tiene una forma concreta y puede seguir distintas distribuciones.

Estas distribuciones pueden ajustarse a diversos contextos en el campo de la investigación médica. A continuación, se muestran una serie de ejemplos de la función de riesgo $h(t)$ en base al entorno de aplicación (Figura 3) (Kleinbaum & Klein, n.d.).

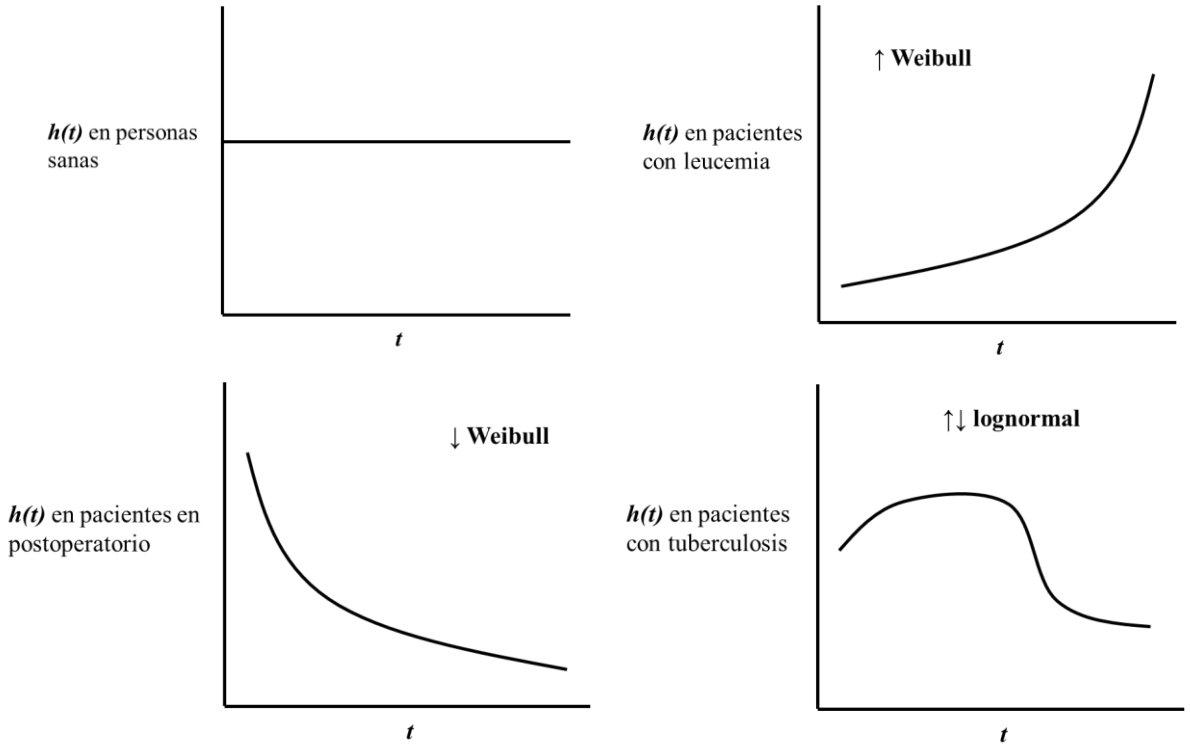

Figura 3: Distribuciones de funciones de riesgo en diversos contextos en investigación médica. Se lee de izquierda a derecha, de arriba abajo. Gráfico adaptado de Kleinbaum, D. G., & Klein, M. (2012). Kaplan-Meier Survival Curves and the Log-Rank Test (pp. 55–96). https://doi.org/10.1007/978-1-4419-6646-9_2.

El primer gráfico muestra un riesgo constante en un estudio con personas sanas, de manera que, independientemente del valor de t, $h(t)$ mantiene el mismo valor.

En segundo lugar, el gráfico de $h(t)$ que describe el riesgo en paciente con leucemia sigue una distribución de Weibull creciente. La evolución de dicha función en este caso es lógica en el caso de pacientes con leucemia sin respuesta al tratamiento, donde el evento de interés es la muerte, y el riesgo del paciente aumenta con el tiempo.

En el tercer gráfico, a diferencia del anterior, el riesgo disminuye con el paso del tiempo, describiendo una distribución de Weibull decreciente. Esto se da, por ejemplo, en pacientes en una etapa de postoperatorio, en la que el evento es la muerte y su riesgo es alto en etapas tempranas de la recuperación.

Otra posible distribución se muestra en el cuarto gráfico de $h(t)$, en el que se describe un incremento del riesgo, seguido de una disminución momentánea, que es ejemplo de un modelo de supervivencia con una distribución lognormal. Esto es esperable, por ejemplo, en pacientes con tuberculosis, dado que su probabilidad de muerte aumenta en etapas tempranas de la enfermedad, experimentando una reducción del riesgo con el tiempo.

1.2.1 Modelización no paramétrica

La función de supervivencia y su representación gráfica es el recurso más utilizado en el análisis de supervivencia. Esta técnica se incluye en la modelización no paramétrica del tiempo de supervivencia, para la cual se conocen diversas técnicas. Entre ellas, destaca la curva de Kaplan Meier (Kaplan & Meier, 1958), la cual, adicionalmente, facilita la comparación de distintas funciones de supervivencia anteriormente mencionadas (Kleinbaum & Klein, n.d.).

Curvas de Kaplan Meier:

Sean $T_1 < T_2 < \cdots < T_k$ un conjunto de tiempos de supervivencia dados hasta el evento de interés para $N(K \leq N)$ individuos. Para cada tiempo hasta el evento $T_j(j = 1, 2, \dots, K)$ el número de eventos observados es $d_j \geq 1$, y r_j el número de muestras en riesgo, cuyo tiempo de censura o de supervivencia es mayor o igual que T_j (Wang et al., 2019).

Para la obtención de r_j se emplea la siguiente expresión (Wang et al., 2019):

$$r_j = r_{j-1} - d_{j-1} - c_{j-1} \qquad (4)$$

Donde c_{j-1} es el número de casos censurados durante el periodo de tiempo entre T_j y T_{j-1}. La probabilidad de supervivencia para cada instante se definirá como (Wang et al., 2019):

$$p(T_j) = \frac{r_j - d_j}{r_j} \qquad (5)$$

Empleando dicha expresión de forma recursiva para el cálculo de la función de supervivencia, se obtiene su estimación, tal que(Wang et al., 2019):

$$\hat{S}(t) = \prod_{j:T_j<t} p(T_j) \qquad (6)$$

Una forma común de emplear la estimación de dicha función en el análisis de supervivencia es su comparación en base a un factor de interés mediante la prueba estadística de *logrank* para dos grupos.

Dicha prueba evalúa si ambas curvas de supervivencia de Kaplan Meier son equivalentes, o análogamente, que no existe evidencia para afirmar que las poblaciones que constituyen cada curva sean distintas, planteando así la hipótesis nula correspondiente (Kleinbaum & Klein, 2012):

$$H_0: No\ hay\ diferencias\ entre\ las\ curvas\ de\ supervivencia$$

En caso de que se pueda asumir que la H_0 es cierta, el estadístico *logrank* seguirá una distribución χ^2 con un grado de libertad, y se obtiene de acuerdo con la siguiente expresión (Kleinbaum & Klein, 2012):

$$X^2 = \sum_{i}^{número\ de\ grupos} \frac{(O_i - E_i)^2}{E_i} \qquad (7)$$

En este caso, se dispone de dos grupos ($i = 2$), y se obtiene el número esperado de eventos para cada instante j en cada grupo siguiendo las siguientes expresiones (Kleinbaum & Klein, 2012):

$$e_{1j} = \left(\frac{r_{1j}}{r_{1j} + r_{2j}}\right) \cdot \left(d_{1j} + d_{2j}\right) \qquad (8)$$

$$e_{2j} = \left(\frac{r_{2j}}{r_{1j} + r_{2j}}\right) \cdot (d_{1j} + d_{2j}) \qquad (9)$$

De esta forma, se obtienen los parámetros para la obtención del estadístico (Kleinbaum & Klein, 2012):

$$O_i - E_i = \sum_{j}^{K} (d_{ij} - e_{ij}) \qquad (10)$$

$$E_i = \sum_{j}^{K} (e_{ij}) \qquad (11)$$

Este test estadístico puede emplearse por tanto para la determinación de factores de interés que, a priori, tienen un efecto sobre el tiempo de supervivencia. Estos factores, en el contexto del campo de investigación médica, pueden ser, por ejemplo, la aplicación de distintos tratamientos o la aplicación de un tratamiento en comparación con un placebo (Kleinbaum & Klein, 2012).

Cabe tener en cuenta que este tipo de herramientas estadísticas no paramétricas deben ser complementarias a otros tipos de análisis de supervivencia, dada la posibilidad de que se obtengan estimaciones inexactas (Wang et al., 2019).

1.2.2 Modelos de Riesgos Proporcionales de Cox

El modelo de Cox es el modelo más comúnmente utilizado en el análisis de supervivencia dentro de la categoría de modelos semiparamétricos (Wang et al., 2019). En comparación a las técnicas estadísticas no paramétricas, tiene la ventaja de que no es necesaria la asunción o conocimiento de la distribución

que sigue el tiempo de supervivencia, aunque eso haga más difícil su interpretación. Además, admite la incorporación de variables explicativas para el ajuste del modelo.

Definición del modelo:

Para una observación *i* dada ($i = 1, 2, \dots, N$), se tiene el vector de variables explicativas $\boldsymbol{x_i}$, junto al indicador binario de censura δ_i ($\delta_i = 0$ para casos censurados y $\delta_i = 1$ para casos no censurados) y el tiempo observado y_i, tal que (Wang et al., 2019):

$$y_i = \begin{cases} T_i \ \ si\ \delta_i = 1 \\ C_i \ \ si\ \delta_i = 0 \end{cases} \qquad (12)$$

Por tanto, para cada individuo *i*, se define la función de riesgo $h(t, \boldsymbol{x_i})$ del modelo de Cox, que sigue la asunción de riesgos proporcionales, dada por la siguiente expresión (Wang et al., 2019):

$$h(t, X_i) = h_0(t) \exp(x_i \beta) \qquad (13)$$

Donde $h_0(t)$ es la función de riesgo base y $\boldsymbol{\beta^T} = (\beta_1,\ \beta_2, \dots, \beta_p)$ el vector de coeficientes asociados a las *p* variables explicativas. Dicha expresión combina una función base de riesgo tiempo dependiente y una componente exponencial que involucra las variables explicativas, sin incluir el tiempo (Wang et al., 2019).

Por tanto, el modelo de Cox consiste en la modelización de la función de riesgo, considerándola como función de las covariantes. Así, es posible estudiar los efectos de las variables explicativas sobre el tiempo de vida. Asimismo, este modelo se define como semi – paramétrico dado que la función de riesgo base no tiene una forma específica (Wang et al., 2019).

Puede obtenerse el cociente de riesgos, o *hazard ratio*, entre dos individuos 1 y 2 como (Wang et al., 2019):

$$HR = \frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t) \exp(x_1 \beta)}{h_0(t) \exp(x_2 \beta)} = \exp[(x_1 - x_2)\beta] \qquad (14)$$

En caso de que dicho cociente sea mayor que 1, el riesgo en el instante *t* del primer individuo es mayor. Además, se puede observar cómo es totalmente independiente de la función de riesgo base, por lo que el modelo de Cox se define de riesgos proporcionales dado que dicho cociente es una constante y todos los sujetos comparten la misma función de riesgo base (Wang et al., 2019).

Bajo esta asunción, la función de supervivencia puede obtenerse como (Wang et al., 2019):

$$S(t) = S_0(t)^{\exp(X\beta)} \qquad (15)$$

Donde $S_0(t)$ es la función base de supervivencia.

La estimación de los coeficientes se lleva a cabo por medio de la función de verosimilitud parcial, o *partial likelihood*, que únicamente dependerá del parámetro de interés β. Dicha función se define a continuación (Wang et al., 2019):

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{\exp(X_j \beta)}{\sum_{i \in R_j} \exp(X_i \beta)} \right]^{\delta_j} \qquad (16)$$

Donde $j = 1, \dots, N$. Cuando $\delta_j = 0$, es decir, cuando se produce censura, el término no tendrá ningún efecto sobre el ajuste final. Finalmente, el vector de coeficientes $\widehat{\boldsymbol{\beta}}$ se obtiene con la minimización del *log partial likelihood* negativa, tal que (Wang et al., 2019):

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \left[X_j \beta - log(\sum_{i} \exp(X_j \beta)) \right] \qquad (17)$$

Que es equivalente a la maximización de la verosimilitud parcial, cuyo estimador resultante recibe el nombre de MPLE (*Maximum Partial Likelihood – Estimator)*.

Una vez se hayan estimado los coeficientes, se emplea la prueba de Wald o la prueba de la ratio de verosimilitudes para la inferencia estadística sobre dichos estimadores.

2. Objetivos

El **objetivo general** del presente trabajo es la comparación, en términos de poder predictivo, detección de factores pronóstico y coste computacional de modelos novedosos en análisis de supervivencia para datos multiómicos. Con la finalidad de llevarlo a cabo, se han definido los siguientes **objetivos específicos**:

- Recopilar modelos basados multivariantes o de aprendizaje automático para el análisis de supervivencia con bases de datos de alta dimensionalidad.
- Aplicar dichos modelos y sus versiones multi – bloque, cuando las haya, a dos conjuntos de datos multiómicos.
- Evaluar los modelos seleccionados y comparar el poder predictivo mediante medidas de rendimiento estándar en el análisis de supervivencia.
- Valorar el poder predictivo que aportan las distintas ómicas, así como los modelos multi – bloque para la predicción de la supervivencia.
- Identificar posibles factores pronóstico y contrastar su papel en la patología de estudio, en este caso, cáncer de mama invasivo y cáncer de cuello y cabeza, ya sean factores transcriptómicos, genómicos, proteómicos o epigenómicos.
- Aplicar los estimadores Kaplan Meier sobre los factores identificados para la comparación e interpretación de curvas de supervivencia.
- Obtener el coste computacional de cada modelo, incluyendo la fase de entrenamiento y optimización de los hiperparámetros.

3. Materiales y métodos

3.1 Bases de datos: *The Cancer Genome Atlas*

Las bases de datos seleccionadas para llevar a cabo los objetivos establecidos anteriormente proceden del proyecto *The Cancer Genome Atlas* (TCGA), introducido por los Institutos Nacionales de la Salud (NIH), con el fin de crear un atlas comprehensivo de perfiles ómicos del cáncer, facilitando así el estudio de las complejas relaciones en un sistemas biológico (Tomczak et al., 2015).

Este proyecto abarca más de 30 tipos diferentes de tumores, con sus respectivas bases de datos, las cuales contienen información sobre los cambios dinámicos en el genoma debido al cáncer (Hanahan & Weinberg, 2000), otras medidas ómicas, e información sobre aspectos clínicos de cada uno de los pacientes (Tomczak et al., 2015).

Las bases de datos seleccionadas para este estudio incluyen, además de los bloques de variables mencionados, el tiempo de supervivencia de cada individuo y su correspondiente indicador de censura.

A continuación, se detallarán las características de dichas bases de datos.

3.1.1 Breast Invasive Carcinoma

En primer lugar, se ha seleccionado la base de datos de *Breast Invasive Carcinoma* (BRCA), que destaca por el gran número de estudios dedicados a la investigación del cáncer de mama invasivo y sus dinámicas. Entre estos, no sólo sobresalen los enfocados en la supervivencia de los pacientes, sino también aquellos centrados en imagen biomédica para la clasificación de tumores (Bebis et al., n.d.) o en determinar perfiles genéticos según los resultados histológicos (Ciriello et al., 2015).

Un resumen de los bloques y sus respectivas variables se muestran en la siguiente tabla (Tabla 1):

Tabla 1: Descripción de la base de datos BRCA. En columnas, de izquierda a derecha, se enuncian los bloques que la forman, sus correspondientes dimensiones (filas x columnas) y su descripción.

Bloque	Dimensiones	Descripción
X.Clinical	571 x 13	➢ 'ajcc_pathologic_stage': categórica ordinal. Estadio general del cáncer (Stage I – Stage III). ➢ 'ajcc_pathologic_t': categórica ordinal. Estadio del tumor (T1 – T4). ➢ 'ajcc_pathologic_n': categórica ordinal. Estadio de los nódulos linfáticos (N0 – N3). ➢ 'ajcc_pathologic_m': categórica ordinal. Estadio de la metástasis (M0 – MX). ➢ 'prior_malignancy': categórica. Primer diagnóstico sobre la malignidad del cáncer (Yes / No). ➢ 'paper_BRCA_Pathology': categórica. Diagnóstico histológico (IDC, ILC, Mixed, NA, Other). ➢ 'paper_BRCA_Subtype_PAM50': categórica. Estado de receptor (Basal, Her2, LumA, LumB, Normal).

		➢ 'treatment_type': categórica, Tipo de tratamiento recibido (Radiation Therapy, Pharmaceutical Therapy). ➢ 'age_at_diagnosis': numérica. Edad de diagnóstico de cáncer del individuo.
X.Genes	571 x 19318	Variables numéricas. Cuantificación de la expresión génica.
X.miRNA	571 x 642	Variables numéricas. Cuantificación de la expresión de miRNA.
X.Methyl	571 x 17436	Variables numéricas. Cuantificación de la metilación del DNA.

Con un total de 37409 variables y 571 pacientes.

Esta base de datos descrita anteriormente proviene de un procesado previo mediante diversas transformaciones. Éstas se enumeran a continuación:

- Normalización *Trimmed Mean of M-values* (TMM) de *X.miRNA* y transformación logarítmica.
- Normalización de *X.Genes* con *Conditional Quantile Normalization* (CQN).
- *X.Methyl* agrupado a nivel de gen mediante sus asociaciones y media de valores.

Cabe añadir que se eliminaron pacientes con las siguientes características:

- En presencia de valores faltantes.
- Tiempos de supervivencia nulos o negativos.
- Eliminación de muestras pertenecientes a hombres (consecuentemente, eliminación de la variable 'gender').

Además de la variable 'age_at_index' al estar correlacionada con la variable 'age_at_diagnosis'.

Todas las variables categóricas fueron transformadas a *dummies*, obteniendo de estas 'k-1' nuevas variables o categorías, siendo 'k' los posibles valores de cada una.

