

ESCUELA TÉCNICA SUPERIOR DE
INGENIERÍA INFORMÁTICA

Aprendizaje computacional y Análisis de Supervivencia
en pacientes de cáncer de mama.

Machine Learning and Survival Analysis
of breast cancer patients.

Realizado por
Naira María Chiclana García
Tutorizado por
José Manuel Jerez Aragonés

Departamento
Lenguajes y Ciencias de la Computación,
UNIVERSIDAD DE MÁLAGA

MÁLAGA, SEPTIEMBRE 2019



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERÍA DE LA SALUD

Aprendizaje computacional y Análisis de Supervivencia en pacientes de
cáncer de mama.

Machine Learning and Survival Analysis of breast cancer patients.

Realizado por

Naira María Chiclana García

Tutorizado por

José Manuel Jerez Aragonés

Departamento

Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, SEPTIEMBRE 2019

Fecha defensa:

Fdo.: El Secretario del Tribunal

Declaración de originalidad del trabajo

D^a.: Naira María Chiclana García , con DNI 44717497T , estudiante del Grado en Ingeniería de la Salud con mención en Bioinformática, de la Universidad de Málaga.

DECLARO QUE:

El Trabajo Fin de Grado denominado: *Aprendizaje automático y Análisis de Supervivencia en pacientes de cáncer de mama con terapia adyuvante*, es de mi autoría, inédito (no ha sido difundido por ningún medio, incluyendo internet) y original (no es copia ni adaptación de otra), no habiendo sido presentado anteriormente por mí ni por ningún otro autor o autora, ni en parte ni en su totalidad. Así mismo, se ha desarrollado respetando los derechos intelectuales de terceros, para lo cual se han indicado las citas necesarias en las páginas donde se usan, y sus fuentes originales se han incorporado en la bibliografía. Igualmente se han respetado los derechos de propiedad industrial o intelectual que pudiesen afectar a cualquier institución o empresa.

Para que así conste, firmo la presente declaración en Málaga, a de Septiembre de 2019 .

Fdo.: D^a

Resumen:

El cáncer es la segunda causa de muerte a nivel mundial actualmente y se prevé que los casos diagnosticados no dejarán de aumentar. El cáncer de mama es el que con más frecuencia se diagnostica en mujeres, y, tratado de forma correcta, tiene uno de los índices de supervivencia más altos. La efectividad y consecuente supervivencia del paciente depende de cada caso particular, por lo que es de vital importancia poder conocer el modelado de su supervivencia de forma previa.

La inteligencia artificial es un campo que no deja de crecer, y con él, su aplicación en el ámbito clínico (habiendo superado a los humanos en muchas tareas médicas basadas en evidencia). En análisis de supervivencia, emergió el primer estimador de riesgo estable en 1989 (*Gail Model*) basado en métodos estadísticos. Desde entonces, son abundantes las versiones del mismo que se han realizado y escaso el uso del aprendizaje automático en estas. Los métodos estadísticos hacen muy sencillo evaluar las relaciones no lineales entre variables y el impacto que estas causan sobre la variable a predecir (evento). Sin embargo, por no ser más que ecuaciones matemáticas, presentan carencias que limitan la calidad de sus resultados.

En este proyecto proponemos algoritmos de aprendizaje automático supervisado como alternativa a los métodos estadísticos para predecir y modelar la supervivencia, habiendo probado su mejor funcionamiento. Sabiendo que la mayor dificultad con la que se lida en datos clínicos es su cantidad y calidad, además de replicarlos, hemos realizado un extenso paso previo para garantizar la calidad de los mismos. Sin ignorar la facilidad en análisis multivariante que el método estadístico de Cox ofrece, lo hemos usado junto con otros métodos y tests estadísticos para encontrar el mejor conjunto de datos posible con el que entrenar nuestro modelo. Entre las variables a evaluar, teníamos las estratificaciones usadas de forma estándar en la práctica médica frente a estratificaciones definidas por métodos de aprendizaje automático no supervisado, las cuales, gracias a la capacidad de ajustarse a la particularidad de cada conjunto de datos, han demostrado ser más significativas y precisas para explicar el riesgo. Finalmente, hemos realizado una interfaz donde introduciendo los datos de la paciente podemos observar su curva de supervivencia ajustada en el tiempo para cada posible tratamiento de hormonoterapia. Hemos realizado este estudio trabajando paralelamente con dos conjuntos de datos similares, en uno de ellos el evento es la recurrencia y en otro la muerte.

Palabras clave: cancer de mama, análisis de supervivencia, predicción, inteligencia artificial, método estadístico, aprendizaje automático supervisado, aprendizaje automático no supervisado.

Abstract:

Nowadays cancer is the second leading cause of death worldwide, and the number of cases that will be diagnosed is expected to grow. Breast cancer is the most commonly diagnosed cancer in women, but with appropriate treatment, it has one of the highest survival rates among all other types of cancer. The effectivity of treatment and the consequent survival of the patient depend on the personal circumstances of each person, which is why being able to estimate a single woman's survival curve in advance is of vital importance.

The field of artificial intelligence is growing exponentially, especially in the healthcare environment, where algorithms are becoming better than humans experts in many tasks (particularly regarding evidence-based medicine). In survival analysis, the first reliable risk prediction model emerged in 1989 (Gail Model), which is based on statistical methods. From then on, a large number of versions have been made, but the use of machine learning in them is scarce. Statistical methods evaluate non-linear relations between predictor features incredibly easy, as well as measuring the impact of them on the event class. Nevertheless, because they are only mathematical equations, the accuracy of their performance is limited.

In this project, we suggest the use of supervised machine learning methods as an alternative to statistical methods in the prediction of survival curves, and moreover, we prove that they perform better. Knowing that the main difficulty faced when dealing with clinical data is the shortage and poor quality of it, in addition to obtaining a data replication method, we have made a huge previous step for guaranteeing the quality of the features with which the model will be trained. Not ignoring the easiness in a multivariate analysis that Cox provides, we have used it together with other statistical tests and methods to find the best possible combination of features. Among the features to evaluate, we con-

sidered the standard stratifications used in clinical practice and stratifications found out by unsupervised machine learning. The latter stratifications have shown to be better in explaining the risk factor, thanks to their ability to fit the peculiarities of each dataset. Lastly, we have carried out the design of an interface, where giving the patient characteristics as inputs, we can see a fitted survival curve for each possible hormonotherapy treatment. We have concurrently worked with two similar datasets, one in which the event refers to death, and another one in which the event means recurrence.

Keywords: breast cancer, survival analysis, prediction, artificial intelligence, statistical method, supervised machine learning, unsupervised machine learning.

Índice

1. Introducción	1
1.1. Motivación	1
1.2. Estado del arte: Aprendizaje computacional y análisis de supervivencia en el ámbito clínico	5
1.2.1. Aprendizaje computacional y análisis de supervivencia en el cáncer de mama	5
1.3. Objetivos	7
1.4. Estructura de la memoria	8
2. Contexto Análisis de Supervivencia	9
2.1. Preliminares matemáticos	9
2.1.1. Censura	9
2.1.2. Análisis de supervivencia discreto	12
2.1.3. Análisis de supervivencia continuo	13
2.1.4. Estimadores estadísticos	14
2.2. Métodos estadísticos para estimar supervivencia	16
2.2.1. Métodos estadísticos no-paramétricos	17
2.2.2. Cox Proportional Hazards Model	17
2.2.3. Índice de concordancia	19
3. Contexto Aprendizaje Computacional	20
3.1. Aprendizaje no supervisado	20
3.1.1. Métricas clusters	21
3.1.2. K-means clustering	22
3.1.3. Hierarchical clustering	23
3.2. Modelos de Clasificación Probabilística	24
3.2.1. Métricas para modelos de clasificación	25
3.2.2. Deep Learning (DL)	27
3.2.3. Random Forest (RF)	29
3.2.4. Gradient Boosting Machine (GBM)	30
3.2.5. Extreme gradient boosting (XGBoost)	30

3.2.6. Generalized Linear Models (GLM)	30
3.3. K-Fold Cross Validation	31
4. Conjuntos de datos	33
4.1. Hospital	33
4.1.1. Atributos	33
4.1.2. Distribución y valores atípicos	34
4.1.3. Valores eliminados	36
4.2. TCGA	37
4.2.1. Atributos	37
4.2.2. Distribución y valores atípicos	38
4.2.3. Valores eliminados	40
5. Métodos	42
5.1. Ingeniería de características	42
5.1.1. Creación de variables	43
5.1.2. Análisis y elección de variables	47
5.1.3. Correlación e hipótesis	51
5.1.4. Conjuntos finales	53
5.2. Estudio de los efectos de la quimioterapia y hormonoterapia	55
5.3. Propuesta de tratamiento hormonal basado en la supervivencia	57
5.3.1. Replicación datos	57
5.3.2. Usando <i>mlr</i> : Elección del tipo de predictor	60
5.3.3. Usando <i>h2o</i> : Análisis exploratorio de modelos clasificación	62
5.3.4. Decisión de hormonoterapia en base a la supervivencia predicha	63
6. Resultados	64
7. Conclusión	66
7.1. Trabajos Futuros	68
8. Recursos	69
8.1. Software	69
8.2. Hardware	69

Índice de figuras

1.	Estadísticas globales de porcentaje de diagnosis y número de muertes por tipo de cancer en 2017. [4]	2
2.	Incidencia en cancer de mama por pais entre mujeres de todas las edades en 2018. [6]	3
3.	Esquema general de un estudio de supervivencia. [34]	9
4.	Censura derecha de tipo III. [35]	11
5.	Esquema de aprendizaje de un algoritmo supervisado. [36]	20
6.	Esquema de aprendizaje de un algoritmo no supervisado. [36]	20
7.	Cohesión y separación clusters de forma gráfica. [36]	21
8.	Funcionamiento de <i>Hierarchical Clustering</i> aglomerativo y divisivo. [37]	24
9.	Matriz de confusión.	25
10.	Arquitectura de una red neuronal artificial. [39]	27
11.	Neurona biológica (izquierda) y su modelo matemático (derecha). [40]	27
12.	Tipos de funciones de activación. [36]	28
13.	Arquitectura de un árbol de decisión. [41]	29
14.	Esquema de iteraciones de una máquina Gradient Boosting. [42]	30
15.	Cross validation 5 Folds. [45]	32
16.	Histogramas distribución variables numéricas datos Hospital.	34
17.	Diagramas de cajas variables numéricas datos Hospital.	34
18.	Frecuencia variables categóricas datos Hospital.	35
19.	Diagrama de tarta variable <i>evento</i> datos Hospital.	36
20.	Histogramas distribución variables numéricas datos TCGA.	38
21.	Diagramas de cajas variables numéricas datos TCGA.	39
22.	Frecuencia variables categóricas dataset TCGA.	39
23.	Diagrama de tarta variable <i>evento</i> datos TCGA.	40
24.	Grafico dispersión variables a clusterizar Hospital.	44
25.	Grafico dispersión variables a clusterizar TCGA.	44
26.	Dendogramas <i>hierarchical clustering</i> Hospital	46
27.	Histogramas clusters Hospital	46
28.	Histogramas clusters TCGA	47
29.	Valores de correlación <i>Spearman</i> a dataset final Hospital.	52

30.	Valores de correlación <i>Spearman</i> a dataset final TCGA.	53
31.	Distribución evento en el tiempo en conjunto <i>Hospital</i>	59
32.	Distribución evento en el tiempo en conjunto <i>TCGA</i>	59
33.	Previsualización interfaz.	63
34.	Curvas de supervivencia conjunto <i>Hospital</i> toda la población	70
35.	Curvas de supervivencia conjunto <i>Hospital</i> (1).	71
36.	Curvas de supervivencia conjunto <i>Hospital</i> (2).	72
37.	Curvas de supervivencia conjunto <i>Hospital</i> (3).	73
38.	Curvas de supervivencia conjunto <i>TCGA</i> (2).	73
39.	Curvas de supervivencia conjunto <i>TCGA</i> toda la población	74
40.	Curvas de supervivencia conjunto <i>TCGA</i> (1).	75
41.	Curvas de supervivencia conjunto <i>TCGA</i> (2).	76