3.1.2 Head and Neck Squamous Cell Carcinoma

Por otra parte, la base de datos *Head and Neck Squamous Cell Carcinoma* (HNSC) proviene de un estudio de cohorte que permitió el registro de información sobre tumores situados en la cavidades oral, laríngea y orofaringe (Lawrence et al., 2015). Está enfocado al análisis del impacto del consumo usual de tabaco en la aparición de los tumores mencionados, incluyendo en el estudio individuos no fumadores que padecen del virus del papiloma humano (HPV) (Lawrence et al., 2015).

De nuevo, un resumen de los bloques y sus respectivas variables se muestran en la siguiente tabla (Tabla 2):

Tabla 2: Descripción de la base de datos HNSC. En columnas, de izquierda a derecha, se enuncian los bloques que la forman, sus correspondientes dimensiones (filas x columnas) y su descripción.

Bloque	Dimensiones	Descripción
X.Clinical	443 x 7	➢ 'tobacco_smoking_history': categórica ordinal. Número de paquetes de tabaco diaríos (1 – 5). ➢ 'gender': categórica. Género del paciente (Male, Female) ➢ 'alcohol_history_documented': categórica. Ingesta o no de alcohol (Yes, No). ➢ 'lymphnode_neck_dissection': categórica. Realización o no de cirugía para retiro de nódulos linfáticos (Yes, No). ➢ 'stage_event': categórica ordinal. Estadio general del cáncer (Stage I – Stage IV). ➢ 'neoplasm_histologic_grade': categórica ordinal. Grado de anormalidad de células cancerígenas (G0 – G3). ➢ 'age': numérica. Edad del individuo.
X.Genes	443 x 21520	Variables numéricas. Cuantificación de la expresión génica.
X.miRNA	443 x 793	Variables numéricas. Cuantificación de la expresión de miRNA.
X.Mutation	443 x 116	Variables categóricas binarias. Indica mutación genética ('1') o no ('0') para cada gen.
X.cnv	443 x 57964	Variables numéricas. Cuantificación de las variaciones en el número de copias del ADN.

Con un total de 80400 variables y 443 muestras o individuos, superando con creces la base de datos de cáncer de mama, en cuanto a número de variables.

Debe tenerse en cuenta que se llevó a cabo un filtrado del bloque *X.Mutation* dado el alto porcentaje de variables con una frecuencia por debajo del 1% en la presencia de mutación, pues su implementación en los modelos no supondría mejora alguna y se descartan como posibles factores pronóstico. Además, no se encontraron muestras con datos faltantes o tiempos de supervivencia nulos o negativos.

Todas las variables categóricas fueron transformadas a *dummies*, obteniendo de estas 'k-1' nuevas variables o categorías, siendo 'k' los posibles valores de cada una.

3.2 Metodología

3.2.1 Modelos seleccionados

3.2.1.1 Basados en regresión penalizada

Tal y como se ha comentado anteriormente, la estadística tradicional presenta problemas en la regresión de estimadores para bases de datos de alta dimensionalidad. Esto ha llevado a un desarrollo de modelos de regresión, basados en la penalización de los coeficientes de regresión, que optimiza el poder predictivo del modelo mediante la selección de aquel subconjunto de variables explicativas de una mayor importancia (Kyung et al., 2010).

A continuación, se introducirán las diversas modalidades de la regresión penalizada adaptada al modelo Cox para datos de carácter multiómico.

3.2.1.1.1 Regresión Elastic Net

La regresión *Least Absolute Shrinkage and Selection Operator* (LASSO) (Tibshiranit, 1996) en su modalidad Cox (Tibshirani, 1997) puede definirse mediante la estimación del vector de coeficientes $\boldsymbol{\beta} = \{\beta_j\}$, siendo 'j' cada una de las variables explicativas estandarizadas, tal que $j = 1, \dots, p$. Dicha estimación trata de minimizar la función de pérdidas siguiente:

$$-\rho\ell(\beta) + \lambda \cdot \sum_{j=1}^{p} |\beta_j| \qquad (18)$$

Donde $\lambda \geq 0$ es el factor de penalización y $\|\cdot\|_1 = \sum_{j=1}^{p} |\beta_j|$ la norma L_1, ambos términos correspondientes a la penalización LASSO.

El término $-\rho\ell(\cdot)$ se refiere a la *partial log – likelihood* del modelo de Cox:

$$-\rho\ell(\beta) = -\sum_{i=1}^{N} \delta_i \left[\beta_j x_{ij} - log_e\left(\sum_{i} e^{\beta_j x_{ij}}\right)\right] \qquad (19)$$

Donde, para $i = 1, \dots, n$ muestras, se tiene el indicador de censura δ_i y las variables explicativas estandarizadas $\boldsymbol{x_{ij}}$.

Este modelo, mediante las penalizaciones introducidas en la función de pérdidas, es por tanto capaz de reducir los estimadores mínimo cuadráticos a cero, siendo adicionalmente un método comúnmente usado para la selección de variables (Li & Sillanpää, 2012).

Otro enfoque se emplea en el caso de la regresión Ridge, en la que se emplea una penalización de los coeficientes basada en la norma L_2, tal que $\|\cdot\|_2 = \sum_{j=1}^{p} {\beta_j}^2$ (Li & Sillanpää, 2012).

Comparando ambos métodos, el modelo LASSO puede presentar algunas desventajas (Li & Sillanpää, 2012): (1) en casos de multicolinealidad entre variables explicativas, la regresión LASSO tiende a seleccionar una única variable dentro del conjunto de variables de alta correlación. (2) Cuando se

implementa en bases de datos de elevada dimensionalidad ($p > n$), se pueden seleccionar hasta n variables explicativas para el modelo final.

Una forma de superar estas limitaciones se introduce mediante la combinación de ambas penalizaciones, llevada a cabo por *Elastic Net* (Zou & Hastie, 2005). La función de penalización adoptada por este método es la siguiente:

$$\lambda[(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}{\beta_j}^2 + \alpha\sum_{j=1}^{p}|\beta_j|] \qquad (20)$$

Donde:

$$\begin{cases} \alpha = 0 & : Ridge\ Regression \\ 0 < \alpha < 1 & : Elastic\ Net \\ \alpha = 1 & : LASSO \end{cases} \qquad (21)$$

De esta forma, se mantienen las propiedades de selección de variables de LASSO sin descartar variables correlacionadas, seleccionándolas todas y asignando normalmente un mismo coeficiente de regresión a variables con alta correlación (Li & Sillanpää, 2012).

Dadas las características de los modelos mencionados, se ha decidido implementar en el proyecto la modelización del modelo LASSO estándar ($\alpha = 1$) al igual que un modelo basado en la premisa de *Elastic Net*, en el que se les ha dado el mismo peso a las dos penalizaciones ($\alpha = 0.5$).

3.2.1.1.2 IPF LASSO

Tal y como se ha mencionado en la introducción del trabajo, existen modelos enfocados al análisis de la supervivencia que tratan de implementar la estructura por bloques de las bases de datos de alta dimensionalidad.

Uno de estos modelos es el *Integrative LASSO with Penalty Factors* (IPF – LASSO) (Boulesteix et al., 2017), basado en la penalización LASSO ya explicada.

Definiendo en primer lugar el número bloques M, se tienen los distintos conjuntos de variables explicativas $\boldsymbol{X_1^{(m)}}, \dots, \boldsymbol{X_{p_m}^{(m)}}$, donde p_m es el número de variables en el bloque, y $m = 1, \dots, M$. Se propone la estimación de los M vectores de coeficientes $\boldsymbol{\beta^{(m)}} = (\beta_1^{(m)}, \dots, \beta_{p_m}^{(m)})$ mediante la minimización de la siguiente función de pérdidas (Boulesteix et al., 2017):

$$-\rho\ell(\beta^{(m)}) + \sum_{m=1}^{M}\lambda_m \cdot \left\|\beta^{(m)}\right\|_1 \qquad (22)$$

En la que el término $-\rho\ell(\cdot)$ se refiere a la *partial log – likelihood* del modelo de Cox, $\|\cdot\|_1$ a la norma L_1 y λ_m al término de penalización para cada bloque de variables.

De esta forma, cuanto mayor sea el factor de penalización asociado a cada conjunto de coeficientes de cada bloque m, mayor penalización tendrá en la selección de sus variables en comparación al resto de bloques.

3.2.1.1.3 Priority LASSO

Finalmente, dentro de la modelización de tipo LASSO, se empleará el método *Priority* LASSO (Klau et al., 2018), que al igual que IPF LASSO, implementa la regresión Cox teniendo en cuenta la estructura en bloques de los datos.

Se define el vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$ como el orden de prioridad a la hora de implementar cada uno de los bloques de variables en el modelo: π_1 denota el índice del bloque con mayor prioridad, mientras que π_M el índice del bloque de menor prioridad. Si, por ejemplo, se tienen 4 bloques de variables ($M = 4$) y un vector $\boldsymbol{\pi} = (3, 1, 4, 2)$, el bloque de mayor prioridad será el tercero, y el de menor prioridad, el segundo (Klau et al., 2018).

En la primera etapa, se ajustan las variables asociadas al bloque definido de mayor prioridad π_1. Se estiman por tanto los coeficientes $\beta_1^{(\pi_1)}, \dots, \beta_{p_{\pi_1}}^{(\pi_1)}$ por medio de la minimización de la función de pérdidas:

$$-\rho\ell(\beta^{(\pi_1)}) + \lambda^{(\pi_1)} \cdot \sum_{j=1}^{p_{\pi_1}} \left|\beta_j{}^{(\pi_1)}\right| \qquad (23)$$

Y se obtiene la predicción lineal para cada muestra ‘i’ de este bloque, tal que:

$$\hat{\eta}_{1,i}(\pi) = \hat{\beta}_1^{(\pi_1)} x_{i1}^{(\pi_1)} + \cdots + \hat{\beta}_{p_{\pi_1}}^{(\pi_1)} x_{ip_{\pi_1}}^{(\pi_1)} \qquad (24)$$

De forma iterativa, dicha predicción lineal se utiliza como un *offset* en al ajuste de los coeficientes asociados al resto de bloques, eliminando en cada iteración la influencia de cada uno de los bloques de variables sobre la supervivencia o predicción de cada uno de los individuos (Klau et al., 2018).

Tras la estimación de los coeficientes asociados a cada uno de los bloques, se obtienen las predicciones lineales finales siguientes:

$$\hat{\eta}_{M,i}(\pi) = \sum_{m=1}^{M} \sum_{j=1}^{p_{\pi_m}} \hat{\beta}_j^{(\pi_m)} x_{ij}^{(\pi_m)} \qquad (25)$$

3.2.1.2 Basados en técnicas de *boosting*

El *boosting* es una técnica iterativa basada en la estimación de parámetros, en este caso, los coeficientes del modelo de Cox, mediante la actualización de sus valores paso a paso. En cada iteración, se ajusta un estimador "débil" mediante la minimización de una función de pérdidas preestablecida. La estimación final proviene de las contribuciones obtenidas en cada una de las iteraciones (De Bin, 2016).

3.2.1.2.1 Model Based Boosting

El algoritmo de la técnica *Model Based Boosting* (Hothorn & Bühlmann, 2006) es una implementación directa del algoritmo de gradiente descendente. En el caso de la regresión Cox, la estimación de los coeficientes se realiza conforme al siguiente algoritmo:

1. Inicialización $\widehat{\boldsymbol{\beta}} = (0, \dots, 0)$;
2. Cómputo del vector de gradiente negativo u para cada muestra 'i':

$$u^{(i)} = \delta^{(i)} - \sum_{l \epsilon R^{(i)}} \delta^{(l)} \frac{\exp\left\{X^{(l)^T} \hat{\beta}\right\}}{\sum_{k \epsilon R^{(l)}} \exp\left\{X^{(k)^T} \hat{\beta}\right\}} \qquad (26)$$

3. Cómputo de las posibles contribuciones aplicando el estimador mínimo cuadrático al vector de gradiente negativo:

$$\hat{b}_j = (X_j^T X_j)^{-1} X_j^T u \qquad (27)$$

4. Selección de la contribución óptima j^* minimizando la suma cuadrática de residuos:

$$j^* = argmin_j \sum_{i=1}^{n} (u^{(i)} - X_j^{(i)} \hat{b}_j)^2 \qquad (28)$$

5. Actualización del estimador, penalizándolo con el factor ν ($\nu \in [0,1]$), tal que:

$$\hat{\beta}_{j^*} = \hat{\beta}_{j^*} + \nu \hat{b}_{j^*} \qquad (29)$$

Donde $\widehat{\boldsymbol{\beta}}$ es el vector de coeficientes, $\boldsymbol{X_j}$ el vector columna que contiene los valores de cada variable 'j', δ el indicador de censura ($\delta^{(i)} = 0$ indica censura) y $\boldsymbol{R^{(i)}}$ el conjunto de observaciones en un tiempo de supervivencia dado $t^{(i)}$.

Los pasos de 2 a 5 se repetirán m_{stop} veces, siendo m_{stop} el número de iteraciones realizadas.

Se puede comprobar cómo se ajustan múltiples regresiones lineales univariantes del vector u para cada X_j. Además, dicho método no incluye la estructura en bloques de las variables explicativas.

3.2.1.2.2 Likelihood Based Boosting

En segundo lugar, el algoritmo *Likelihood Based Boosting* (Harald Binder & Harald Binder, 2015), que implementa una función de pérdidas que incluye la norma L_2:

$$pl_{pen}(\beta) = pl(\beta) - 0.5\lambda\beta^T P\beta \qquad (30)$$

Donde $\boldsymbol{P}$ es una matriz $p \times p$ correspondiente a la matriz identidad, siendo p el número de variables explicativas, y λ el factor de penalización. El término $pl(\beta)$ se corresponde a la función *partial log – likelihood*:

$$pl_{pen}^{LB}(\beta|\hat{\beta}) = \sum_{i=1} \delta^{(i)}[\hat{\eta}^{(i)} + (X^{(i)^T}\beta) - \log(\sum_{l \epsilon R^{(i)}} \exp\{\hat{\eta}^{(l)} + X^{(l)^T}\beta\})] - 0.5\lambda\beta^T P\beta \qquad (31)$$

Donde $\hat{\eta}^{(i)}$ es la predicción lineal para cada muestra 'i'.

En cada iteración, las funciones de *partial log – likelihood* restringidas se desplazan hacia $\hat{\beta}_j$, obteniendo el valor de la función $pl_{pen}^{LB}(\beta_j|\hat{\beta})$. Los argumentos de los máximos de estas funciones serán las contribuciones candidatas, y aquella que maximice dicha función de pérdidas será añadida al término de *offset*. El algoritmo, de forma más detallada, es el siguiente (De Bin, 2016):

1. Inicialización $\widehat{\boldsymbol{\beta}} = (0, \ldots, 0)$;
2-3. Cómputo de las posibles contribuciones mediante el MPLE:

$$\hat{b}_j^{LB} = \frac{pl_{\beta_j}^{LB}(0|\hat{\beta})}{-pl_{\beta_j\beta_j}^{LB}(0|\hat{\beta})} \qquad (32)$$

4. Selección de la contribución óptima:

$$j^* = argmin_j \, pl_{\beta_j}^{LB}(0|\hat{\beta})^2 / \left[-pl_{\beta_j\beta_j}^{LB}(0|\hat{\beta})\right] \qquad (33)$$

5. Actualización del estimador:

$$\hat{\beta}_{j^*} = \hat{\beta}_{j^*} + \hat{b}_{j^*}^{LB} \qquad (34)$$

Donde $pl_{\beta_j}^{LB}(\beta_j|\hat{\beta})$ denota el *score* y $pl_{\beta_j\beta_j}^{LB}(\beta_j|\hat{\beta})$ la información observada. La ecuación (33) de la cuarta etapa se ha implementado en el algoritmo por cuestiones computacionales (De Bin, 2016).

Los pasos de 2 a 5 se repetirán m_{stop} veces, siendo m_{stop} el número de iteraciones realizadas.

Al igual que la técnica *Model Based Boosting*, este método tampoco incluye la estructura en bloques de las variables explicativas en la obtención del conjunto de coeficientes.

3.2.1.3 Basados en Random Forest

El algoritmo de *random forest* (Breiman, 2001) ha demostrado ser un potente método de predicción, capaz de capturar patrones complejos entre las variables explicativas y su *output*. Su adaptación al análisis de supervivencia ha demostrado ser prometedor en bases de datos multiómicas o de alta dimensionalidad.

3.2.1.3.1 Random Survival Forest

Una versión de este método en el *random survival forest* (RFSC) (Ishwaran et al., 2008), adaptación directa del algoritmo original que no implementa la estructura en bloques de la base de datos.

La descripción del algoritmo RFSC se muestra a continuación (Ishwaran et al., 2008):

- Paso 0: Fijado del número de árboles (*ntree*) y número de variables candidatas a elegir en cada nodo (*mtry*).
- Paso 1: Tomar *ntree* muestras de *bootstrap* con reemplazamiento del conjunto de entrenamiento.
- Paso 2: Para cada muestra de *bootstrap*, se lleva a cabo el siguiente procedimiento:
  - Tomar *mtry* variables explicativas.
  - Cálculo del estadístico *logrank* para cada posible punto de división de cada variables candidata *mtry*; selección la combinación de variables que maximicen el estadístico *logrank*; división de las observaciones en dos nodos hijos.
  - Crecimiento de cada árbol siempre que el número de observaciones no censuradas en cada nodo es mayor que un mínimo preestablecido, denominado *nodesize*.
- Paso 3: Dada una nueva observación, se introduce en cada uno de los *ntree* hasta sus nodos terminales. Se obtiene una predicción promedio usando el estimador CHF (*Cumulative Hazard Function*) en cada nodo terminal y se promedian los resultados de todos los *ntrees* y puntos en el tiempo para obtener un único *score*.
- Paso 4: Para un conjunto de test, se evalúa el poder predictivo obteniendo el *score* del Paso 3 para cada observación. El *score* seleccionado en este caso es el *cindex*, que será explicado en la sección de Estrategia de validación.

Se puede observar el uso del estadístico *logrank* para la división de muestras en cada nodo de cada uno de los árboles entrenados. Este estadístico facilita la selección de aquellas variables entre las *mtry* candidatas que lo maximizan, lo cual lleva a la maximización de las diferencias entre curvas de supervivencia entre los nodos hijos.

El error de predicción usado para la optimización de los hiperparámetros se define como $1 - C$, donde C hace referencia al *cindex*. Este se calcula a partir de las observaciones no seleccionadas en el Paso 1, que reciben la notación de *out of bag data* (OOB).

3.2.1.3.2 Block Forest

El algoritmo *Block Forest* (Hornung & Wright, 2019a) consiste en una modificación del RFSC para la implementación de la estructura de bloques, como la de los datos multiómicos. Más concretamente, define una serie de estrategias de selección de *mtry* variables candidatas en el Paso 2 de dicho algoritmo con priorización de bloques en base a distintos criterios.

Esto permite solucionar una desventaja del algoritmo RFSC, que es la poca representación que pueden tener bloques con un número de variables muy reducido, como puede ser el conjunto de variables clínicas en comparación a las ómicas. De esta forma, mediante esta modificación, puede implementarse el enfoque de que bloques con un menor número de variables normalmente contienen una mayor densidad de información para la predicción de la supervivencia.

Las estrategias definidas en este método se enumeran a continuación.

- *VarProb*: Define una probabilidad v_m de muestreo de cada bloque de variables *m*, añadiendo la restricción $\sum_{m=1}^{M} p_m \, v_m$, donde p_m es el número de variables en cada bloque *m* y *M* el número total de bloques. Se define el valor de *mtry* como $\sum_{m=1}^{M} \sqrt{p_m}$.
- *SplitWeights*: Utiliza pesos específicos w_m para cada bloque *m*, tal que $\max\{w_1, \dots, w_M\} = 1$. Tras extraer $mtry = \sum_{m=1}^{M} \sqrt{p_m}$ número de variables candidatas de todo el conjunto de variables, y calcular el criterio de división (*logrank*) asociado a cada punto de división, se ponderan estos valores utilizando dichos pesos específicos. Posteriormente, se elige el punto de división óptimo. Así, las variables de bloques con un valores altos de w_m tienen prioridad sobre las variables de bloques con valores bajos.
- *BlockVarSel*: Se define una selección de variables candidatas fija para cada bloque, forzando la representación de todos los bloques en el modelo, tal que $mtry_m = \sqrt{p_m}$. Tras esta selección, se prosigue con el enfoque de *SplitWeights* mediante pesos específicos para elección del punto de división óptimo.
- *RandomBlock*: En primer lugar, se elige un bloque de acuerdo con probabilidades de selección b_m, donde $\sum_{m=1}^{M} b_m = 1$. Posteriormente, se selecciona un conjunto de variables candidatas $\sqrt{p_m}$ de dicho bloque. El cómputo del estadístico de división en dicho nodo se hace de acuerdo con el algoritmo estándar de RFSC.
- *BlockForest*: Primero, se obtiene un *subset* de los *M* bloques, con una probabilidad de 0.5 de seleccionar cada uno de estos. En caso de no haber escogido ninguno, se vuelve a realizar la selección. Finalmente, se lleva a cabo el procedimiento explicado para *BlockVarSel*.

Dado que éste último enfoque es el que mejor resultados ha proporcionado en análisis de supervivencia para bases de datos de alta dimensionalidad (Hornung & Wright, 2019a), se ha decidido usar esta versión del modelo en el presente trabajo. Además, esta estrategia de selección de variables asegura una representación de todos los bloques de variables, haciendo al *Block Forest* comparable con el resto de los modelos que incluyen la estructura en bloques de la base de datos.

3.2.1.4 Modelo de Riesgos Proporcionales de Cox

El modelo Cox PH (*Proportional Hazards*), cuya base estadística está detallada en la sección Análisis de Supervivencia: Modelización Cox, es la aproximación clásica al análisis de datos en estudios de supervivencia.

Por esto último, se ha decidido su implementación en el presente proyecto como un modelo de referencia, comparándolo así con el resto de modelos. Dado que este modelo es incompatible con bases de datos de alta dimensionalidad (con más variables que observaciones) y con la consecuente multicolinealidad, únicamente se ajustará con las variables explicativas pertenecientes al bloque *X.Clinical* de cada base de datos, siguiendo el enfoque de que contiene una mayor cantidad de información sobre la supervivencia de los individuos para un menor número de variables.

Así, además, podrá relativizarse la incorporación de variables de carácter ómico y su impacto sobre el rendimiento para la predicción de la supervivencia.

3.2.2 Evaluación de los modelos

3.2.2.1 Estrategia de validación

A continuación, se especificarán los detalles más importantes en cuanto a configuración de cada tipología de modelo.

- *Standard LASSO:* La optimización del factor de penalización λ se ha realizado de forma interna por medio de una validación cruzada de 10 *k - fold*, sin repeticiones, maximizando el *cindex*. Este modelo no implementa la estructura en bloques de la base de datos.

- *Elastic Net:* La optimización del factor de penalización λ se ha realizado de forma interna por medio de una validación cruzada de 10 *k - fold*, sin repeticiones, maximizando el *cindex*. Este modelo no implementa la estructura en bloques de la base de datos. El parámetro de *Elastic Net* α se ha fijado en 0.5, dándole el mismo peso a la regresión LASSO y a la regresión *Ridge.*

- *IPF LASSO:* La optimización de los factores de penalización λ_m no se incluye en el propio flujo de trabajo de la función disponible. Se lleva a cabo la optimización de un único factor de penalización λ mediante validación cruzada de 10 *k - fold* sin repeticiones, y mediante el argumento *'pf'* se atribuye mayor o menor penalización a cada bloque. Se utiliza la función de pérdidas de *partial likelihood* para dicha optimización. Se ha considerado que este enfoque dificulta una comparación justa entre modelos, por lo que se ha utilizado el conjunto de entrenamiento con un escalado de tipo pareto, asignándole así una mayor penalización, o menor variabilidad, a aquellos bloques con un mayor número de variables, y manteniendo por tanto el argumento *'pf'* sin implementar.

- *Priority LASSO*: La optimización del factor de penalización λ se ha realizado de forma interna por medio de una validación cruzada de 10 *k - fold*, sin repeticiones, empleando la función de pérdidas de *partial likelihood*. La penalización se introduce modificando el orden de los bloques e introduciendo los índices de variables, asignados en función de su pertenencia a cada bloque. Con el fin de facilitar la comparabilidad de los modelos, el orden de bloques se ha introducido de menor a mayor número de variables, asignando una mayor penalización a estos últimos. El orden para cada base de datos se muestra a continuación:

  - BRCA: $X.Clinical \rightarrow X.miRNA \rightarrow X.Methyl \rightarrow X.Genes$

  - HNSC: $X.Clinical \rightarrow X.Mutation \rightarrow X.miRNA \rightarrow X.Genes \rightarrow X.cnv$

- *Random Survival Forest*: La optimización de los parámetros *mtry*, *nodesize* y *ntrees* se ha realizado de forma externa, sin repeticiones. Posteriormente, se ha entrenado el modelo con hiperparámetros optimizados y se ha extraído la importancia asociada a cada variable. Este modelo no implementa la estructura en bloques de la base de datos.

- *Block Forest*: Se ha hecho uso de la optimización de *ntrees* realizada en el modelo de *Random Survival Forest*. La optimización de *nodesize*, *mtry* y de los pesos $w_1 \dots w_M$ asociados a la variante *BlockForest* se lleva a cabo de forma interna. Se ha incluido la estrategia de selección de variables recomendada en cada nodo de *BlockForest*, la cual además mantiene el enfoque mencionado para asegurar la comparabilidad de los modelos que implementan la estructura en bloques de la base de datos. Los conjuntos de entrenamiento y test usados no han sido escalados ni centrados.

- *CoxBoost*: La optimización del factor de penalización se ha realizado de forma externa con una validación cruzada de 10 *k - fold*, empleando la función de pérdidas de *penalized partial likelihood*, con un número de iteraciones fijo de 150 de acuerdo con las recomendaciones de la literatura (Herrmann et al., 2021a). Posteriormente, se ha entrenado el modelo con hiperparámetros óptimos. Este modelo no implementa la estructura en bloques de la base de datos.

- *MBoost*: La optimización del factor de penalización se ha realizado de forma externa, utilizando la función de pérdidas de *partial likelihood*, con un número de iteraciones fijo de 150 de acuerdo con las recomendaciones de la literatura (Herrmann et al., 2021a). Posteriormente, se ha entrenado el modelo con hiperparámetros óptimos. Este modelo no implementa la estructura en bloques de la base de datos.

El entrenamiento y optimización de los modelos se ha llevado a cabo con un conjunto de entrenamiento formado por el 70% de las muestras, previamente centrado y escalado. El 30% restante para ambas bases de datos se ha empleado para la obtención de las medidas de rendimiento con las respectivas predicciones. Este conjunto se ha centrado y escalado con las medias y desviaciones típicas empleadas para el conjunto de entrenamiento. Los conjuntos de entrenamiento y test usados no han sido escalados ni centrados en los modelos *Random Survival Forest* y *Block Forest*.

Se han empleado dos tipos de escalados en los modelos, el estándar y el pareto, en los que se ha concatenado las distintas matrices ómicas por no disponer de opción multibloque. Los pesos aplicados a cada una de las variables en dichos escalados se muestran a continuación:

$$Estándar: \quad \frac{1}{S_k} \qquad (35)$$

$$Pareto: \quad \frac{1}{S_k \cdot (m_b)^{\frac{1}{4}}} \qquad (36)$$

Donde S_k es la desviación típica de la variable k y m_b el número de variables del bloque b. El escalado estándar proporciona el mismo peso a todas las variables, y, por tanto, a mayor el número de variables en cada bloque, mayor representación tendrá en el modelo. Por otra parte, el escalado de tipo pareto asigna un mayor peso a aquellas variables pertenecientes al bloque *b* de menor dimensionalidad.

Una tabla resumen de los modelos entrenados y sus principales características en su validación y evaluación se puede ver a continuación (Tabla 3).

Tabla 3: Descripción de la validación en cada modelo. En cada columna, de izquierda a derecha, se enuncia el modelo, si considera o no la estructura de bloque de la base de datos, la estrategia de optimización y cómo se evalúa la selección de variables en cada modelo. El sufijo 'Par' indica un escalado de tipo pareto sobre los conjuntos de datos empleados.

Modelo	Bloque	Optimización	Selección de variables
Elastic Net (α=0.5)	No	*k – fold* CV (k = 10)	Coeficientes ajustados
Elastic Net (α=0.5) Par	Sí		
Standard LASSO	No		
Standard LASSO Par	Sí		
IPF LASSO	Sí		
Priority LASSO	Sí		
CoxBoost	No	*k – fold* CV (k = 10)	Coeficientes ajustados
CoxBoost Par	Si		
MBoost	No		
MBoost Par	Si		
Random Survival Forest	No	OOB	Importancia de variables
Block Forest	Sí	OOB	
Cox PH (referencia)	No	-	-

Como se puede observar, el escalado de tipo pareto ha permitido comparar los modelos en los que únicamente se ha realizado una concatenación de variables con los que sí utilizan información por bloques. Esto se debe a que asigna un mayor peso, o mayor variabilidad, a aquellos bloques con un menor número de variables, tal y como se ha llevado a la práctica con los modelos *IPF LASSO*, *Priority LASSO* y *Block Forest*.

Este punto de vista se adopta comúnmente en el contexto del análisis de la supervivencia en bases de datos de alta dimensionalidad, figurando como una técnica de preprocesado típica en estos contextos (De Bin et al., 2014).

3.2.2.2 Medidas de rendimiento

La evaluación de los modelos se llevará a cabo de acuerdo con tres dimensiones: rendimiento en la predicción, selección de factores pronósticos y tiempo de computación.

El rendimiento en la predicción se evaluará por medio de dos medidas usadas comúnmente en análisis de supervivencia (Herrmann et al., 2021b): el *Brier score* integrado (Graf et al., 1999) y el *cindex* (Harrell et al., 1996).

En primer lugar, el *Brier score* integrado es usado como parámetro de calibración, con el fin de cuantificar la habilidad de cada modelo de realizar predicciones lo más cercanas posibles a la ocurrencia real de los eventos (Herrmann et al., 2021b) (Khene et al., 2023).

Esta medida se basa en el uso directo de las probabilidades estimadas de supervivencia $\hat{\pi}(t^*|X)$, dado un conjunto de p variables explicativas $X = (X^1, \dots, X^p)$ de cada paciente, comparándolas con su *status*

real de evento $Y = I(T > t^*)$, donde T es una variable, mayor que cero, que representa el tiempo desde un punto de inicio $t = 0$ hasta el evento de interés. La función $I(T > t^*) \in \{0,1\}$ representa por tanto el *status* del evento tras un tiempo dado t^*, tomando como valor 0 o 1 en función de si se produce o no el evento, respectivamente (Graf et al., 1999). En el caso de las bases de datos seleccionadas, dicho evento es la muerte del sujeto.

Una primera expresión de este parámetro se muestra a continuación:

$$\mathrm{BS}(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t^*) - \hat{\pi}(t^*|X)\right)^2 \qquad (37)$$

Obteniéndose una estimación del *Brier score* para cada instante definido t^*.

Dado que se está tratando con datos censurados, es necesario incorporar este cambio a la anterior expresión. Para cada paciente, se observa un tiempo $\tilde{T}_i = \min(T_i, C_i)$ y el indicador de censura $\delta_i = I(T_i \leq C_i)$, donde C_i es el tiempo hipotético bajo observación (o de censura) de cada paciente. Se asume una distribución $G(t) = P(C > t)$, de forma análoga a $S(t) = P(T > t)$, que son las probabilidades marginales de que no se produzca un evento o de supervivencia hasta el tiempo t (Graf et al., 1999).

La censura se introduce mediante distintas contribuciones, que se dividen en tres categorías:

$$Categoría\ 1:\ \tilde{T}_i \leq t^*\ y\ \delta_i = 1\ con\ contribución\ (0 - \hat{\pi}(t^*|X))^2 \qquad (38)$$

$$Categoría\ 2:\ \tilde{T}_i > t^*\ y\ \delta_i = 1\ o\ \delta_i = 0\ \ con\ contribución\ (1 - \hat{\pi}(t^*|X))^2 \qquad (39)$$

$$Categoría\ 3:\ \tilde{T}_i \leq t^*\ y\ \delta_i = 0\ \mathrm{sin}\ contribución \qquad (40)$$

Para las muestras sin censura (categoría 1) el evento sucede antes de t^*, y el *status* del evento en t^* equivale a $I(T_i > t^*) = 0$. En cambio, en la categoría 2, el *status* del evento observado en t^* equivale a $I(T_i > t^*) = 1$. Finalmente, dado que la censura sucede antes de t^*, el *status* del evento es desconocido para t^*, por lo que su contribución no puede calcularse (Graf et al., 1999).

Por esto último, para el cálculo del *Brier score* en datos censurados, es importante añadir un peso distinto para cada una de las contribuciones mencionadas: $1/\hat{G}(\tilde{T}_i)$ en la categoría 1, $1/\hat{G}(t^*)$ en la categoría 2 y cero para la categoría 3. Divididos entre n, estos pesos suman 1, y el nuevo parámetro de *Brier score* se define como:

$$\mathrm{BS}(t^*) = \frac{1}{n}\sum_{i=1}^{n}\{\left(0 - \hat{\pi}(t^*|X)\right)^2 I(\tilde{T}_i \leq t^*, \delta_i = 1)(1/\hat{G}(\tilde{T}_i)) + \left(1 - \hat{\pi}(t^*|X)\right)^2 I(\tilde{T}_i > t^*)(1/\hat{G}(t^*))\} \qquad (41)$$

Obteniéndose una estimación del *Brier score* con datos censurados para cada instante definido t^*.

El parámetro que se empleará para la evaluación de la calibración de los modelos de supervivencia es el *Brier score* integrado, que se obtiene a partir de la integración del vector resultante de la ecuación (22) mediante una función de pesos $W(t) = t/t^*$, tal que:

$$IBS = \int_0^{t^*} BS(t^*)\, dW(t) \qquad (42)$$

Por tanto, se tendrá un único marcador evaluable por cada uno de los modelos que oscilará entre 0 y 1, donde un *score* de 0 indica una calibración perfecta y un *score* de 1 la peor calibración posible (Graf et al., 1999).

Por otra parte, se empleará el índice de concordancia, o *c-index* de Harrell (Harrell et al., 1996), utilizado en el contexto de supervivencia para la determinación de la capacidad discriminativa de cada modelo. Está diseñado, más concretamente, para estimar la probabilidad de concordancia $P(\eta_j > \eta_i | T_i > T_j)$, que compara las posiciones de dos pares independientes de tiempo de supervivencia T_i y T_j y las predicciones η_j y η_i, donde los índices i y j se refieren a pares de observaciones. De esta forma, se evalúa si elevadores valores de riesgo o probabilidad de evento, η_i, están asociados con valores bajos de T_i, y viceversa (Schmid et al., 2015).

La expresión que sigue este índice es la siguiente:

$$C = \frac{\sum_{i,j} I\left(\tilde{T}_i > \tilde{T}_j\right) \cdot I\left(\eta_j > \eta_i\right) \cdot \Delta_j}{\sum_{i,j} I\left(\tilde{T}_i > \tilde{T}_j\right) \cdot \Delta_j} \qquad (43)$$

Donde Δ_j es el indicador de censura, tal que $\Delta_j = 1$ en caso de que T_i haya sido totalmente observado y $\Delta_j = 0$ en caso contrario. Por tanto, se cumple que $\Delta_j = I(T_i \leq C_i)$.

Este parámetro consiste, por dicha razón, en el cociente entre pares concordantes de observaciones entre el número total de pares comparables (Schmid et al., 2015). Es considerado análogo al área ROC bajo curvas tiempo dependientes (Heagerty & Zheng, 2005), y será aplicado tanto sobre el conjunto de entrenamiento como en el de test, lo cual facilitará la detección adicional de modelos con sobre entrenamiento.

Tras llevar a cabo una evaluación en términos de rendimiento de los modelos entrenados con los parámetros mencionados, se realizará un cribado de aquellos modelos con mejores resultados, seleccionando los mejores dentro de cada familia (basados en regularización, basados en técnicas *boosting* y basados en *random forest*). Tal y como se describe en la Tabla 3 de la sección de Estrategia de validación, en los modelos basados en *random forest* se empleará la importancia de variables para la selección de posibles factores pronóstico. Se ha decidido entrenar primeramente los modelos basados en *random forest* con todas las variables, para posteriormente seleccionar aquellas 50 con una mayor importancia y volver a entrenarlos únicamente con dichas variables. Esto dará como resultado un conjunto de modelos más sencillos y comparables, en términos de rendimiento, selección de variables y coste computacional.