Índice de cuadros

1.	Resumen conjuntos de datos	41
2.	Agregaciones variables numéricas	43
3.	Análisis calidad variables Hospital a partir de $p - value$ y 95 %CI	49
4.	Test de riesgos proporcionales de Cox conjuntos Hospital.	49
5.	Tests significancia global variables Cox PH model conjuntos Hospital	50
6.	Análisis calidad variables TCGA a partir de $p - value$ y 95 %CI	51
7.	Test de riesgos proporcionales de Cox conjuntos TCGA.	51
8.	Tests significancia global variables Cox PH model conjuntos TCGA	51
9.	Pruebas de hipótesis correlaciones Hospital	52
10.	Pruebas de hipótesis correlaciones TCGA	53
11.	Ajuste supervivencia Kaplan Meier del conjunto Hospital.	53
12.	Ajuste supervivencia Kaplan Meier del conjunto TCGA.	53
13.	Índice de impacto, error estándar, índice de riesgo, p-value e intervalos de confianza 95 % usando el modelo de Cox en conjunto Hospital.	54
14.	Índice de impacto, error estándar, índice de riesgo, p-value e intervalos de confianza 95 % usando el modelo de Cox en conjunto TCGA.	54
15.	Tasa de muerte por grupo riesgo según exposición a quimioterapia conjunto Hospital	56
16.	Valoración utilidad exposición a quimioterapia	56
17.	Aceptación de tratamientos hormonoterapia	56
18.	Estado conjuntos antes y después de la replicación.	60
19.	Resultados clasificadores <i>Survival</i> usando <i>mlr Hospital</i>	61
20.	Resultados clasificadores <i>Survival</i> usando <i>mlr TCGA</i>	61
21.	Resultados clasificadores <i>Clasificación</i> usando <i>mlr Hospital</i>	62
22.	Resultados clasificadores <i>Clasificación</i> usando <i>mlr TCGA</i>	62
23.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>Hospital</i> con variables estandar y sin replicar.	64
24.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>Hospital</i> replicado y sin balancear	64
25.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>Hospital</i> replicado y balanceado.	64

26.	Modelo de mayor precisión para conjunto <i>Hospital</i> : <i>Deep Learning</i> con los datos balanceados, hiperparámetros. <i>Learning rate=0.005, 2611 epochs.</i>	64
27.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>TCGA</i> con variables estándar y sin replicar.	65
28.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>TCGA</i> replicado y sin balancear.	65
29.	Resultados con desviación estándar de predictores <i>Clasificación</i> usando <i>h2o</i> en <i>TCGA</i> replicado y balanceado.	65
30.	Modelo de mayor precisión para conjunto <i>TCGA</i> : <i>Extreme Gradient Boosting</i> con los datos sin balancear, hiperparámetros. <i>160 epochs.</i>	65
31.	Valores curvas de supervivencia conjuntos <i>Hospital</i> .	70
32.	Valores curvas de supervivencia conjuntos <i>TCGA</i> .	74

1. Introducción

1.1. Motivación

El **cancer** es la segunda causa de muerte a nivel mundial, con una estimación de 9.6 millones de muertes en 2018 (1 de cada 6 muertes) , y los casos de diagnosis y muertes por el mismo no dejan de aumentar cada año. [1]. Se espera que para 2030 los casos de cancer detectados por año aumenten en 23.6 millones. [3]

Así como el impacto social, el impacto económico global también es altísimo, con una cifra media de \$895 billones gastados anualmente, mientras que a su vez, el coste de los tratamientos necesarios para su tratamiento sigue creciendo un 10% cada año. [2]

Los cánceres más comunes registrados en los últimos años (en orden descendente) han sido el cancer de mama, el cancer de próstata, y el cancer de colon y rectal. [4]

Como podemos ver en la Figura 1 con datos estadísticos del año 2017 , el **cancer de mama** se encuentra, con significativa diferencia, el primero de todos los cánceres diagnosticados (alcanzando la cifra de un 0,21 % de los casos). Es la enfermedad maligna no dermatológica más frecuente y la que más muertes causa en las mujeres de todo el mundo, con más de 1 millón de nuevos casos diagnosticados cada año [5]. Lo realmente interesante es que, a pesar de ser este el más padecido, se encuentra en la quinta posición en el ranking de muertes por cancer.

Concretamente, España se encuentra entre los países del mundo con más casos de cancer de mama (como podemos ver en la Figura 2) y donde la incidencia más ha aumentado en los últimos años. [6]

Los tratamientos para el cancer existentes actualmente tienen fuertes y dañinos efectos secundarios, efectos que los pacientes no suelen estar dispuestos a sufrir a menos que tengan la esperanza de vivir lo suficiente para disfrutar de los beneficios que estos les aportarían. Hasta que curas definitivas para el cancer empiecen a aparecer, es primordial saber cómo va a ser la esperanza de vida de cada paciente, es decir, la **predicción o análisis de supervivencia**. Esta correcta predicción, además de ayudar a los pacientes y sus familiares a tomar una decisión acerca del tratamiento, haría más efectivo el uso de recursos médicos y económicos. Haciendo uso de la inteligencia artificial, además de

la supervivencia, también podríamos estimar el tratamiento más correcto en cada caso, con el objetivo de, gracias a estrategias personalizadas, incrementar la supervivencia en pacientes de alto riesgo a la vez que decrecen costes y complicaciones en mujeres de bajo riesgo.