Además, esto facilitará el posterior cruzado de variables seleccionadas, o factores pronóstico, es decir, cuyo coeficiente es no nulo en los modelos de regularización y basados en *boosting* y aquellas variables con una mayor importancia en los basados en *random forest*. Mediante la herramienta de búsqueda de información genética *Ensembl* (Martin et al., 2023), se llevará a cabo una búsqueda de los factores pronóstico obtenidos de los bloques de variables ómicos y su relación conocida con ciertas patologías. Así, se valorará si el uso de estos modelos en el análisis de supervivencia facilita la detección de biomarcadores en patologías como el cáncer.

Asimismo, en caso de identificar factores pronóstico, se empleará el estimador de Kaplan Meier para la comparación de curvas de supervivencia. De esta forma, y mediante la prueba *logrank*, se analizará y cuantificará su impacto en la evolución de la probabilidad de supervivencia a lo largo del seguimiento de los pacientes.

Finalmente, se comparará el tiempo de cómputo para todos los modelos entrenados. Este tiempo incluirá el proceso de optimización o *tuning* de cada uno, independientemente de si se ha realizado de forma externa o en la propia función de entrenamiento de cada modelo. Este detalle se aclara para cada modelo en la sección Implementación en RStudio. Además, dado que los tiempos de cómputo son independientes del tipo de escalado aplicado, no se hará esta distinción en su comparación, dado que el principal objetivo de esta sección será la de evaluar dicho coste computacional según las dimensiones de las bases de datos utilizadas.

3.2.3 Implementación en RStudio

Todos los modelos enumerados en la Tabla 3 de la sección de Estrategia de validación han sido entrenados, validados y evaluados en el *software* RStudio (versión R 4.2.2), en un equipo informático de las siguientes características:

- RAM: 8,00 GB
- Procesador: Intel® Core™ i5-6400 @ 2.70 Mhz
- Gráfica: NVIDIA GeForce GTX 1050 Ti

A continuación, se muestra una tabla con sus correspondientes librerías en R y sus funciones, incluyendo las de las medidas de rendimiento (Tabla 4). Aquellas técnicas con más de dos funciones asignadas (*Cox – Boost*, *MBoost* y *Random Survival Forest*) indica que su optimización y entrenamiento se ha realizado de forma separada.

Para el estimador Kaplan Meier, en cambio, se ha empleado una función para la representación de las curvas de supervivencia (‘survfit’) y otra para la realización del test estadístico *logrank* (‘survdiff’).

Tabla 4: Resumen de la implementación en RStudio de los modelos mencionados. Contiene las funciones empleadas, tanto para optimización y entrenamiento, como la correspondiente librería en R.

Técnica	**Función**	**Librería en R**
Standard Lasso y *Elastic Net*	cv.glmnet	*glmnet* (Friedman et al., 2010)
IPF – Lasso	cvr.ipflasso	*ipflasso* (Anne-Laure Boulesteix et al., 2022)
Priority – Lasso	cvm_prioritylasso	*prioritylasso* (Klau et al., 2023)
Cox – Boost	CoxBoost optimCoxBoostPenalty	*CoxBoost* (Harald Binder & Harald Binder, 2015)
MBoost	glmboost cvrisk	*mboost* (Bühlmann & Hothorn, 2007)
Random Survival Forest	rfsrc tune.rfsrc	*RandomForestSRC* (*Package "randomForestSRC" Title Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2023)
Block Forest	blockfor	*blockForest* (Hornung & Wright, 2019b)
Cox – PH	coxph	*survival* (*Package "Survival,"* 2023)
Kaplan Meier	survfit survdiff	*survcomp* (Haibe-Kains et al., 2017)
cindex Brier *Score*	concordance.index sbrier.score2proba	*survcomp*

4. Resultados y Discusión

4.1 Poder predictivo: BRCA

A continuación, se muestra una tabla que contiene los resultados en términos de poder predictivo para cada uno de los modelos ajustados en la base de datos de BRCA (Tabla 5).

Tabla 5: Resumen del rendimiento para BRCA. Se enuncia el modelo, si considera o no la estructura de bloque de la base de datos y los parámetros extraídos. El sufijo 'Par' indica un escalado de tipo pareto sobre los conjuntos de datos empleados.

Modelo	Bloque	Brier *score*	*Cindex train*	*Cindex test*
Cox PH (referencia)	No	0,386	0,801	0,816
Elastic Net	No	0,472	1	0,552
Elastic Net Par	Sí	0,258	0,953	0,804
Standard LASSO	No	0,474	1	0,471
Standard LASSO Par	Sí	0,192	0,671	0,630
IPF LASSO	Sí	0,199	0,889	0,797
Priority LASSO	Sí	0,530	0,893	0,498
CoxBoost	No	0,201	0,954	0,481
CoxBoost Par	Sí	0,201	0,759	0,685
MBoost	No	0,502	0,886	0,577
MBoost Par	Sí	0,157	0,911	0,577
Random Survival Forest	No	0,086	0,976	0,609
Block Forest	Sí	0,195	0,985	0,720

En primer lugar, y prestando atención a los resultados obtenidos tanto de calibración como de poder discriminativo, se puede apreciar como en aquellos modelos que consideran la estructura de bloques de la base de datos BRCA se obtienen, en general, mejores resultados. Esto se evidencia, principalmente, en los modelos basados en la estrategia LASSO, en los que se observa un Brier *score* menor y un aumento del *cindex* sobre el conjunto de test, exceptuando el modelo de *Priority LASSO.*

Algo similar sucede en los modelos basados en el algoritmo de *boosting*, donde el escalado de tipo pareto ha proporcionado una mejora a los modelos de *CoxBoost* y de *MBoost.*

Es interesante como los modelos que consideran dicha estructura de bloque muestran un *cindex* sobre el conjunto de entrenamiento menor, aunque un *cindex* sobre el conjunto de test superior, revelando que han conseguido una mejor estimación de los coeficientes, al no alcanzar un sobre ajuste. Esto último es apreciable en los modelos de *Elastic Net* y *Standard LASSO*, cuyo valor de *cindex* sobre el conjunto de entrenamiento es óptimo.

Por otra parte, la tabla muestra que los modelos basados en el algoritmo de *random forest* tienen los mejores valores de calibración, o Brier *score*. El *Random Survival Forest* supera con creces al resto de modelos en dicho parámetro, aunque la capacidad discriminativa se ve superada por el modelo de *Block Forest*, el cual implementa la estructura de bloques de la base de datos correspondiente.

Considerando el modelo de referencia de *Cox PH*, se puede comprobar que, únicamente implementando las variables explicativas de carácter clínico, se pueden obtener buenos resultados en términos de poder discriminativo y calibración. Aun así, modelos como *Elastic Net Par* y *IPF LASSO* demuestran que la

implementación de variables ómicas aseguran un mejor ajuste del modelo a los datos, al tener un Brier *score* inferior a *Cox PH*, asegurando al mismo tiempo una capacidad discriminativa similar al enfoque clásico, dado que el valor de *cindex* sobre el conjunto de test es similar.

A continuación, se listan los modelos que se han seleccionado para la obtención de factores pronóstico en la sección de Detección de factores pronóstico: BRCA, en base a su poder predictivo:

- *ElasticNet Par*
- *Standard LASSO Par*
- *IPF LASSO*
- *CoxBoost Par*
- *MBoost Par*
- *Random Survival Forest*
- *Block Forest*

Lo cual permitirá relacionar directamente las variables seleccionadas por estos modelos con potenciales biomarcadores, al estar seleccionando únicamente aquellos modelos que mejor explican la supervivencia de pacientes a las patologías de estudio.

Finalmente, se considera que el ajuste de los modelos planteados a la base de datos BRCA, y la capacidad discriminativa adquirida por estos, es adecuada con respecto a los parámetros empleados, alcanzando valores de Brier *score* por debajo de 0,2 y del *cindex* sobre el conjunto de test por encima de 0,7.

4.2 Poder predictivo: HNSC

A continuación, se muestra una tabla que contiene los resultados en términos de poder predictivo para cada uno de los modelos ajustados en la base de datos de HNSC (Tabla 6).

Tabla 6: Resumen del rendimiento para HNSC. Se enuncia el modelo, si considera o no la estructura de bloque de la base de datos y los parámetros extraídos. El sufijo 'Par' indica un escalado de tipo pareto sobre los conjuntos de datos empleados.

Modelo	**Bloque**	**Brier *score***	***Cindex train***	***Cindex test***
Cox PH (referencia)	No	0,128	0,621	0,552
Elastic Net	No	0,207	1	0,667
Elastic Net Par	Sí	0,209	0,681	0,517
Standard LASSO	No	0,115	0,832	0,669
Standard LASSO Par	Sí	0,255	0,711	0,495
IPF LASSO	Sí	0,165	0,639	0,579
Priority LASSO	Sí	0,143	0,771	0,681
CoxBoost	No	0,132	0,904	0,636
CoxBoost Par	Sí	0,132	0,835	0,474
MBoost	No	0,115	0,836	0,670
MBoost Par	Sí	0,153	0,861	0,557
Random Survival Forest	No	0,142	0,901	0,605
Block Forest	Sí	0,135	0,949	0,575

Tal y como se expone en la tabla, y a diferencia de lo que se ha obtenido para la base de datos de BRCA, los modelos en los que se incluye la información de la estructura de bloques de la base de datos tienen un rendimiento ligeramente inferior al resto, exceptuando el *Priority LASSO*. Este último destaca por tener el mejor poder discriminativo, es decir, el mayor valor del parámetro *cindex* para el conjunto de test, en comparación al resto de modelos, seguido por el *MBoost.* Por tanto, podría afirmarse que el escalado de tipo pareto no ha beneficiado, en este caso, un mejor ajuste del modelo a los tiempos de supervivencia, o lo que es lo mismo, no ha facilitado la selección de variables que proporcionan una mayor información para la predicción de la supervivencia.

Estos resultados también muestran que los modelos basados en el algoritmo de *random forest* tienen valores adecuados de calibración, pero con un poder discriminativo por debajo de la mayoría de los modelos. Además, aunque el modelo de *Elastic Net* tenga un valor de 1 para la métrica *cindex* para el conjunto de entrenamiento, su *cindex* en el conjunto de test es superior a la mayoría de los modelos, por lo que no se puede confirmar la presencia de sobre ajuste en ninguno de éstos.

Considerando el modelo de referencia *Cox PH*, se puede observar cómo, con la implementación de únicamente variables clínicas, no se alcanza un valor de capacidad discriminativa comparable con modelos que introducen variables ómicas, como *Priority LASSO* o *Standard LASSO.* Por tanto, en el caso de la base de datos HNSC, se puede afirmar con mayor facilidad que la inclusión de variables ómicas ha favorecido a la mejora de la capacidad discriminativa de los modelos.

A continuación, se listan los modelos que se han seleccionado para la obtención de factores pronóstico en la sección de Detección de factores pronóstico: HNSC, en base a su poder predictivo:

- *ElasticNet*
- *Standard LASSO*
- *Priority LASSO*
- *CoxBoost*
- *MBoost*
- *Random Survival Forest*
- *Block Forest*

Finalmente, se considera que el ajuste de los modelos planteados a la base de datos HNSC es adecuado, dado que todos los valores de Brier *score* están contenidos entre 0,1 y 0,2. Algunos de estos modelos han adquirido una capacidad discriminativa adecuada, alcanzando valores cercanos al 0,7 sobre el parámetro de *cindex* sobre el conjunto de test.

En comparación a los resultados obtenidos en BRCA, es clara la diferencia en cuanto al rendimiento de los modelos. Esto es posible que se deba al preprocesado del que ha sido posible disponer en el caso de BRCA, dado que se tratan de métodos para la eliminación de ciertos sesgos, aunque no se puede descartar la probabilidad de que las variables explicativas incluidas en HNSC no proporcionen tanta información sobre la supervivencia de los pacientes como en BRCA por diferencias de calidad entre ambas.

Cabe tener en cuenta que los modelos de *Random Survival Forest* y *Block Forest* se han entrenado únicamente con las primeras 50 variables con la mayor importancia, obtenida de un primer proceso de optimizado y entrenamiento de estos. De esta forma, y tal y como se ha comentado a lo largo de la metodología, se asegura una comparabilidad al hacerlos más accesibles a su aplicación clínica. Sin embargo, en caso de que el objetivo del uso de estos modelos sea conseguir un método con el mayor

rendimiento para la predicción de la supervivencia, dejando de un lado la interpretabilidad, el modelo de *Block Forest* demuestra ser la mejor opción (Herrmann et al., 2021a).

Por otra parte, y tal como se ha analizado en esta sección, para la selección de un modelo que facilite una interpretabilidad sobre las variables que explican mejor la supervivencia de los pacientes, se necesita de un proceso de entrenamiento, optimización y evaluación de todos los métodos planteados. De esta forma, no solo se consigue uno o varios modelos adecuados para la predicción, sino que también se da pie al análisis de las variables seleccionadas o coeficientes no nulos, y, por tanto, al análisis de posibles biomarcadores.

4.3 Detección de factores pronóstico: BRCA

En la siguiente tabla, se muestra el número de variables seleccionadas por cada uno de los modelos, además de su distribución para cada uno de los bloques de la base de datos de BRCA (Tabla 7).

Tabla 7: Resumen de las variables seleccionadas para BRCA. El sufijo 'Par' indica un escalado de tipo pareto sobre los conjuntos de datos empleados, y el símbolo '' indica qué modelos se han seleccionado en la sección de Poder Predictivo: BRCA.*

Modelo	**Bloque**	**N.º de variables**	***X.Clinical***	***X.Genes***	***X.miRNA***	***X.Methyl***
Elastic Net	No	526	1	367	33	125
*Elastic Net Par**	Sí	60	1	0	8	51
Standard LASSO	No	122	1	79	10	32
*Standard LASSO Par**	Sí	130	1	1	17	111
*IPF LASSO**	Sí	27	1	0	6	20
Priority LASSO	Sí	33	5	0	5	23
CoxBoost	No	49	1	29	1	18
*CoxBoost Par**	Sí	6	1	0	2	3
MBoost	No	23	2	15	1	5
*MBoost Par**	Sí	18	1	11	0	6
*Random Survival Forest**	No	50	1	37	3	9
*Block Forest**	Sí	50	23	14	13	0

En dicha tabla, se puede observar que, de entre los modelos seleccionados tras la evaluación de su rendimiento (marcados con '*' en Tabla 7), se han incluido principalmente variables pertenecientes al bloque de *X.Methyl.* Además, mientras que los modelos pertenecientes a la regresión penalizada han incluido también variables de *X.miRNA*, *MBoost Par* y los modelos basados en *random forest* tienen un mayor porcentaje de variables pertenecientes al bloque de *X.Genes.* Asimismo, es apreciable como la mayoría de los modelos únicamente han incorporado una única variable clínica, a pesar de que dicho bloque posee una menor penalización en los modelos que incorporan la estructura de bloques. Sin embargo, el *Block Forest* ha llegado a seleccionar en su totalidad el bloque de variables clínicas.

También es notable como los modelos basados en *Elastic Net* y su implementación del parámetro α ha favorecido la selección de una mayor cantidad de variables. Dada la naturaleza de dicho modelo, es muy

probable que las variables seleccionadas tengan, entre ellas, cierta correlación para su inclusión en el modelo final.

Por otra parte, la implementación del escalado de tipo pareto no solo ha favorecido la mejora en el rendimiento de ciertos modelos que no incorporan la estructura de bloques, sino que también ha favorecido la reducción del número de variables seleccionadas en los bloques de variables ómicas. Estos modelos son *Elastic Net, Standard LASSO, CoxBoost y MBoost.*

Los modelos seleccionados tras la evaluación de su rendimiento tienen, exceptuando el modelo *Standard LASSO Par*, un número de variables o coeficientes no nulos reducido. Partiendo de la regla EPV (*Events per variable*) (J. Zheng et al., 2021) por la que se recomiendan, aproximadamente, 10 eventos por la incorporación de cada variable al modelo, y teniendo en cuenta que la base de datos BRCA cuenta con hasta 58 eventos, el número de variables incluidas en los modelos, aun siendo reducido, es superior al número recomendado (≈ 6), exceptuando el modelo *CoxBoost Par*. Sin embargo, dados a sus resultados en las métricas de estudio, se considera que los modelos seleccionados son adecuados.

Tras comentar los resultados generales sobre las variables seleccionadas por cada modelo, se procede a analizar los distintos subconjuntos que se crean de éstas a partir de la intersección de todas las variables de cada uno de los modelos elegidos. Esto se realiza por medio de un gráfico *upset* (Conway et al., 2017), cuyo resultado se muestra a continuación (Figura 4).