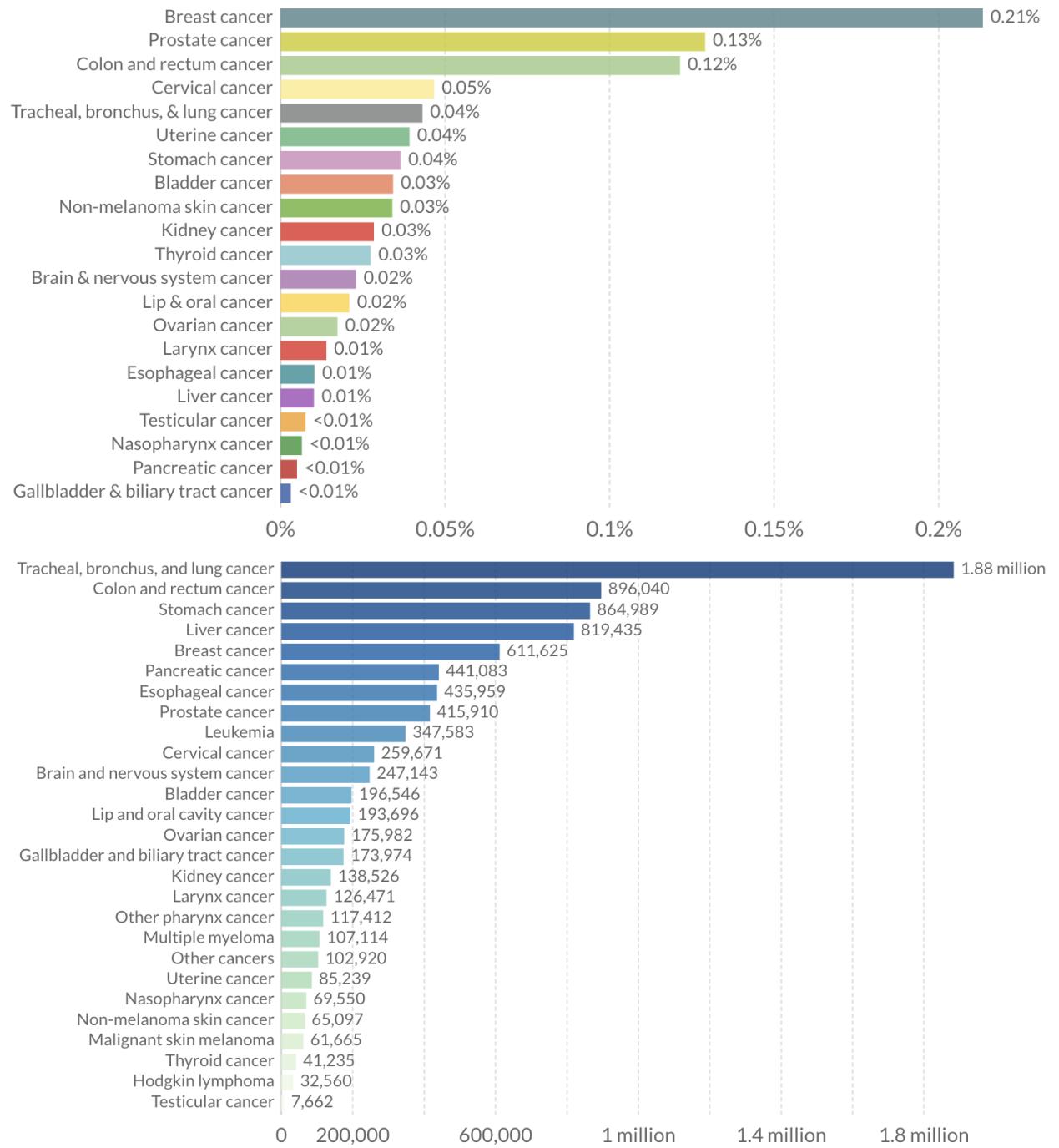


Figura 1: Estadísticas globales de porcentaje de diagnosis y número de muertes por tipo de cancer en 2017. [4]

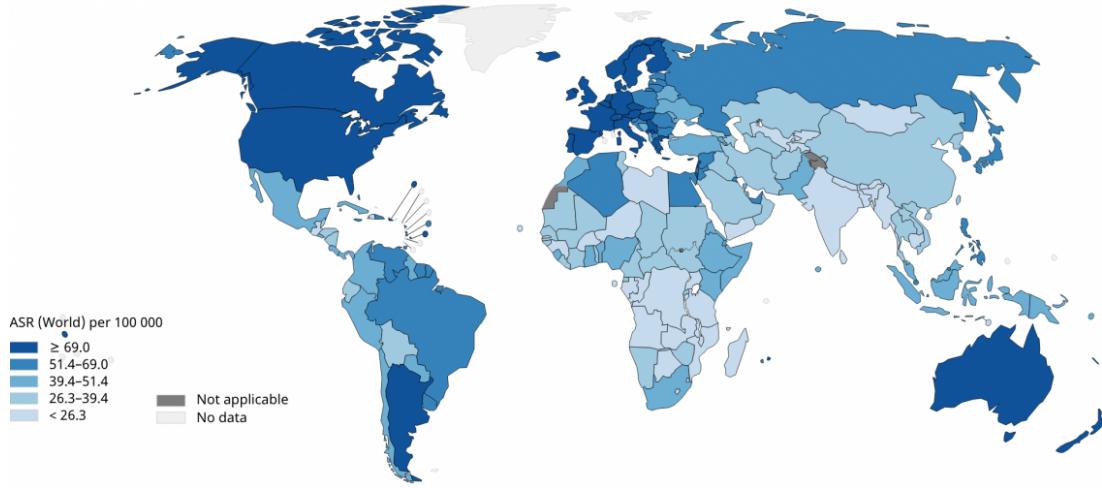


Figura 2: Incidencia en cancer de mama por pais entre mujeres de todas las edades en 2018. [6]

La **inteligencia artificial (IA)** está cada vez más presente en nuestras vidas, cambiando radicalmente todas las áreas de nuestra sociedad. Cada vez más, prácticamente cualquier tarea humana repetitiva puede ser automatizada, y no solo automatizada, sino, en muchos casos, puede mejorar con creces la capacidad humana cuando de extraer relaciones en alto número de dimensiones o de tomar decisiones basadas en grandes cantidades de datos se trata.

Uno de los campos en los que la inteligencia artificial más nos ha beneficiado ha sido la medicina, área en la que más se ha invertido de entre todos los sectores (\$1.3 billones en 2017, \$320 millones tan solo en el primer cuatrimestre de 2018, y continua creciendo [7]). Hablamos de lo que se conoce como *Inteligencia Artificial Médica*.

La esencia de la medicina basada en evidencia es tomar decisiones basadas en los datos pasados, tarea en la que, como acabamos de comentar, la IA puede superar con creces a los humanos. Estas máquinas, no solo pueden tomar decisiones y aprender como lo haría un humano, sino que además, pueden hacerlo con más evidencia, gracias a que en solo unos instantes pueden exponerse y aprender más casos de lo que un médico lo haría en toda su vida [8]. Algunos ejemplos donde estos algoritmos han superado la precisión y especificidad de los humanos los encontramos mayormente en tests diagnósticos a partir de imágenes, como la correcta detección de lesiones o cancer en la piel[9], la identificación de tuberculosis pulmonar [10], etc...

“AI can’t replace doctors. But it can make them better.” - Rahul Parikh

“We need to let the machine do what it does well — such as ingesting a whole lot of scientific papers and organizing them — and let the clinician make the final decision on what should be done to treat a particular patient.” [11]

Es evidente la gran efectividad que la IA supone en la medicina y diagnóstico. El aprendizaje computacional cuenta con técnicas capaces de extraer complejas asociaciones que difícilmente una ecuación podría hacer. Sin embargo, tradicionalmente se ha abordado la tarea de extraer conclusiones basadas en datos con métodos estadísticos, los cuales caracterizan patrones mediante ecuaciones matemáticas.

Concretando en el problema médico al que nos enfrentamos en este proyecto (la predicción de supervivencia), comúnmente se usan métodos estadísticos semi-paramétricos y no paramétricos. Estos cuentan con importantes carencias y riesgos que pueden llevarnos a estimadores de resultados inválidos e inconsistentes, como veremos más adelante. A pesar de ser tan prometedora la posible mejora, son solo algunos los casos en los que recientemente se han empezado a sustituir los modelos estadísticos por los de aprendizaje automático para esta causa.

Además, sabemos que el problema más difícil con el que se lida en los datos clínicos son aquellos censurados, donde el evento de interés no se observa 2.1.1. El alto porcentaje de estos datos frente a los casos si observados (normalmente 80 %-90 % frente a un 10 %-20 %) puede crear sesgo en el modelo. Los modelos de aprendizaje automático están más preparados para lidar con estos datos censurados de los que lo están los estadísticos mencionados previamente, teniendo la posibilidad de alcanzar resultados más precisos. Una de las grandes barreras que encuentra la IA en el ámbito médico es la falta de datos y la poca calidad de estos.