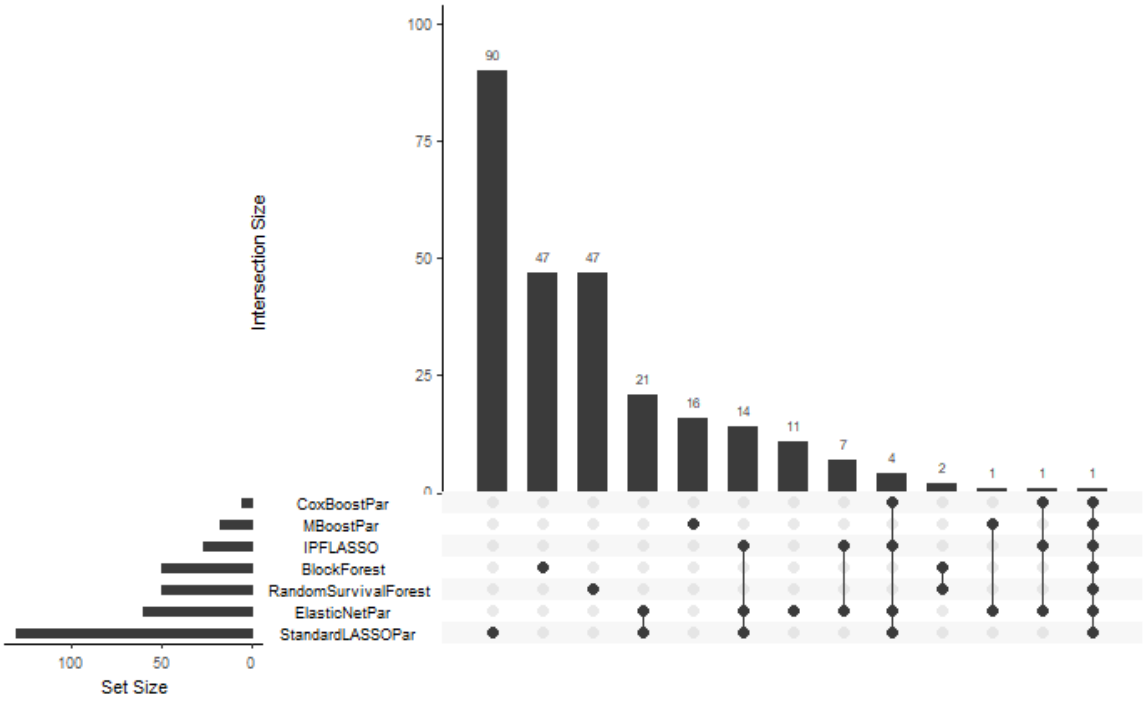

Figura 4: Gráfico upset resultante de las variables seleccionadas por los modelos de mejor rendimiento para BRCA. En filas, a la izquierda del gráfico, se tienen los modelos seleccionados junto al número de variables de cada uno (Set Size). La matriz de puntos conecta aquellos modelos para los que se tienen intersecciones. De izquierda a derecha, en cada columna, se indican el número de intersecciones de mayor a menor (Intersection Size). Aquellos puntos sin unir indican variables únicamente seleccionadas por el modelo en cuestión.

Las variables que hacen intersección, ordenadas de mayor a menor aparición para las distintas combinaciones de modelos (Tabla 9, Anexo I) revela ciertas características de los posibles factores pronóstico del BRCA.

En primer lugar, gran parte de las variables que coinciden entre los modelos elegidos forman parte del bloque de *X.Methyl*, constituyendo aproximadamente un 80% del total. También se puede observar la presencia de variables pertenecientes al bloque de *X.miRNA*, siendo las dos restantes de los bloques de *X.Genes* y de *X.Clinical*, donde la edad de diagnóstico es la variable clínica seleccionada por todos los modelos.

Cabe mencionar que, en casos de carcinoma invasivo de mama durante la práctica clínica, la edad del paciente está considerada un factor de riesgo, siendo esencial en la etapa de diagnóstico, junto a los correspondientes marcadores genéticos BRCA1 y BRCA2, para la orientación de cada caso al tipo de cirugía más adecuada (Beattie et al., n.d.).

Haciendo uso del estimador Kaplan Meier, se obtienen las curvas de supervivencia en función de la edad de los pacientes. Categorizando dicha variable en binaria, se obtienen dos curvas de supervivencia: una para pacientes por encima de 60 años, y otra para pacientes de edades por debajo de dicha edad. Esta edad se ha fijado de acuerdo con la mediana de la variable 'age_at_diagnosis'.

Así, se obtiene el siguiente resultado (Figura 5):

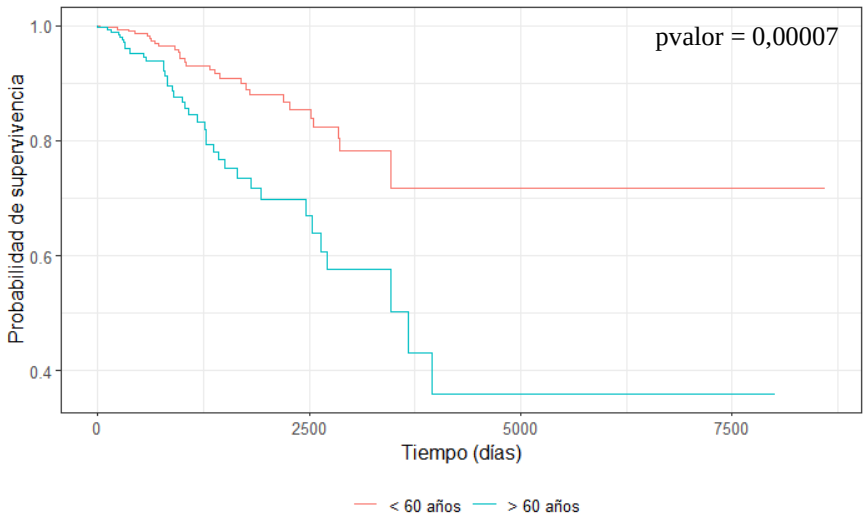

Figura 5: Curvas de supervivencia Kaplan Meier en función de la edad en BRCA. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo, estimación para sujetos con menos de 60 años. En azul, estimación para sujetos mayores de 60 años.

Tal y como se puede apreciar en dicho gráfico, los individuos más jóvenes tienen una mayor probabilidad de supervivencia ante el carcinoma invasivo de pecho, teniendo una probabilidad de aproximadamente el 70% de que no se produzca el evento de interés, o la muerte, al final del estudio. Por otra parte, la población de individuos con una edad por encima de 60 años alcanza una probabilidad menor del 40% de supervivencia. Además, la reducción de dicha probabilidad en el caso de los individuos de mayor edad es más pronunciada en etapas más tempranas.

Dado que el test estadístico *logrank* devuelve un pvalor por debajo del riesgo de primera especie de 0,05, se puede afirmar con un 95% de confianza que existen diferencias estadísticamente significativas entre ambas curvas de supervivencia. Por tanto, la selección de dicha variable clínica como factor pronóstico se ve reforzado por los resultados obtenidos mediante el estimador Kaplan Meier.

Se han obtenido además las curvas Kaplan Meier de las variables ómicas más frecuentemente seleccionadas por los modelos (Frecuencia 1 en Tabla 9, Anexo I), categorizando previamente dichos predictores empleando su mediana. A continuación, se muestran las curvas de supervivencia para una de estas variables (Figura 6).

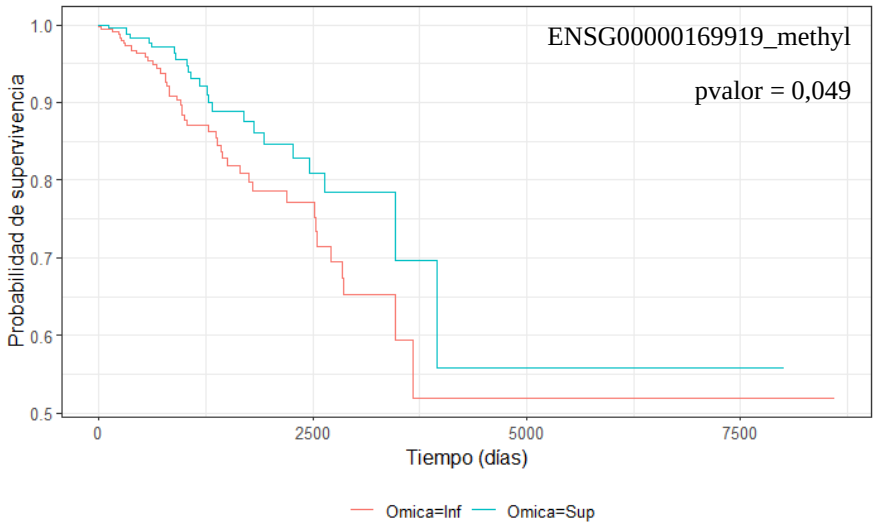

Figura 6: Curvas de supervivencia Kaplan Meier en función de ENSG00000169919_methyl en BRCA. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

Tal y como se puede apreciar en dicho gráfico, sujetos con niveles bajos de metilación del gen 'ENSG00000169919' muestran una probabilidad de supervivencia significativamente menor a lo largo del tiempo, con un pvalor del test *logrank* menor que 0,05. Por tanto, al igual que la edad de diagnóstico, la selección de dicha variable ómica se ve reforzada por el estimador Kaplan Meier.

Por otra parte, las curvas de supervivencia obtenidas para la segunda variable ómica seleccionada (Figura 12, Anexo I) indican que no se presentan diferencias estadísticamente significativas. A pesar de esto, cabe tener en cuenta la posibilidad de que el análisis multivariante llevado a cabo mediante los modelos implementados y el estimador Kaplan Meier no coincidan, dado que en este último no se tiene en cuenta el posible efecto del resto de variables.

Mediante el buscador de genes *Ensembl*, ha sido posible clasificar las variables ómicas que se listan en la Tabla 9 (Anexo I), y cuya función puede ser de ayuda para evaluar la detección de factores pronóstico de los modelos propuestos.

La relación de dichos genes o microRNAs con condiciones genéticas o patologías (Tabla 11, Anexo I) muestra como la cuantificación de la metilación de cinco de los genes posee relación con la supervivencia de pacientes que padecen carcinoma o carcinoma invasivo de mama. Por lo tanto, se han detectado biomarcadores de la patología de interés ya reconocidos, lo cual indica que los modelos elegidos han sido capaces de detectar una alteración de la metilación (Choudhuri, 2014) en un porcentaje

importante de los genes implementados en el estudio. El resto de las variables ómicas o bien poseen relación con otras patologías (≈ 64%) o bien es desconocida (≈ 24%), lo cual da pie a posibles estudios sobre dichos factores.

Adicionalmente, los microRNAs 'hsa-let-7d-5p', 'hsa-miR-26a-5p' y 'hsa-miR-148b-5p' han sido identificados en otros estudios como biomarcadores del carcinoma invasivo de mama. Sus respectivas funciones en esta patología podrían ser su influencia sobre la sensibilidad a ciertas terapias o fármacos (Tormo et al., 2017) (Uhr et al., 2019) o la recurrencia de la enfermedad y su progresión patológica (Ning et al., 2023), respectivamente. Las moléculas 'hsa-miR-340-3p' y 'hsa-miR-589-5p' también han sido identificados en otros estudios como posibles biomarcadores (Uhr et al., 2019) (Martinez-Gutierrez et al., 2020).

Finalmente, se considera que la detección de factores pronóstico para la base de datos de BRCA por medio de modelos basados en análisis de supervivencia ha sido satisfactoria, al identificar una serie de variables ómicas las cuales la literatura científica, junto a *Ensembl*, ha permitido considerarlas como potenciales biomarcadores, tanto en el campo de la genómica como en el de la transcriptómica. Adicionalmente, el estimador Kaplan Meier ha permitido corroborar el impacto del factor de riesgo clínico reconocido (edad de diagnóstico) sobre la probabilidad de supervivencia de los individuos.

4.4 Detección de factores pronóstico: HNSC

En la siguiente tabla, se muestra el número de variables seleccionadas por cada uno de los modelos, además de su distribución para cada uno de los bloques de la base de datos de HNSC.

Tabla 8: Resumen de las variables seleccionadas para HNSC. El sufijo 'Par' indica un escalado de tipo pareto sobre los conjuntos de datos empleados, y el símbolo '' indica qué modelos se han seleccionado en la sección de Poder Predictivo: BRCA.*

Modelo	Bloque	N.º de variables	*X.Clinical*	*X.Genes*	*X.miRNA*	*X.cnv*	*X.Mutation*
*Elastic Net**	No	703	1	375	37	290	0
Elastic Net Par	Sí	16	0	0	16	0	0
*Standard LASSO**	No	40	1	29	4	6	0
Standard LASSO Par	Sí	22	0	1	21	0	0
IPF LASSO	Sí	7	0	0	7	0	0
*Priority LASSO**	Sí	24	1	22	0	0	1
*CoxBoost**	No	66	1	46	6	13	0
CoxBoost Par	Sí	62	1	12	49	0	0
*MBoost**	No	38	1	28	4	5	0
MBoost Par	Sí	44	0	32	1	11	0
*Random Survival Forest**	No	50	1	22	6	21	0
*Block Forest**	Sí	50	12	0	38	0	0

En dicha tabla, se puede observar que, de entre los modelos seleccionados tras la evaluación de su rendimiento (marcados con * en Tabla 8), se han incluido principalmente variables pertenecientes al bloque de *X.Genes.* Por el otro lado, las variables del bloque de *X.Mutation* no han sido seleccionadas en ningún modelo, exceptuando el *Priority LASSO.* Modelos en los que se ha realizado el escalado de tipo pareto ha permitido la incorporación de un mayor porcentaje de variables del bloque de *X.miRNA,* como *Elastic Net Par, Standard LASSO Par* o *CoxBoost Par*.

Además, estos métodos, junto a *MBoost Par,* no han incorporado ninguna variable clínica en sus respectivos modelos, a diferencia del resto, los cuales han incluido únicamente una variable del bloque *X.Clinical.* Sin embargo, el *Block Forest,* al igual que en la base de datos de BRCA, ha llegado a seleccionar en su totalidad el bloque de variables clínicas.

Asimismo, los modelos cuyos conjuntos de datos han sido tratados con el escalado de tipo pareto ha resultado en la selección de un menor número de variables. Esto es más notable comparando los modelos *Elastic Net* y *Elastic Net Par,* donde el primero, gracias al factor α, ha incluido un mayor número de variables correlacionadas, viéndose limitado por el posterior escalado en su versión *Par.*

Partiendo de la regla EPV (J. Zheng et al., 2021) anteriormente mencionada, y teniendo en cuenta que la base de datos de HNSC cuenta con 152 eventos, el número de variables o coeficientes un nulos incluidos por los modelos es, de nuevo, superior al recomendado (≈ 15), siendo el modelo más adecuado en cuanto a dicha regla el *Priority LASSO.* Sin embargo, dados sus resultados en rendimiento, se considera que los modelos seleccionados son adecuados.

Tras comentar los resultados generales sobre las variables seleccionadas por cada modelo, se procede a analizar las variables coincidentes entre los modelos elegidos previamente. El gráfico *upset* (Conway et al., 2017) obtenido en este caso se muestra a continuación (Figura 7).

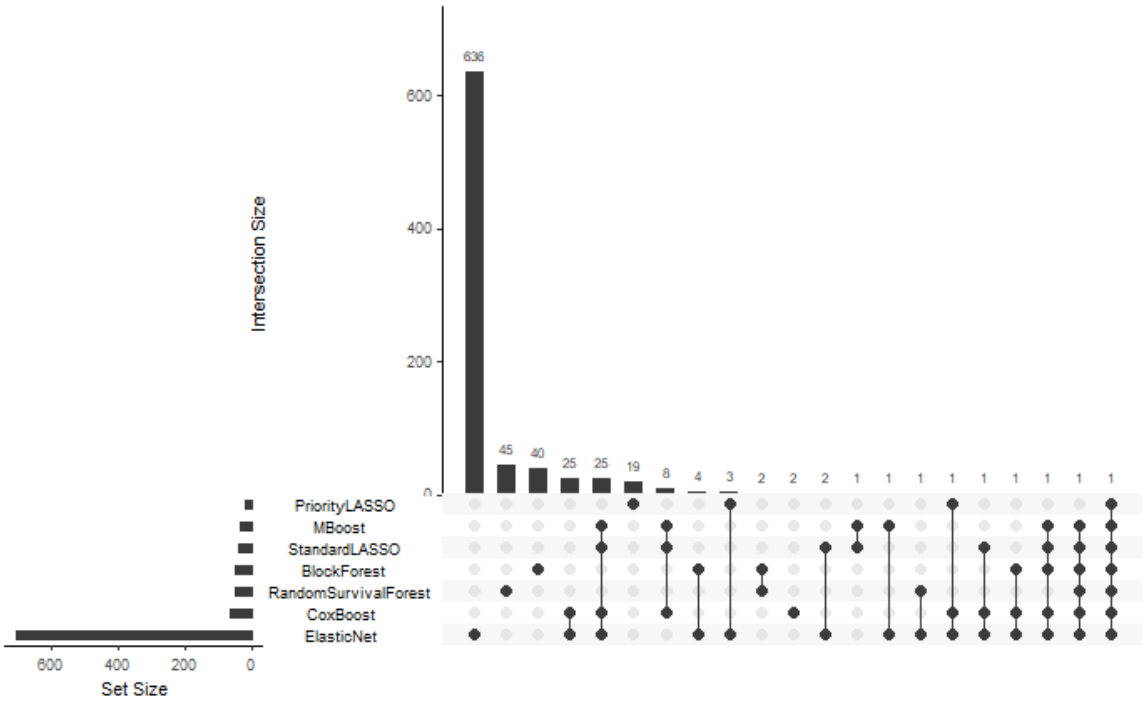

Figura 7: Gráfico upset resultante de las variables seleccionadas por los modelos de mejor rendimiento para HNSC. En filas, a la izquierda del gráfico, se tienen los modelos seleccionados junto al número de variables de cada uno (Set Size). La matriz de puntos conecta aquellos modelos para los que se tienen intersecciones. De izquierda a derecha, en cada columna, se indican el número de intersecciones de mayor a menor (Intersection Size). Aquellos puntos sin unir indican variables únicamente seleccionadas por el modelo en cuestión.

Las variables que hacen intersección, ordenadas de mayor a menor aparición para las distintas combinaciones de modelos (Tabla 10, Anexo I) revela ciertas características de los posibles factores pronóstico del HNSC.

Primeramente, gran parte de las variables que forman las intersecciones que se muestran en el gráfico (Figura 6) son del bloque de variables *X.Genes*. Asimismo, las variables que forman parte de la mayoría de los modelos son de los bloques *X.genes* y *X.miRNA*, siendo la minoría de estas pertenecientes al bloque *X.cnv*. Además, de igual forma que en BRCA, la edad del individuo es la única variable clínica de dicho listado.

Haciendo uso del estimador Kaplan Meier, se obtienen las curvas de supervivencia en función de la edad de los pacientes. Categorizando dicha variable en binaria, se obtienen dos curvas de supervivencia: una para pacientes por encima de 62 años, y otra para pacientes de edades por debajo de dicha edad. Esta edad se ha fijado de acuerdo con la mediana de la variable 'age'.