En este proyecto intentaremos demostrar cómo, para dos conjuntos de datos de supervivencia en cancer de mama distintos, trabajando su calidad y cantidad, y usándolos para entrenar modelos de aprendizaje automático (también no paramétricos) podemos obtener resultados de precisión que superen notablemente a los que nos daría un método estadístico.

1.2. Estado del arte: Aprendizaje computacional y análisis de supervivencia en el ámbito clínico

Como ya hemos visto en el punto anterior, las aplicaciones del aprendizaje computacional en este campo no dejan de sorprendernos con nuevos avances cada día, transformando a gran velocidad el paradigma médico. Son muchas las aplicaciones que ya se le están dando, desde investigación básica hasta aplicaciones clínicas. Entre todas ellas, destacan dos grandes campos: la diagnosis y el descubrimiento y creación de fármacos.

Algunos sorprendentes ejemplos son compañías como *Exscientia* [13] que llevan a cabo su descubrimiento de fármacos de forma *AI – Driven*, o el algoritmo desarrollado en 2019 que pudo detectar un lunar imperceptible al ojo humano como maligno con 100 % de confianza [14].

El análisis de supervivencia es muy popular debido a su simplicidad, y, además de usarse en la estadística médica, tiene una extensa aplicación en otros campos como la economía, educación, biología o industria. El impago de un crédito, cuando dejará de funcionar un electrodoméstico o el abandono de los estudios por parte de un estudiante son algunos de los eventos que se estudian con esta técnica. [15]

La aplicación principal que se le dá en la medicina es para analizar los eventos de interés: muerte, recaída, reacción adversa a un fármaco o el desarrollo de una nueva enfermedad. Para todos estos casos, se puede modelar y saber el riesgo de que el evento tenga lugar en un rango de tiempo desde semanas a años según el caso. [16]

1.2.1. Aprendizaje computacional y análisis de supervivencia en el cáncer de mama

Son muchos los predictores que se han desarrollado para estimar el riesgo individual de mujeres con cáncer de mama. Algunos de los más conocidos y aceptados de forma estándar son *Gail Model* por Gail MH.en 1989 [17], ,*Colditz Model* en 1996 , *Tyrer-Cuzick model* en 2004 por Tyrer J. o *BCRAT (Breast Cancer Risk Assesment Tool)* en 2007 como una de las numerosas modificaciones que existen del modelo de *Gail* . Estos modelos son Modelos de Predicción de Riesgo, es decir, herramientas estadísticas que estiman la probabilidad de que un individuo con factores de riesgo específicos desarrolle una condición futura (evento) en un periodo de tiempo específico.

Gail es, sin duda, del que mayor fama y uso goza. Su estudio implicó a 280,000 mujeres de entre 35 y 74 años de edad. El modelo original estimaba el riesgo de desarrollar cancer en los siguientes 5 años y en el tiempo de vida en mujeres blancas. Las posteriores adaptaciones han incluido mujeres africanas, hispánicas y asiáticas. Son muchos los estudios que podemos encontrar que validan este modelo con diferentes conjuntos de datos, por ejemplo [18], [19], [20],[21],[22],[23],[24],[25] o [26]. A lo largo de los estudios se han encontrado diversas carencias para casos particulares de variables de riesgo no incluidas en el modelo original, el cual cuenta con muy pocas y conlleva el posible riesgo de descalibrarse en poblaciones variadas. Se han hecho casi tantas versiones del mismo como estudios se han realizado. Estas versiones suelen medir su precisión con el Índice de Concordancia ($C - Index$) 2.2.3. Los valores alcanzados de este indice rondan entre 0,5 y 0,8, raramente sobrepasándolo.

Estas nuevas versiones de modelos de riesgo se han realizado en su gran mayoría usando el modelo de riesgos proporcionales *Cox Proportional Hazards Model*, u otros métodos estadísticos similares semi-paramétricos o no paramétricos como *Kaplan-Meier*, *Exponential Weibull*, etc...

El modelo de Cox infiere y representa fácilmente los efectos y relaciones no lineales entre la variable de salida (supervivencia) y el resto (características clínicas). A pesar de su simplicidad, una de las fuertes carencias que presenta, (sin olvidar la de datos censurados expuesta en el apartado anterior) es que depende de la suposición de que las variables son independientes del tiempo, lo cual no suele ser cierto en muchos de estos casos. Además, tiende a predecir la supervivencia para grupos de pacientes con riesgo similar, mientras que métodos de aprendizaje computacional pueden hacerlo de forma individual y personalizada. Las técnicas de aprendizaje automático también reducen el riesgo potencial de no tener los requisitos suficientes para el modelo, lo cual evita estimadores de resultados inválidos e inconsistentes.

Los estimadores de desarrollo futuro de la enfermedad son esenciales ya que permiten aplicar tratamientos preventivos como fármacos o cirugía. Además, estos calculadores son muy rápidos de usar e interpretar. Pero, como hemos expuesto en el punto anterior, conocer la supervivencia y como a este le afecta el uso de un fármaco u otro una vez ya se padece la enfermedad también es muy importante. A pesar de ello, los estudios centrados

en esta circunstancia se reducen notablemente frente a los anteriores.

En los estudios de predicción de recurrencia, aunque, como en el caso anterior priman los métodos estadísticos (como [27], [28], [29] o [30]). Con algo de dificultad podemos encontrar algunos estudios que han aplicado aprendizaje computacional (como [31] o [32]), y los resultados de precisión de estos aumentan de forma notable.

En este proyecto, analizaremos el comportamiento de modelos de aprendizaje automático para predecir la muerte y recurrencia frente a métodos estadísticos, mientras que sacaremos provecho de las ventajas que estos últimos ofrecen para seleccionar un conjunto de variables óptimo con el que entrenar a nuestro modelo, tarea en la que han dado muy buenos resultados a lo largo del tiempo.