Así, se obtiene el siguiente resultado (Figura 8):

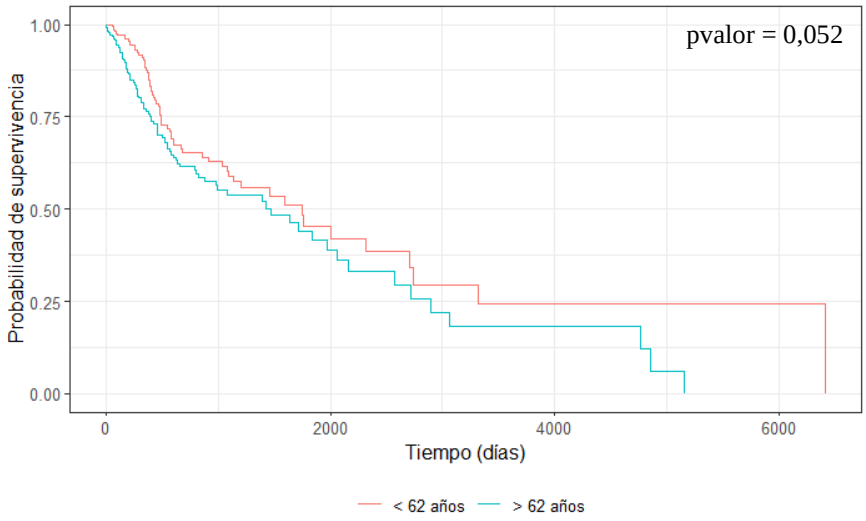

Figura 8: Curvas de supervivencia Kaplan Meier en función de la edad en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo, estimación para sujetos con menos de 62 años. En azul, estimación para sujetos mayores de 62 años.

Observando el gráfico, se puede apreciar como los individuos más jóvenes tienen una mayor probabilidad de supervivencia ante el cáncer de cuello y cabeza, dado que la evolución de dicha probabilidad es, a lo largo de la totalidad del estudio, menor que en pacientes de mayor edad.

Dado que el test estadístico *logrank* devuelve un pvalor muy cercano al riesgo de primera especie de 0,05, se puede afirmar con un 95% de confianza que existen diferencias estadísticamente significativas entre ambas curvas de supervivencia. Por esto, la selección de la edad como factor pronóstico se ve reforzado por los resultados obtenidos mediante el estimador Kaplan Meier, además de considerarse un factor de riesgo en la práctica clínica (Garavello et al., 2006).

Se han obtenido además las curvas Kaplan Meier de las variables ómicas más frecuentemente seleccionadas por los modelos (Frecuencia 1 en Tabla 10, Anexo I), categorizando previamente dichos predictores empleando su mediana. Así, se representa la evolución de la probabilidad de supervivencia en función del nivel de expresión de cada ómica. A continuación, se muestran las curvas de supervivencia para una de estas variables (Figura 9).

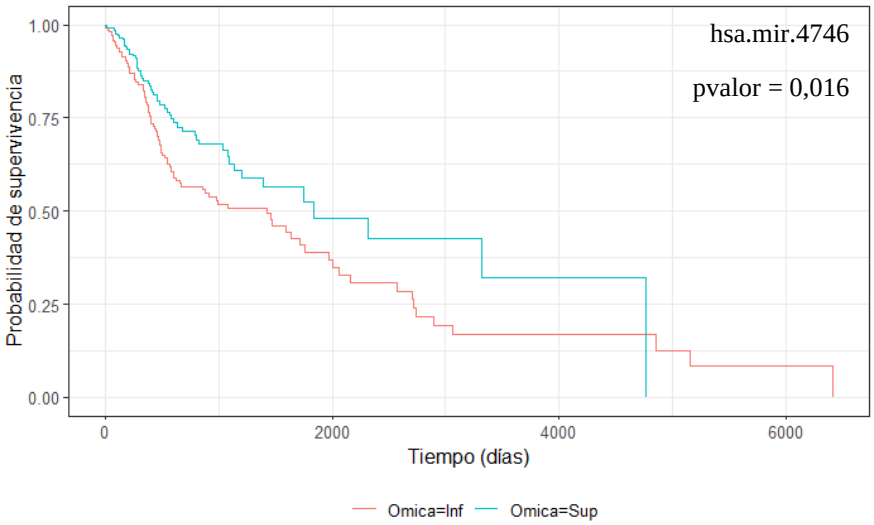

Figura 9: Curvas de supervivencia Kaplan Meier en función de hsa.mir.4746 en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

Tal y como se puede apreciar en dicho gráfico, sujetos con niveles bajos del biomarcador 'hsa.mir.4746' muestran una probabilidad de supervivencia significativamente menor a lo largo del tiempo, con un pvalor del test *logrank* menor que 0,05. Lo mismo sucede con otras de las variables ómicas seleccionadas ('hsa.mir.3664', 'hsa.mir.552', 'ENSG00000095370', 'ENSG00000010310'), en las que su sobreexpresión puede indicar un aumento o disminución de la probabilidad de supervivencia (Figuras 13, 14, 16 y 17, Anexo I). Así, de igual forma que en BRCA, se muestra que la selección de posibles factores pronóstico llevada a cabo se ve reforzada por el estimador Kaplan Meier, aunque algunas de las variables ómicas seleccionadas no presenten dicha significatividad (Figuras 15, 18 y 19, Anexo I).

Con la ayuda de la herramienta de búsqueda de *Ensembl*, se ha verificado la relación entre la expresión de los genes y de las CNV seleccionadas y sus patologías asociadas (Tabla 12, Anexo I). Tal y como se puede observar en la categorización realizada, tres de los genes seleccionados tiene una relación directa con el cáncer de cuello y cabeza (HNSC). Del resto de variables ómicas, o bien tienen relación con otras patologías (≈ 23%) o bien es desconocida (≈ 72%), lo cual da pie a posibles estudios de dichos factores. Aun así, algunos de estos genes figuran como posibles biomarcadores en la literatura científica, producto de estudios llevados a cabo sobre la misma base de datos (Jiang et al., 2020) (Vega-Benedetti et al., 2019) (Z. Zheng et al., 2022).

Mediante las variables del bloque *X.miRNA* coincidentes se han encontrado ciertas relaciones de dichos transcritos con el cáncer de cuello y cabeza. Estos son 'hsa.mir.4746' y 'hsa.mir.135a.2', reconocidos como posibles biomarcadores (Hu et al., 2018) (Lopez-Rincon et al., 2019).

Por tanto, se considera que la detección de factores pronóstico para la base de datos de HNSC ha sido satisfactoria, al identificar una serie de variables ómicas las cuales la literatura científica, junto a *Ensembl*, ha permitido considerarlas como potenciales biomarcadores.

Aunque en ambos casos de estudio, tanto BRCA como HNSC, los modelos han sido capaces de detectar genes o microRNAs cuya expresión irregular o mutación pueden tener una relación con la patología, sería necesaria una mayor evidencia, junto a conocimiento experto y métodos complementarios, para

considerar las variables ómicas seleccionadas como biomarcadores. De todas formas, se ha conseguido demostrar la utilidad de las diversas técnicas estadísticas para la modelización de datos de supervivencia en la orientación del estudio de biomarcadores. Estas han permitido, por medio de distintos criterios o algoritmos, reducir drásticamente el abanico de posibles variables ómicas candidatas, facilitando un posterior análisis más preciso y el descarte de ciertos bloques de variables sin suficiente información para la predicción de la supervivencia.

4.5 Coste computacional

En el siguiente gráfico (Figura 10) se muestra el coste computacional, medido en segundos, de cada uno de los modelos entrenados. Estos tiempos comprenden no sólo su entrenamiento, sino también su optimización.

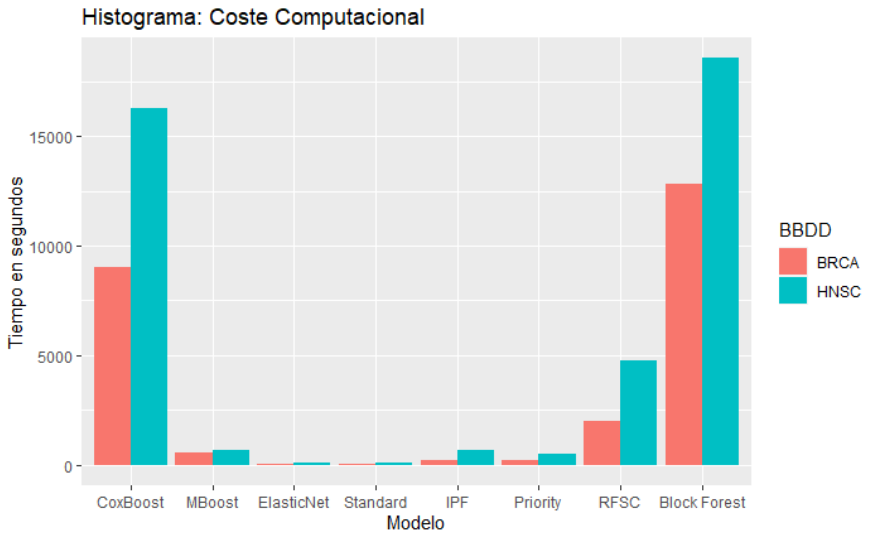

Figura 10: Gráfico de barras del tiempo de optimización y entrenamiento de cada modelo. En rojo, para la BBDD BRCA, en azul, para HNSC.

Partiendo de la gran diferencia de número de variables explicativas entre las bases de datos de BRCA y HNSC, con unas dimensiones de 421 x 37421 y 311 x 80398 de los respectivos conjuntos de entrenamiento, es lógico que el coste computacional sea mayor, e incluso llegue a duplicarse, en el caso de la base de datos de HNSC.

Este incremento del coste computacional es más notable en el caso de los modelos de *Block Forest*, *CoxBoost* y RFSC, donde el máximo tiempo de entrenamiento y optimización alcanza, aproximadamente, las 6,38 horas (23.000 segundos) para el modelo de *Block Forest* en la base de datos de HNSC. Estos tiempos se ven seguidos por los correspondientes a los modelos basados en regresión penalizada y *MBoost*, que alcanzan, en promedio, tiempos entre 500 y 1000 segundos.

Con el fin de comparar con mayor facilidad los modelos de mayor a menor coste computacional, se ha obtenido el mismo gráfico, pero en escala logarítmica (Figura 11).

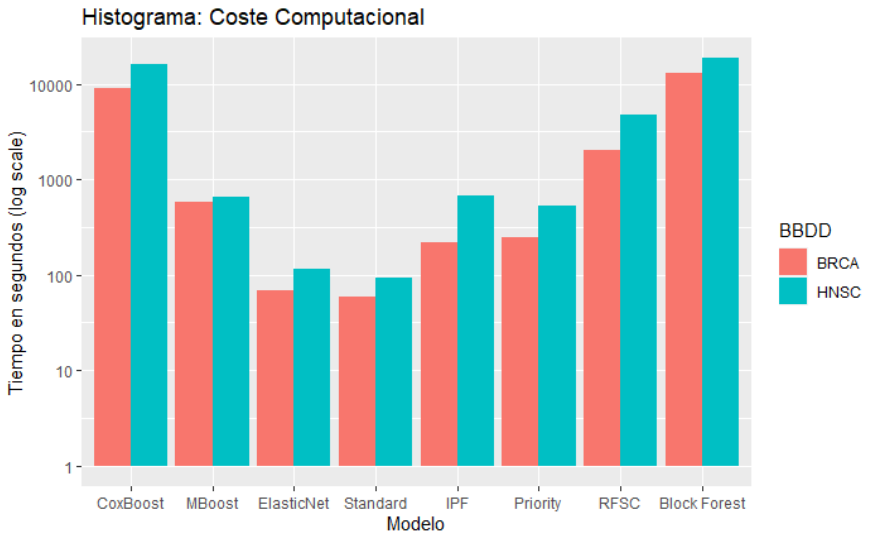

Figura 11: Gráfico de barras del tiempo de optimización y entrenamiento de cada modelo. En rojo, para la base de datos (BBDD) BRCA, en azul, para HNSC. El tiempo está en escala logarítmica.

Esta figura muestra cómo todos los modelos basados en la regresión penalizada poseen un tiempo de cómputo menor en comparación al resto. Entre estos, tanto *IPF LASSO* como *Priority LASSO*, que son los que implementan en un algoritmo la estructura en bloque de la base de datos, poseen un coste computacional superior.

Este coste computacional viene seguido por los modelos de *MBoost* y RFSC, siendo el primero mencionado el que mayor coste computacional describe de entre los modelos que no implementan dicha estructura de bloques.

Junto a los resultados obtenidos durante la evaluación del rendimiento de los modelos, se pueden llegar a algunas conclusiones con respecto a la relación rendimiento coste computacional de estos. El modelo *Cox PH*, con un entrenamiento instantáneo y con únicamente información clínica, ha demostrado ser competitivo con modelos que incluyen variables ómicas. Sin embargo, la inclusión de este tipo de variables facilita el análisis de posibles biomarcadores, como se ha analizado en la sección de Detección de factores pronóstico para ambas bases de datos, y se han podido obtener modelos con variables ómicas y un buen rendimiento sin un coste computacional demasiado elevado, como *MBoost* o los basados en regresión penalizada.

Aun así, modelos que incluyen estas bases de datos de alta dimensionalidad experimentan un mayor coste computacional, viéndose acentuado este incremento por norma general en caso de emplear un modelo que implementa su estructura de bloques durante el entrenamiento de este. Además, tal y como se ha mencionado, los métodos tradicionales en análisis de supervivencia, como la modelización *Cox PH* o el estimador Kaplan Meier, gozan de una mayor flexibilidad en términos de coste computacional. Tal y como otros estudios similares han demostrado, estas aproximaciones tienen capacidades similares en términos de rendimiento e interpretabilidad a los modelos adaptados a las bases de datos multiómicas. Sin embargo, una práctica recomendada es la implementación de ambos enfoques, dadas las ventajas de cada uno (Herrmann et al., 2021a).

5. Conclusiones

En el presente trabajo se ha llevado a cabo una comparación de diversas técnicas estadísticas para el análisis de supervivencia que se ha abordado en varias etapas, sobre bases de datos de alta dimensionalidad de carácter ómico, en términos de rendimiento, identificación de factores pronóstico y coste computacional. Las dos bases de datos seleccionadas incluyen información genómica y transcriptómica, además de variables clínicas, y describen dos patologías: cáncer invasivo de mama (BRCA) y cáncer de cuello y cabeza (HNSC).

A continuación, se responde a cada uno de los objetivos específicos definidos al inicio del estudio:

- Tras realizar la búsqueda pertinente, se ha recopilado una lista de modelos que se adapta a los datos de interés, incluyendo entre estos técnicas estadísticas capaces de implementar la estructura de bloques en el ajuste del modelo, perteneciente a 3 posibles familias: basados en regularización, *boosting* o *random forest.*

- Asimismo, el criterio común para la comparación de los modelos teniendo en cuenta la naturaleza multi – bloque de los datos se ha basado en el principio de que, a mayor el número de variables de cada bloque, menor es la información que contiene para la explicación de la supervivencia de los pacientes. Además de la implementación del escalado de tipo pareto, que también se ajusta a este criterio, y que posteriormente se vería que puede llegar a mejorar el rendimiento de ciertos modelos.

- La evaluación del desempeño de los modelos sugiere que el proceso de obtención de un modelo óptimo debe estar constituido por una optimización y entrenamiento de los diversos enfoques actuales. Mientras que para la base de datos de BRCA, el modelo tradicional de *Cox PH* arroja buenos resultados, seguido por las técnicas basadas en regresión penalizada, los mejores modelos para la base de datos de HNSC están basados en técnicas *boosting* y en regresión penalizada. La inclusión de modelos que consideran la estructura multi – bloque de la base de datos ha resultado en una mejora general de los resultados en su rendimiento. Asimismo, la implementación de las variables ómicas ha resultado en una mejora general del desempeño de los modelos. Por último, se recomienda el uso de técnicas de preprocesado como el escalado de tipo pareto, que ha proporcionado a los modelos de la base de datos de BRCA una mejora en la discriminación de las muestras según sus predicciones.

- A partir de la selección de modelos con mejor desempeño se han estudiado los factores pronóstico identificados por los modelos para cada patología, siendo por tanto una etapa esencial del estudio. La selección de dichas variables ómicas y su búsqueda en la base de datos genómica *Ensembl* ha mostrado la utilidad de las técnicas estadísticas aplicadas para el análisis de potenciales biomarcadores, al haber detectado variables con una relación consolidada con cada una de las patologías. Del mismo modo, dicha identificación de posibles factores pronóstico ha permitido descartar bloques de variables con poca información predictiva para la supervivencia de los pacientes de estudio, orientando así el estudio genético de cada patología.

- El estimador de Kaplan Meier ha permitido conocer la relación entre las variables de riesgo seleccionadas con la probabilidad de supervivencia a lo largo del tiempo, observando una reducción de esta en pacientes de mayor edad y en distintos niveles, altos o bajos, de los factores ómicos. Además, el test estadístico realizado reafirma la selección llevada a cabo por los modelos para la identificación de factores pronóstico.

- La estimación del coste computacional de cada modelo ha facilitado la categorización de cada modelo en función de su ratio tiempo de ejecución – rendimiento. Los modelos que poseen una mejor ratio entre dichas características son el *Cox PH*, con un tiempo de cómputo prácticamente nulo, los basados en regresión penalizada y el *MBoost*. Además, el coste computacional de todos los modelos cambia de forma coherente ante bases de datos con un mayor número de variables explicativas, llegando a casi duplicarse en modelos como *Block Forest* o *Random Survival Forest* cuando se implementan los datos de HNSC.

Se puede concluir que el estudio ha permitido abordar una comparación adecuada de modelos de supervivencia para datos de alta dimensiones sobre bases de datos ómicas, al haber sido implementados mediante estrategias de validación y evaluación correctas, y cuyo análisis ha demostrado su potencial para la predicción de la probabilidad de supervivencia y para la detección de potenciales biomarcadores de carcinoma invasivo de mama y de cáncer de cuello y cabeza.

Esta evaluación llevada a cabo sobre los tres aspectos mencionados ha permitido deducir un conjunto de métodos que se recomiendan de forma general, entre los que figuran *IPF LASSO*, *Priority LASSO*, *MBoost* y *Cox PH*, siendo los basados en *random forest* los menos recomendados. Las diferencias en términos de desempeño de los modelos para las dos bases de datos empleadas es probable que se deban a la eliminación de sesgos llevada a cabo sobre la base de datos BRCA, aunque no se descarta que el origen se trate de una diferencia de calidad entre ambas, con un peor registro de las variables ómicas en el caso de HNSC.