1.3. Objetivos

Por lo visto en los puntos anteriores, el objetivo principal de este proyecto consistirá en encontrar un modelo de aprendizaje computacional óptimo para predecir la supervivencia individualizada de pacientes de cancer de mama así como el correspondiente tratamiento más adecuado. Este objetivo principal lo podemos dividir en las siguientes tareas:

1. Analizar los efectos de la quimioterapia frente a la hormonoterapia para descartar esta primera como efectiva en el tratamiento del cancer de mama. 5.2
2. Proponer nuevas estratificaciones de variables numéricas. Se realizarán con métodos de aprendizaje automático no supervisado que dividan a los pacientes según el riesgo en cada conjunto de datos particular. Comprobaremos su mayor significancia frente a las agrupaciones standar. 5.1
3. Usando técnicas de aprendizaje automático, intentar encontrar el modelo predictor de supervivencia más preciso. Usaremos para este el conjunto de variables significativo encontrado en el paso anterior. Compararemos este con los métodos estadísticos convencionales, esperando encontrar una mejora.
4. Proponer el tratamiento hormonal más adecuado, es decir, aquel que maximice la supervivencia. Lo haremos usando el mejor predictor encontrado en punto anterior, y a través de una interfaz sencilla de usar. 5.3.4

1.4. Estructura de la memoria

- Contexto: En esta sección se explican las bases teóricas y matemáticas de los conceptos principales con los que trabajaremos a lo largo del proyecto.
- Conjuntos de datos: Aquí podremos ver los dos conjuntos de pacientes con los que vamos a trabajar y de los que extraeremos conclusiones. Su estado inicial, el procesamiento (*Data Mining*) que se ha llevado a cabo, y el conjunto final de datos a analizar.
- Métodos: Esta sección contiene las fases principales de un proyecto básico de *Machine Learning* partiendo de unos datos ya limpios. Los datos son analizados, transformados y reducidos. Una exploración sobre distintos algoritmos y su comportamiento es llevada a cabo, y finalmente, predictores finales son elegidos y usados para conseguir nuestro propósito final.
- Resultados: Aquí se exponen los resultados finales obtenidos en distintos algoritmos, diferenciando entre los dos tipos de predicción realizada (recurrencia y muerte).
- Conclusiones: Finalmente, la sección final contiene las conclusiones, la discusión tan discutida sobre la mejora de los modelos generados con respecto a los tradicionales y los trabajos futuros.

2. Contexto Análisis de Supervivencia

El análisis de supervivencia tiene como objetivo modelar el tiempo que transcurre hasta que ocurre un determinado evento, relacionando el resultado de un paciente con sus variables biológicas descriptivas asociadas. Incorporan el fenómeno conocido como censura y 2.1.1 dan como resultado la estimación de curvas de supervivencia.

2.1. Preliminares matemáticos

2.1.1. Censura

Se trata de observaciones con información incompleta. Como hemos comentado en la sección 1.1, es uno de los retos más difíciles a enfrentar cuando tratamos con datos clínicos.

La observación de cada paciente se inicia al diagnóstico ($t = 0$) y continua hasta el evento o hasta que el tiempo de seguimiento se interrumpe. Se dice que un tiempo o paciente está censurado cuando el tiempo exacto del evento no se observa, y por lo tanto, no sabemos el momento exacto de ocurrencia del mismo. En este proyecto entendemos el evento como recurrencia en el caso del dataset *Hospital* y muerte en el caso del dataset *TCGA*.

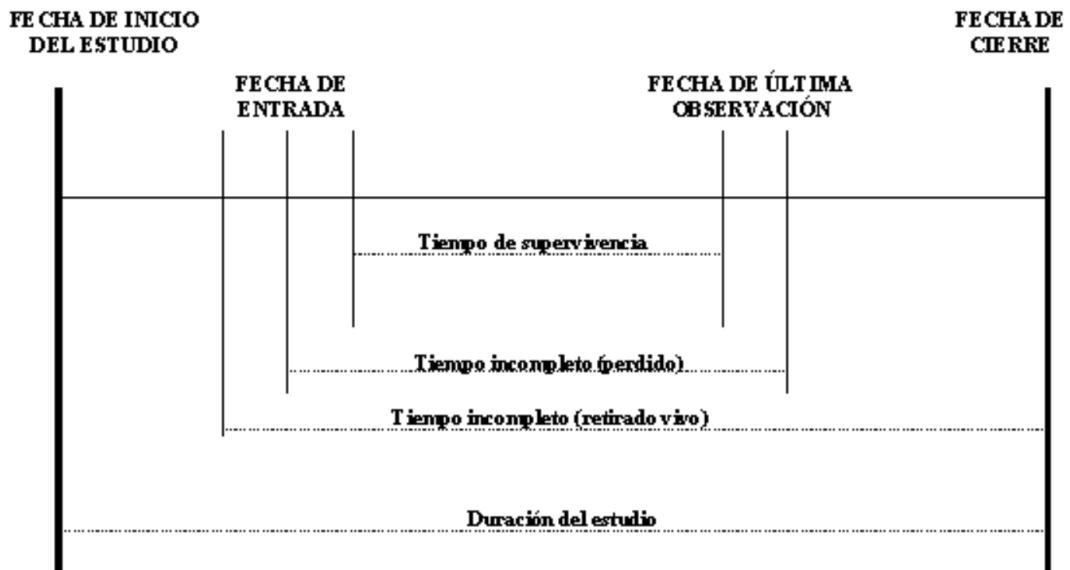


Figura 3: Esquema general de un estudio de supervivencia. [34]

Existen diferentes tipos de censuras: *Right censoring*, *Left censoring*, *Right truncation*, *Left truncation*, *Type I*, *Type II* y *Type III* son las más comunes.

Una observación no está truncada ni censurada cuando el paciente entra en el estudio sano, y ocurre el evento de interés antes de que este se acabe:

$$I * \dots \text{---} +t$$

No truncada, censurada:

$$I * \dots \text{---} \dots$$

Truncada, no censurada:

$$\ast \dots \text{---} +t$$

Truncada, censurada:

$$\ast \dots \text{---} \dots$$

- Censura derecha: Sujeto abandona el estudio o el estudio acaba antes de que el evento haya ocurrido. Si el paciente abandona el evento en un tiempo t_i , el evento ocurre en (t_i, ∞) . Este tipo de censura es la más común. Dentro de esta modalidad de censura podemos diferenciar:
 - Tipo I: Seleccionamos n individuos y los observamos durante un tiempo t especificado previamente. Cualquier caso de no ocurrencia del evento es *right-censored* con tiempo de censura t .
 - Tipo II: Seleccionamos n individuos y los observamos hasta un tiempo t en el que hayan tenido lugar m eventos (donde m se especifica previamente y $m \leq n$). El resto de individuos $(n - m)$ están *right-censored*, siendo el tiempo de censura de estos el tiempo t_m en el que haya tenido lugar el último evento.

- Tipo III: Los individuos entran en el estudio en tiempos distintos, tendrán duraciones distintas en el estudio, y se pueden retirar antes del final de este. Es el caso más común en los estudios clínicos, y asimismo el caso de nuestros conjuntos de datos.

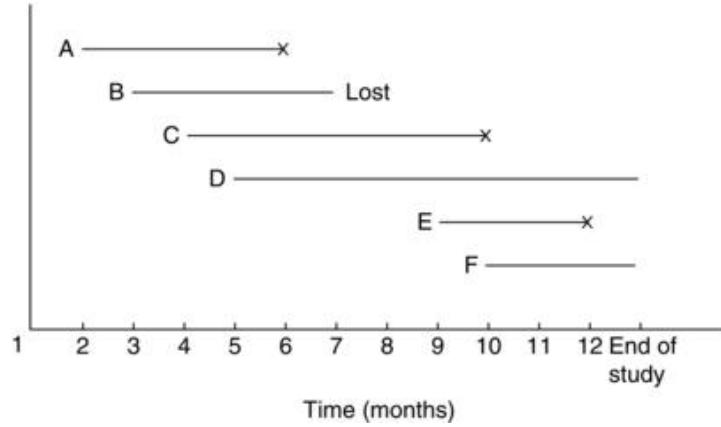


Figura 4: Censura derecha de tipo III. [35]

En la Figura 4 el eje- x representa el tiempo del estudio, desde el inicio al fin, en este caso representado en meses. Cada letra representa un paciente distinto, iniciando el estudio en tiempos distintos. La x representa el evento de interés. Podemos ver que en los pacientes A, C y D hemos observado el evento, y estos no están censurados. Los pacientes B, D y F están censurados de forma derecha; B por abandonar el estudio, y D y F porque el estudio ha llegado a su fin sin observarse el evento.

- Censura izquierda: El evento ya ha ocurrido antes del registro del paciente. Es un caso muy poco común.
- Truncamiento derecho: Toda la población del estudio ha experimentado el evento. En nuestro caso, si el evento fuera ser diagnosticado con cancer, tendríamos un truncamiento derecho, como en su lugar es recaer después de la biopsia o morir, no tenemos truncamiento derecho, pues todas las pacientes tendrían que haber recaido o fallecido respectivamente.
- Truncamiento izquierdo: Los sujetos del estudio han estado en riesgo de experimentar el evento antes de entrar en el estudio.

En el análisis de supervivencia se asume un supuesto básico: los mecanismos del evento y censura son estadísticamente independientes, es decir, los casos no censurados representan bien a los censurados.

2.1.2. Análisis de supervivencia discreto

Tanto el análisis univariante discreto como continuo asumen la hipótesis nula de que la probabilidad de un evento es igual entre la población comparada. También hacen las suposiciones de que la censura no se relaciona con el pronóstico, que la probabilidad de supervivencia no es significativamente diferente para individuos incluidos en diferentes tiempos del estudio, y que los eventos ocurren en tiempos específicos. [33]

El análisis discreto se refiere a los casos donde los datos solo pueden tomar valores discretos.

Sea T una variable discreta con una función de probabilidad de masa $P(T = t_j)$ para un conjunto de valores finito $\{t_1 < t_2 < \dots\}$ con probabilidades

$$f(t_j) = f_j = P\{T = t_j\} \quad (1)$$

Definimos la función de supervivencia en un tiempo t_j para un tiempo de supervivencia T como

$$S(t_j) = S_j = P\{T \geq t_j\} = \sum_{k=j}^{\infty} f_k \quad (2)$$

El riesgo en un tiempo t_j , como la probabilidad condicionada de que ocurra el evento en ese tiempo teniendo en cuenta que no ha ocurrido hasta este se define como

$$\lambda(t_j) = \lambda_j = P\{T = t_j | T \geq t_j\} = \frac{f_j}{S_j} \quad (3)$$

La función de supervivencia 2 en el tiempo t_j también puede verse en términos del riesgo de todos los tiempos anteriores t_1, \dots, t_{j-1}

$$S_j = (1 - \lambda_1)(1 - \lambda_2)\dots(1 - \lambda_{j-1}) \quad (4)$$

2.1.3. Análisis de supervivencia continuo

Es el caso de nuestros datos (aunque podrían transformarse a tiempos discretos).

Sea T el tiempo de vida futuro de un individuo de edad 0, T es una variable continua de valor $\mathbb{R}^+ = [0, \infty)$. En estudios de humanos, el tiempo de vida suele estar limitado a 120, por lo que usaremos de forma estandar $T \in [0, 120]$.

- Función de supervivencia de T : la probabilidad de sobrevivir en el tiempo t se define como

$$S(t) = P(T > t) \quad (5)$$

Donde el tiempo de muerte será mayor que cualquier otro tiempo definido para esa paciente, y donde se asume que $S(0) = 1$, siendo la probabilidad de supervivencia máxima al comienzo del estudio.

- Función de distribución de tiempo de vida de T (y la complementaria a la función de supervivencia 5): la probabilidad de muerte en el tiempo t se define como

$$F(t) = P(T \leq t) = 1 - S(t) \quad (6)$$

- Función de densidad del evento, representa el índice de muertes por unidad de tiempo y se obtiene derivando la función 6

$$f(t) = F'(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \quad (7)$$

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t) \quad (8)$$

- Función de riesgo de T (*Hazard Function*) o riesgo de que ocurra el evento en el tiempo t se define como

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} \quad (9)$$

Esta función nunca tiene valores negativos y su valor se comprende en $\lambda(t) \in [0, \infty]$, es decir, para denominarse función de riesgo (*hazard* o $h(x)$) tiene que cumplir que

$$\forall x \geq 0 \rightarrow h(x) \geq 0 \text{ y } \int_0^\infty h(x)dx = \infty$$

- Función de riesgo de acumulado de T , el índice de eventos en un tiempo t condicionado por la supervivencia hasta ese tiempo siendo $T \geq t$ se define como

$$H(t) = \int_0^t h(x)dx \quad (10)$$

2.1.4. Estimadores estadísticos

Definición y significado de las medidas estadísticas que hemos usado y analizado para diferentes causas a lo largo del desarrollo del proyecto. Encontramos estimadores de similitud de curvas 12 , independencia 13 , correlación 15 y relación 16 entre de variables, significancia de variables 14 y probabilidad del evento debido a una variable 2.1.4. Los veremos más adelante la sección 5.1.2, donde los hemos usado para seleccionar el conjunto de variables final.

- **p-value:** Representa la probabilidad de que la hipótesis nula sea cierta. Se establece un valor umbral de significancia (normalmente 0,05) el cual determina, según el si el $p - value$ es mayor o menor, se acepta o descarta la hipótesis nula H_0 , respectivamente.

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (11)$$

Siendo n el tamaño de la muestra, p la proporción del conjunto, y p_0 la proporción del conjunto asumida.