Finalmente, y a modo de ampliación del trabajo, se propone la inclusión en esta comparativa de modelos basados en PLS (*Projection to Latent Structures*) (Wold et al., n.d.), técnica estadística basada en la regresión de espacios latentes de las matrices de predictores y de variables repuesta. Este métodos es capaz de manejar bases de datos de alta dimensionalidad, además de que, dada la ortogonalidad del espacio latente, se descarta la influencia de la multicolinealidad. Es por esto por lo que es una aproximación muy interesante al análisis de supervivencia, además de que ofrece una amplia variedad de herramientas para la interpretación de coeficientes e identificación de factores pronóstico.

6. Referencias bibliográficas

Anne-Laure Boulesteix, A., Fuchs, M., & Schulze Maintainer Anne-Laure Boulesteix, G. (2022). *Type Package Title Integrative Lasso with Penalty Factors*. https://doi.org/10.5282/ubm/epub.59092

Beattie, M. S., Crawford, B., Lin, F., Vittinghoff, E., & Ziegler, J. (n.d.). *Uptake, Time Course, and Predictors of Risk-Reducing Surgeries in BRCA Carriers*. www.liebertpub.com

Bebis, G., Athitsos, V., Yan, T., Lau, M., Li, F., Shi, C., Yuan, X., Mousas, C., & Bruder, G. (n.d.). Advances in Visual Computing. In *Proceedings*. http://www.springer.com/series/7412

Boulesteix, A. L., De Bin, R., Jiang, X., & Fuchs, M. (2017). IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*, *2017*. https://doi.org/10.1155/2017/7691937

Breiman, L. (2001). *Random Forests* (Vol. 45).

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477–505. https://doi.org/10.1214/07-STS242

Choudhuri, S. (2014). Fundamentals of Genes and Genomes. In *Bioinformatics for Beginners* (pp. 1–25). Elsevier. https://doi.org/10.1016/b978-0-12-410471-6.00001-3

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M. K., Factor, R. E., Collins, L. C., Allison, K. H., … Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506–519. https://doi.org/10.1016/j.cell.2015.09.033

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

De Bin, R. (2016). Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics*, *31*(2), 513–531. https://doi.org/10.1007/s00180-015-0642-2

De Bin, R., Sauerbrei, W., & Boulesteix, A. L. (2014). Investigating the prediction ability of survival models based on both clinical and omics data: Two case studies. *Statistics in Medicine*, *33*(30), 5310–5329. https://doi.org/10.1002/sim.6246

Donoho, D. L. (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Garavello, W., Ciardo, A., Spreafico, R., & Gaini, R. M. (2006). Risk Factors for Distant Metastases in Head and Neck Squamous Cell Carcinoma. In *Arch Otolaryngol Head Neck Surg* (Vol. 132). www.archoto.com

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). ASSESSMENT AND COMPARISON OF PROGNOSTIC CLASSIFICATION SCHEMES FOR SURVIVAL DATA. In *STATISTICS IN MEDICINE Statist. Med* (Vol. 18).

Haibe-Kains, B., Schroeder, M., Olsen, C., Sotiriou, C., Bon-Tempi, G., & Maintainer, J. Q. (2017). *Package "survcomp" Type Package Title Performance Assessment and Comparison for Survival Analysis*. http://www.bordet.be/en/services/medical/array/practical.htm

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. In *Cell* (Vol. 100).

Harald Binder, A., & Harald Binder, M. (2015). *Package "CoxBoost" Title Cox models by likelihood based boosting for a single survival endpoint or competing risks*.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. In *Genome Biology* (Vol. 18, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-017-1215-1

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., & Boulesteix, A. L. (2021a). Large-scale benchmark study of survival prediction methods using multi-omics data. In *Briefings in Bioinformatics* (Vol. 22, Issue 3). Oxford University Press. https://doi.org/10.1093/bib/bbaa167

Hornung, R., & Wright, M. N. (2019b). Block Forests: Random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2942-y

Hothorn, T., & Bühlmann, P. (2006). Model-based boosting in high dimensions. *Bioinformatics*, *22*(22), 2828–2829. https://doi.org/10.1093/bioinformatics/btl462

Hu, Y., Dingerdissen, H., Gupta, S., Kahsay, R., Shanker, V., Wan, Q., Yan, C., & Mazumder, R. (2018). Identification of key differentially expressed MicroRNAs in cancer patients through pan-cancer analysis. *Computers in Biology and Medicine*, *103*, 183–197. https://doi.org/10.1016/j.compbiomed.2018.10.021

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Jenkins, S. P. (2005). *Survival Analysis*.

Jiang, P., He, S., Li, Y., & Xu, Z. (2020). Identification of Therapeutic and Prognostic Biomarkers of Lamin C (LAMC) Family Members in Head and Neck Squamous Cell Carcinoma. *Medical Science Monitor*, *26*. https://doi.org/10.12659/MSM.925735

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 367, Issue 1906, pp. 4237–4253). Royal Society. https://doi.org/10.1098/rsta.2009.0159

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Khene, Z.-E., Bigot, P., Doumerc, N., Ouzaid, I., Boissier, R., Nouhaud, F.-X., Albiges, L., Bernhard, J.-C., Ingels, A., Borchiellini, D., Kammerer-Jacquet, S., Rioux-Leclercq, N., Roupret, M., Acosta, O., De Crevoisier, R., Bensalah, K., Pignot, G., Ahallal, Y., Lebacle, C., … Larre, S. (2023). Application of Machine Learning Models to Predict Recurrence After Surgical Resection of Nonmetastatic Renal Cell Carcinoma. *European Urology Oncology*, *6*(3), 323–330. https://doi.org/10.1016/j.euo.2022.07.007

Klau, S., Hornung, R., Bauer, A., & Maintainer, J. H. (2023). *Package "prioritylasso" Type Package Title Analyzing Multiple Omics Data with an Offset Approach*.

Klau, S., Jurinovic, V., Hornung, R., Herold, T., & Boulesteix, A. L. (2018). Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2344-6

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis : techniques for censored and truncated data*. Springer.

Kleinbaum, D. G., & Klein, M. (n.d.). *Statistics for Biology and Health Survival Analysis A Self-Learning Text Third Edition*. http://www.springer.com/series/2848

Kleinbaum, D. G., & Klein, M. (2012). *Kaplan-Meier Survival Curves and the Log-Rank Test* (pp. 55–96). https://doi.org/10.1007/978-1-4419-6646-9_2

Kyung, M., Gilly, J., Ghoshz, M., & Casellax, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–412. https://doi.org/10.1214/10-BA607

Lawrence, M. S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Gabriel, S. B., Getz, G., Ally, A., Balasundaram, M., Birol, I., Bowlby, R., Brooks, D., Butterfield, Y. S. N., Carlsen, R., Cheng, D., Chu, A., Dhalla, N., Guin, R., Holt, R. A., … Pham, M. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, *517*(7536), 576–582. https://doi.org/10.1038/nature14129

Li, Z., & Sillanpää, M. J. (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. In *Theoretical and Applied Genetics* (Vol. 125, Issue 3, pp. 419–435). Springer Verlag. https://doi.org/10.1007/s00122-012-1892-9

Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A., & Tonda, A. (2019). Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3050-8

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., … Flicek, P. (2023). Ensembl 2023. *Nucleic Acids Research*, *51*(1 D), D933–D941. https://doi.org/10.1093/nar/gkac958

Martinez-Gutierrez, A. D., Cantú de León, D., Millan-Catalan, O., Coronel-Hernandez, J., Campos-Parra, A. D., Porras-Reyes, F., Exayana-Alderete, A., López-Camarillo, C., Jacobo-Herrera, N. J., Ramos-Payan, R., & Pérez-Plasencia, C. (2020). Identification of miRNA Master Regulators in Breast Cancer. *Cells*, *9*(7). https://doi.org/10.3390/cells9071610

Narod, S. A., & Foulkes, W. D. (2004). BRCA1 and BRCA2: 1994 and beyond. In *Nature Reviews Cancer* (Vol. 4, Issue 9, pp. 665–676). https://doi.org/10.1038/nrc1431

Ning, S., Xie, J., Mo, J., Pan, Y., Huang, R., Huang, Q., & Feng, J. (2023). Imaging genetic association analysis of triple-negative breast cancer based on the integration of prior sample information. *Frontiers in Genetics*, *14*. https://doi.org/10.3389/fgene.2023.1090847

Package "randomForestSRC" Title Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). (2023).

Package "survival." (2023). https://github.com/therneau/survival

Schmid, M., Wright, M., & Ziegler, A. (2015). *On the use of Harrell's C for clinical risk prediction via random survival forests*. http://arxiv.org/abs/1507.03092

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, *16*(4), 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tibshiranit, R. (1996). Regression Shrinkage and Selection via the Lasso. In *J. R. Statist. Soc. B* (Vol. 58, Issue 1).

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. In *Wspolczesna Onkologia* (Vol. 1A, pp. A68–A77). Termedia Publishing House Ltd. https://doi.org/10.5114/wo.2014.47136

Tormo, E., Adam-Artigues, A., Ballester, S., Pineda, B., Zazo, S., González-Alonso, P., Albanell, J., Rovira, A., Rojo, F., Lluch, A., & Eroles, P. (2017). The role of miR-26a and miR-30b in HER2+ breast cancer trastuzumab resistance and regulation of the CCNE2 gene. *Scientific Reports*, *7*. https://doi.org/10.1038/srep41309

Uhr, K., Prager-Van der Smissen, W. J. C., Heine, A. A. J., Ozturk, B., van Jaarsveld, M. T. M., Boersma, A. W. M., Jager, A., Wiemer, E. A. C., Smid, M., Foekens, J. A., & Martens, J. W. M. (2019). MicroRNAs as possible indicators of drug sensitivity in breast cancer cell lines. *PLoS ONE*, *14*(5). https://doi.org/10.1371/JOURNAL.PONE.0216400

Vega-Benedetti, A. F., Loi, E., Moi, L., Blois, S., Fadda, A., Antonelli, M., Arcella, A., Badiali, M., Giangaspero, F., Morra, I., Columbano, A., Restivo, A., Zorcolo, L., Gismondi, V., Varesco, L., Bellomo, S. E., Giordano, S., Canale, M., Casadei-Gardini, A., … Zavattari, P. (2019). Clustered protocadherins methylation alterations in cancer. *Clinical Epigenetics*, *11*(1). https://doi.org/10.1186/s13148-019-0695-0

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys*, *51*(6). https://doi.org/10.1145/3214306

Wold, S., Sjostrom, M., Eriksson, L., & Sweden˚˚, S. (n.d.). PLS-regression: a basic tool of chemometrics. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58). www.elsevier.comrlocaterchemometrics

Xiong, Jin. (2006). *Essential bioinformatics*. Cambridge University Press.

Zheng, J., Wang, H., Gao, Y., & Ai, Z. (2021). A Study on the Evaluation of a Risk Score of Osteonecrosis of the Femoral Head Based on Survival Analysis. *Journal of Arthroplasty*, *36*(1), 62–71. https://doi.org/10.1016/j.arth.2020.07.046

Zheng, Z., Xie, W., Chen, X., Wang, F., Huang, L., Li, X., Lin, Q., & Wong, K. C. (2022). Subclass-Specific Prognosis and Treatment Efficacy Inference in Head and Neck Squamous Carcinoma.

IEEE Journal of Biomedical and Health Informatics, *26*(8), 4303–4313. https://doi.org/10.1109/JBHI.2022.3168289

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. In *J. R. Statist. Soc. B* (Vol. 67, Issue 2). https://academic.oup.com/jrsssb/article/67/2/301/7109482

7. Anexo I

Tabla 9: Variables seleccionadas mediante el gráfico upset para BRCA. La frecuencia (columnas) indica el número de variables seleccionadas para cada combinación de modelos (filas). Comprende variables de los bloques X.Methyl (sufijo "_methyl"), X.Genes (sufijo "_gen"), X.miRNA (prefijo "hsa") y X.Clinical.

Frecuencia 1	**Frecuencia 2**	**Frecuencia 4**
age_at_diagnosis	hsa-miR-92a-1-5p ENSG00000185245_gen	ENSG00000204923_methyl ENSG00000103707_methyl hsa-miR-589-5p hsa-miR-769-5p
ENSG00000174243_methyl		
ENSG00000169919_methyl		
Frecuencia 7	**Frecuencia 14**	**Frecuencia 21**
ENSG00000073578_methyl ENSG00000146963_methyl ENSG00000135686_methyl ENSG00000124702_methyl ENSG00000165688_methyl ENSG00000120688_methyl hsa-miR-26a-5p	ENSG00000167632_methyl ENSG00000103319_methyl ENSG00000004897_methyl ENSG00000077235_methyl ENSG00000102786_methyl ENSG00000198231_methyl ENSG00000063177_methyl ENSG00000197785_methyl ENSG00000146247_methyl ENSG00000198951_methyl ENSG00000156873_methyl hsa-let-7d-5p hsa-let-7i-5p hsa-miR-148b-3p	ENSG00000166886_methyl ENSG00000077235_methyl ENSG00000188033_methyl ENSG00000198551_methyl ENSG00000168818_methyl ENSG00000163811_methyl ENSG00000176444_methyl ENSG00000111364_methyl ENSG00000124571_methyl ENSG00000181472_methyl ENSG00000204217_methyl ENSG00000159131_methyl ENSG00000205659_methyl ENSG00000198901_methyl ENSG00000126107_methyl ENSG00000100442_methyl ENSG00000168615_methyl ENSG00000141646_methyl ENSG00000021574_methyl hsa-miR-186-5p hsa-miR-340-3p

Tabla 10: Variables seleccionadas mediante el gráfico upset para HNSC. La frecuencia (columnas) indica el número de variables seleccionadas para cada combinación de modelos (filas). Comprende variables de los bloques X.cnv (sufijo "_cnv"), X.Genes (sufijo "_gen"), X.miRNA (prefijo "hsa") y X.Clinical.

<table>
<tr><th>Frecuencia 1</th><th>Frecuencia 2</th><th>Frecuencia 3</th></tr>
<tr><td>age</td><td rowspan="2">ENSG00000185182_cnv
ENSG00000277865_cnv</td><td rowspan="2">ENSG00000008300_gen
ENSG00000011083_gen
ENSG00000060656_gen</td></tr>
<tr><td>hsa.mir.3664</td></tr>
<tr><td>hsa.mir.552</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>hsa.mir.4746</td></tr>
<tr><td>ENSG00000080166_gen</td><td rowspan="4">hsa.mir.6503
hsa.mir.128.1</td><td rowspan="4"></td></tr>
<tr><td>ENSG00000010310_gen</td></tr>
<tr><td>ENSG00000180386_cnv</td></tr>
<tr><td>ENSG00000185966_cnv</td></tr>
<tr><td>ENSG00000095370_gen</td><td></td><td></td></tr>
<tr><th>Frecuencia 4</th><th>Frecuencia 8</th><th>Frecuencia 25</th></tr>
<tr><td>hsa.mir.6797
hsa.mir.135a.2
hsa.mir.7641.2
hsa.mir.3178</td><td>ENSG00000108947_gen
ENSG00000146574_gen
ENSG00000170522_gen
ENSG00000174652_gen
ENSG00000221823_gen
ENSG00000273729_gen
hsa.mir.412
ENSG00000253642_cnv</td><td>ENSG00000087266_gen
ENSG00000100225_gen
ENSG00000110075_gen
ENSG00000110375_gen
ENSG00000115446_gen
ENSG00000116285_gen
ENSG00000120159_gen
ENSG00000128891_gen
ENSG00000132639_gen
ENSG00000134569_gen
ENSG00000136247_gen
ENSG00000162241_gen
ENSG00000164920_gen
ENSG00000178776_gen
ENSG00000184208_gen
ENSG00000204967_gen
ENSG00000212901_gen
ENSG00000231312_gen
ENSG00000256518_gen
ENSG00000256937_gen
ENSG00000258914_gen
hsa.mir.5690
ENSG00000196912_cnv
ENSG00000275363_cnv
ENSG00000163202_cnv</td></tr>
<tr><th colspan="3">Frecuencia 25</th></tr>
<tr><td colspan="2">ENSG00000067141_gen
ENSG00000106927_gen
ENSG00000119537_gen
ENSG00000125657_gen
ENSG00000140350_gen
ENSG00000143183_gen
ENSG00000145348_gen
ENSG00000159556_gen
ENSG00000162367_gen
ENSG00000165521_gen
ENSG00000180855_gen
ENSG00000182149_gen</td><td>ENSG00000255893_gen
ENSG00000257207_gen
ENSG00000258484_gen
ENSG00000267374_gen
hsa.mir.521.1
ENSG00000223702_cnv
ENSG00000270872_cnv
ENSG00000272815_cnv
ENSG00000233707_cnv
ENSG00000232215_cnv
ENSG00000267247_cnv
ENSG00000244414_cnv
ENSG00000246250_cnv</td></tr>
</table>

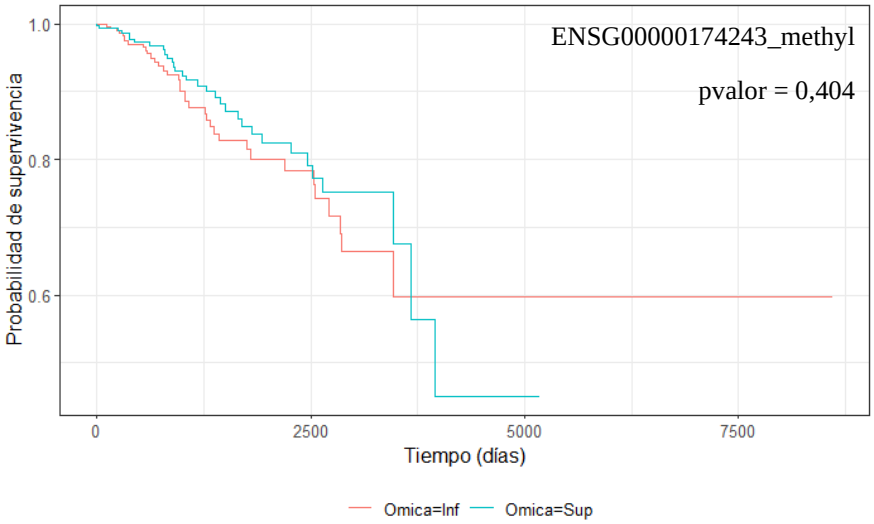

Figura 12: Curvas de supervivencia Kaplan Meier en función de ENSG00000174243_methyl en BRCA. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

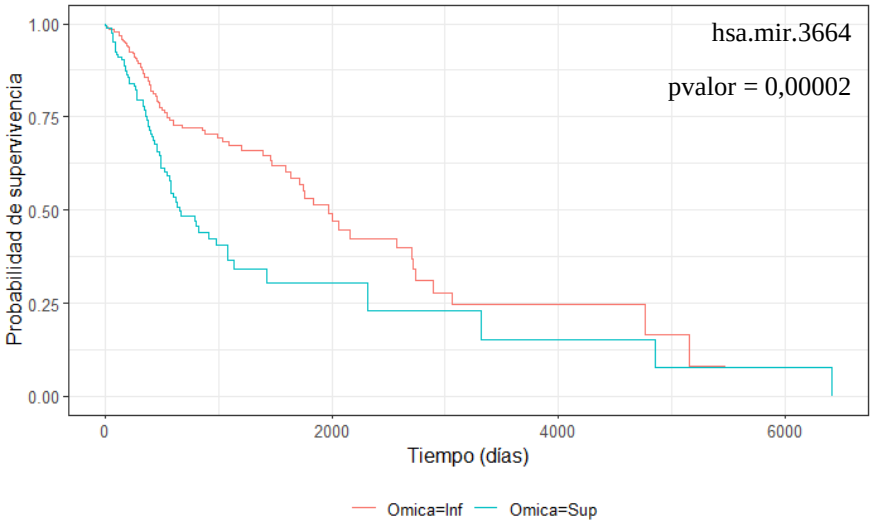

Figura 13: Curvas de supervivencia Kaplan Meier en función de hsa.mir.3664 en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

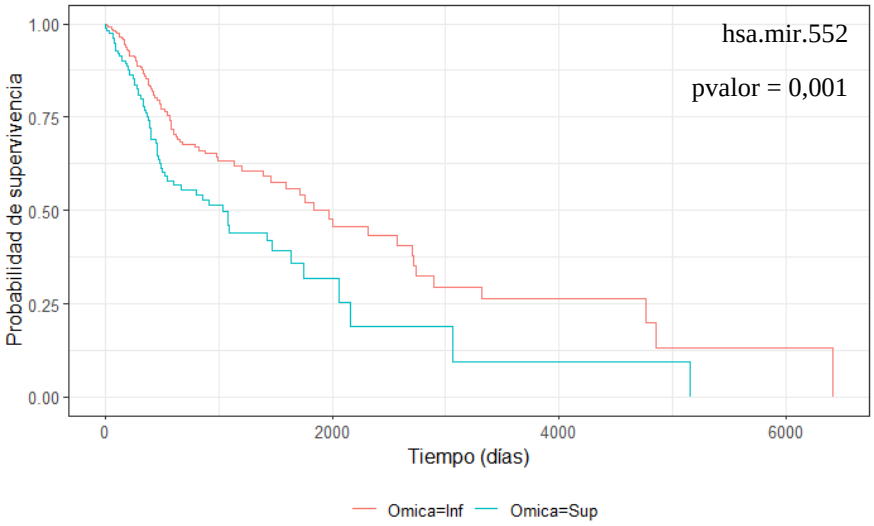

Figura 14: Curvas de supervivencia Kaplan Meier en función de hsa.mir.552 en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

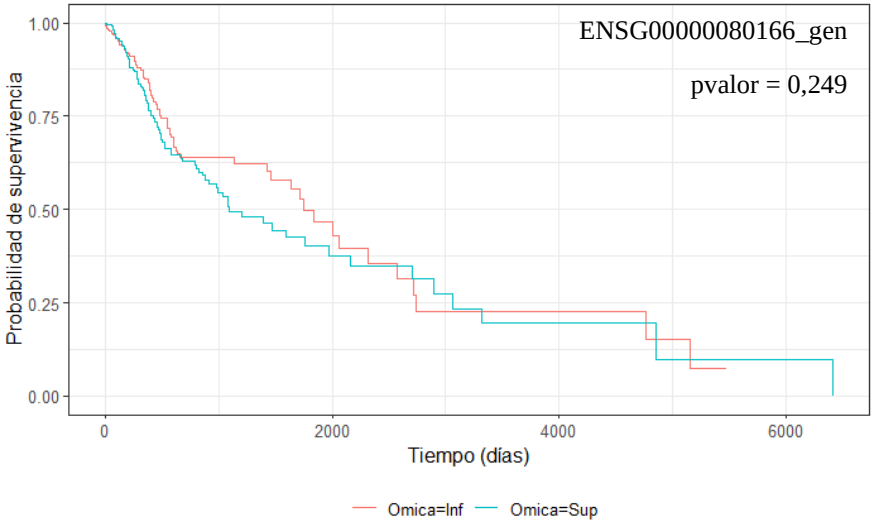

Figura 15: Curvas de supervivencia Kaplan Meier en función de ENSG00000080166_gen en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

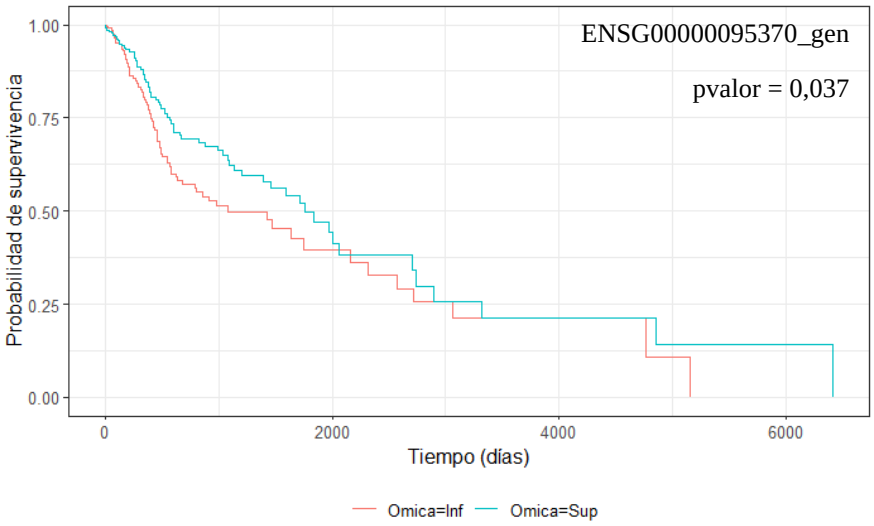

Figura 17: Curvas de supervivencia Kaplan Meier en función de ENSG00000095370_gen en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

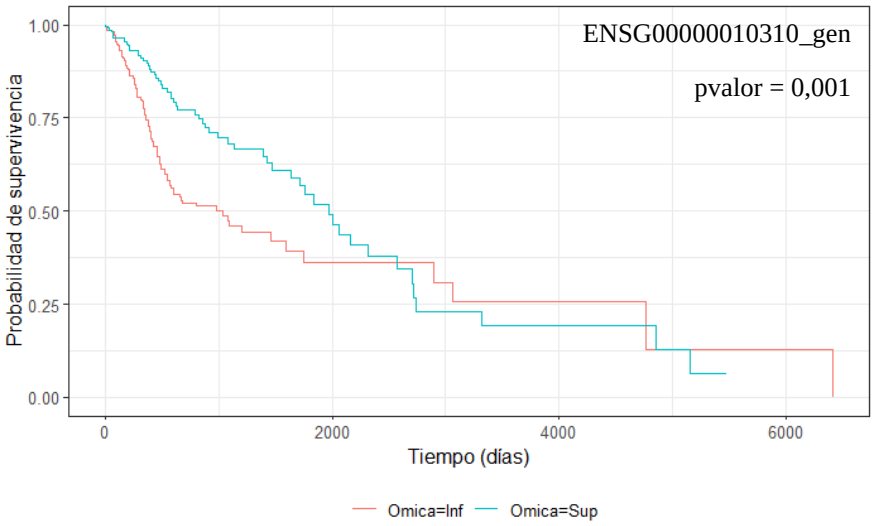

Figura 16: Curvas de supervivencia Kaplan Meier en función de ENSG00000010310_gen en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

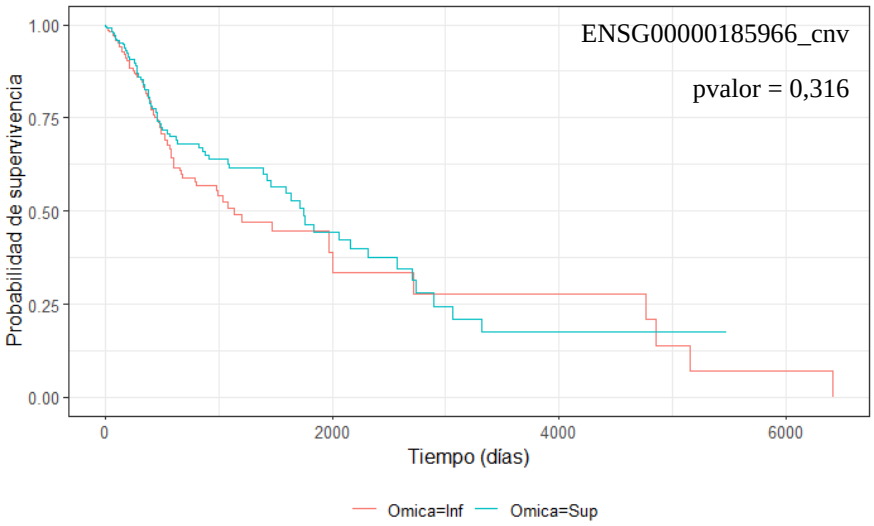

Figura 19: Curvas de supervivencia Kaplan Meier en función de ENSG00000185966_cnv en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

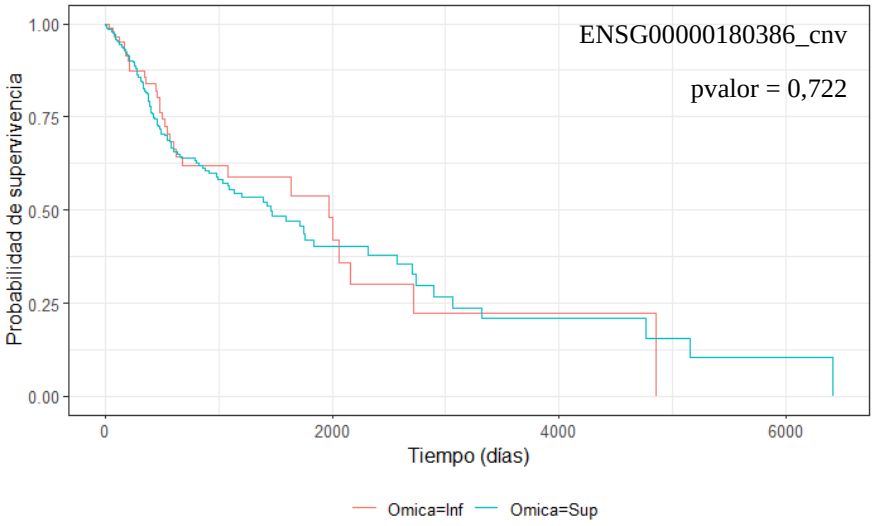

Figura 18: Curvas de supervivencia Kaplan Meier en función de ENSG00000180386_cnv en HNSC. El resultado del test estadístico logrank figura en la esquina superior derecha del gráfico. En rojo (Omica = Inf), estimación para sujetos con niveles de expresión inferiores a la mediana. En azul (Omica = Sup), estimación para sujetos con niveles de expresión superiores a la mediana.

Tabla 11: Categorización de las variables ómicas de los bloques X.Methyl (sufijo "_methyl") y X.Genes (sufijo "_gen") extraídas del gráfico upset para BRCA. El símbolo '-' equivale implica que dicha variable no tiene ninguna patología asociada.

Variable ómica	Patología Asociada	Variable ómica	Patología Asociada
ENSG00000174243_methyl	Otra	ENSG00000166886_methyl	Carcinoma de pecho, adenocarcinoma, y otras
ENSG00000169919_methyl	Otra	ENSG00000077235_methyl	Otra
ENSG00000185245_gen	Otra	ENSG00000188033_methyl	Otra
ENSG00000204923_methyl	Otra	ENSG00000198551_methyl	-
ENSG00000103707_methyl	Otra	ENSG00000168818_methyl	Otra
ENSG00000073578_methyl	Carcinoma de pecho, adenocarcinoma, carcinoma lobular y otras	ENSG00000163811_methyl	Cáncer de pecho
ENSG00000135686_methyl	-	ENSG00000176444_methyl	Otra
ENSG00000146963_methyl	Otra	ENSG00000111364_methyl	Otra
ENSG00000124702_methyl	-	ENSG00000124571_methyl	Otra
ENSG00000165688_methyl	Otra	ENSG00000181472_methyl	Otra
ENSG00000120688_methyl	-	ENSG00000204217_methyl	Otra
ENSG00000167632_methyl	Otra	ENSG00000159131_methyl	Otra
ENSG00000103319_methyl	-	ENSG00000205659_methyl	Otra
ENSG00000004897_methyl	Otra	ENSG00000198901_methyl	Carcinoma de pecho y otras
ENSG00000077235_methyl	-	ENSG00000126107_methyl	-
ENSG00000102786_methyl	-	ENSG00000100442_methyl	Otra
ENSG00000198231_methyl	-	ENSG00000168615_methyl	Otra
ENSG00000063177_methyl	Otra	ENSG00000141646_methyl	Carcinoma invasivo de pecho y otras
ENSG00000197785_methyl	Otra	ENSG00000021574_methyl	Otra
ENSG00000146247_methyl	Otra	ENSG00000156873_methyl	Otra
ENSG00000198951_methyl	Otra		

Tabla 12: Categorización de las variables ómicas de los bloques X.cnv (sufijo "_cnv") y X.Genes (sufijo "_gen") extraídas del gráfico upset para HNSC. El símbolo '-' equivale implica que dicha variable no tiene ninguna patología asociada.

Variable ómica	Patología Asociada	Variable ómica	Patología Asociada
ENSG00000080166_gen	Otra	ENSG00000165521_gen	-
ENSG00000277865_cnv	-	ENSG00000180855_gen	-
ENSG00000095370_gen	Otra	ENSG00000182149_gen	-
ENSG00000108947_gen	-	ENSG00000116285_gen	-
ENSG00000146574_gen	Otra	ENSG00000120159_gen	Otra
ENSG00000170522_gen	-	ENSG00000128891_gen	Otra
ENSG00000174652_gen	-	ENSG00000132639_gen	Otra
ENSG00000221823_gen	-	ENSG00000134569_gen	Otra
ENSG00000273729_gen	-	ENSG00000136247_gen	-
ENSG00000253642_cnv	-	ENSG00000162241_gen	-
ENSG00000087266_gen	Otra	ENSG00000164920_gen	-
ENSG00000100225_gen	Cáncer de cuello y cabeza	ENSG00000178776_gen	-
ENSG00000110075_gen	-	ENSG00000184208_gen	-
ENSG00000110375_gen	-	ENSG00000204967_gen	-
ENSG00000115446_gen	Otra	ENSG00000212901_gen	-
ENSG00000163202_cnv	Otra	ENSG00000231312_gen	-
ENSG00000010310_gen	-	ENSG00000256518_gen	-
ENSG00000180386_cnv	-	ENSG00000256937_gen	-
ENSG00000185966_cnv	-	ENSG00000258914_gen	-
ENSG00000185182_cnv	-	ENSG00000196912_cnv	-
ENSG00000008300_gen	-	ENSG00000275363_cnv	-
ENSG00000011083_gen	-	ENSG00000255893_gen	-
ENSG00000060656_gen	-	ENSG00000257207_gen	-
ENSG00000067141_gen	-	ENSG00000258484_gen	-
ENSG00000106927_gen	-	ENSG00000267374_gen	-
ENSG00000119537_gen	Cáncer de cuello y cabeza	ENSG00000223702_cnv	Otra
ENSG00000125657_gen	-	ENSG00000270872_cnv	-
ENSG00000140350_gen	-	ENSG00000272815_cnv	-
ENSG00000143183_gen	Otra	ENSG00000233707_cnv	-
ENSG00000145348_gen	Otra	ENSG00000232215_cnv	-
ENSG00000159556_gen	Otra	ENSG00000267247_cnv	-
ENSG00000162367_gen	Cáncer de cuello y cabeza	ENSG00000244414_cnv	Otra
ENSG00000246250_cnv	-		

8. Anexo II

El grado de relación del trabajo realizado con los Objetivos de Desarrollo Sostenible (Tabla 13, Anexo II) es elevado con el tercer objetivo, el cual aspira a garantizar una vida sana y promover el bienestar.

Esta estrecha relación se debe a que el análisis de supervivencia ha demostrado una de las muchas herramientas utilizadas para conducir a avances terapéuticos y preventivos y para la reducción de riesgos en pacientes con patologías como el cáncer mediante el reconocimiento de biomarcadores, o moléculas biológicas cuya expresión irregular o mutación puede llevar a características fenotípicas malignas. La identificación de sujetos con estas alteraciones genéticas ha resultado esencial en casos como el cáncer de mama o de ovarios, donde el descubrimiento de los marcadores genéticos BRCA1 y BRCA2 permiten actualmente anticipar el diagnóstico y aumentar la probabilidad de supervivencia (Narod & Foulkes, 2004).

Tabla 13: Objetivos de Desarrollo Sostenible

	Alto	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza**				x
ODS 2. **Hambre cero**				x
ODS 3. **Salud y bienestar**	x			
ODS 4. **Educación de calidad**				x
ODS 5. **Igualdad de género**				x
ODS 6. **Agua limpia y saneamiento**				x
ODS 7. **Energía asequible y no contaminante**				x
ODS 8. **Trabajo decente y crecimiento económico**				x
ODS 9. **Industria, innovación e infraestructuras**				x
ODS 10. **Reducción de las desigualdades**				x
ODS 11. **Ciudades y comunidades sostenibles**				x
ODS 12. **Producción y consumo responsables**				x
ODS 13. **Acción por el clima**				x
ODS 14. **Vida submarina**				x
ODS 15. **Vida de ecosistemas terrestres**				x
ODS 16. **Paz, justicia e instituciones sólidas**				x
ODS 17. **Alianzas para lograr objetivos**				x