- **Logrank test:** Es el método más popular para comparar grupos de supervivencia (y el que usa el estimador *Kaplan – Meier* 2.2.1). Es un test de hipótesis no paramétrico para comparar distribuciones de supervivencia de dos conjuntos. Se basa en la hipótesis nula de que no hay diferencia entre las poblaciones para la ocurrencia de un evento ($H_0 : h_1(t) = h_2(t)$ y $H_1 : h_1(t) \neq h_2(t)$).

$$Z_i = \frac{\sum_{j=1}^J (O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^J V_{i,j}}} \quad (12)$$

Siendo O el número de casos observados y E el número esperado de casos.

- **Chi-squared Test of Independence (χ^2):** Dos variables se consideran independientes si la distribución de probabilidad de una no se ve afectada por la presencia de la otra. Este test determina si hay una diferencia significativa entre las frecuencias esperadas y las frecuencias observadas en las categorías. Cuando la hipótesis nula es cierta, la muestra cumple la distribución de *chi – squared* (distribución de la suma de cuadrados de k variables aleatorias independientes y estandar) ($H_0 : O_i = E_i$, $H_1 : O_i \neq E_i$).

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (13)$$

Siendo O_i el número de casos observados en la categoría i y E_i el número esperado de casos en la categoría i .

- **Wald test o Wald Chi-Squared Test (W):** Sirve para determinar si las variables que definen a una paciente son significativas, es decir, añaden información al modelo. Las variables cuyo resultado de este test sea 0 pueden ser eliminadas sin que afecte de una manera significativa. ($H_0 : R\theta = r$ y $H_1 : R\theta \neq r$)

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})} \quad (14)$$

Siendo $\hat{\theta}$ el error estandar del mayor estimador *likelihood* y θ un valor supuesto.

- **Pearson Correlation Coefficient (ρ):** mide la correlación lineal entre dos variables.

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \quad (15)$$

Para las variables X, Y , el numerador representa la covarianza y el denominador las desviaciones estándar de X e Y .

$\rho \in [-1, +1]$, donde +1 es correlación positiva total, -1 correlación inversa total, y 0 ninguna correlación.

- **Spearman's Rank Correlation Coefficient o Test-rho (r_s):** Es una medida no paramétrica de la correlación de *rank* (mide la dependencia estadística entre los rangos de dos variables) e indica como de bien se puede describir la relación de dos variables usando una función monótona (función que preserva o reversa el orden de conjuntos ordenados). Sería el equivalente a la correlación de Pearson entre rangos.

$$r_s = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \quad (16)$$

Donde el numerador representa la covarianza entre el rango de las variables y el denominador las desviaciones estándar del rango de las variables.

$\rho \in [-1, +1]$ y se interpreta de la misma manera que la correlación de Pearson.

- **Likehood-ratio test (LR):** Evalúa cómo de bueno es el ajuste de dos métodos estadísticos enfrentados, basándose en sus índices *likehood*. Estos expresan la probabilidad de que un conjunto de observaciones sea debido a diferentes valores de parámetros estadísticos. En nuestro caso, nos indica cómo de probable es para un paciente sufrir el evento. A mayor LR , mayor probabilidad. un valor comprendido entre 0 y 1 disminuye la evidencia del evento, mientras que por encima de 1 la incrementa más cuanto mayor es el valor.

2.2. Métodos estadísticos para estimar supervivencia

Los modelos estadísticos de análisis de supervivencia se dividen en paramétricos, semi-paramétricos y no-paramétricos. Todos estos sirven para estimar la función de supervivencia 5

- Los métodos **paramétricos** son más eficientes (porque se estiman menos parámetros), y suelen hacer interpretaciones más relevantes del modelo. Sin embargo, tienen que ajustarse a una distribución de datos para cumplir las condiciones de validez. Algunos de ellos son: *Exponential, Weibull, Log-normal,...*
- Los métodos **no-paramétricos** son más robustos, y no es requisito que datos que los datos cumplan una distribución definida, ya que son los datos observados la que la determinan. Estos métodos no pueden lidiar por sí mismos con regresión multivariante, lo que convierte a los métodos semi-paramétricos en la mejor aproximación.

Algunos métodos no paramétricos son: Estimador de *Kaplain-Meier*, Estimador de *Nelson – Aaelen*, Estimador de *Life – Table*,...

- Los métodos **semi-paramétricos** se conocen por este nombre debido a que estiman la función de riesgo λ de forma no-paramétrica, pero la forma funcional de las variables es paramétrica. El más conocido y usado es *Cox PH model* 2.2.2.

2.2.1. Métodos estadísticos no-paramétricos

Como hemos comentado anteriormente, son modelos estadísticos de análisis de supervivencia para los que el tiempo no sigue necesariamente una distribución normal.

Suponemos n individuos independientes distribuidos de forma idéntica en el conjunto, y siendo $t_1 < t_2 < \dots < t_k$ los tiempos de eventos ordenados para $k \leq n$: $\sum_{j=1}^k d_j = n$ siendo d_j el número de muertes en el tiempo t_j .

El estimador de *Kaplan – Meier* se obtiene a partir de la ecuación de probabilidad de supervivencia discreta y de la máxima estimación *likelihood* y se define como

$$\hat{S}(t) = \prod_{j:t_j \leq t} (1 - \hat{h}_j) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \quad (17)$$

El estimador de *Nelson – Aalen* o función de riesgo acumulada en análisis de supervivencia discreto se define como

$$\hat{H}(t) = \sum_{j:t_j \leq t} \hat{h}_j = \sum_{j:t_j \leq t} \frac{d_j}{n_j} \quad (18)$$

2.2.2. Cox Proportional Hazards Model

Es el modelo de regresión semi-paramétrico más usado en el análisis de supervivencia clínico, su aplicación principal es la de analizar la asociación entre las variables predictoras (covariables) y el tiempo de supervivencia. A diferencia de la mayoría de métodos no-paramétricos, funciona tanto para variables cualitativas como cuantitativas. También tiene la capacidad de realizar un análisis multivariante (evaluar el efecto de más de una variable predictora al mismo tiempo).

En este análisis de variables en el tiempo, nos indica específicamente como afecta cada nivel de las variables categóricas en la ocurrencia del evento.

Se expresa por la función de riesgo variante en el tiempo $h(t)$ determinada por las covariables (x_1, x_2, \dots, x_p) asumidas estas independientes del tiempo como

$$h(t, x(t)) = h_0 \cdot \exp(\beta \cdot x(t)) = h_0 \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) \quad (19)$$

Donde t es el tiempo de supervivencia y β_i los coeficientes que indican el impacto de las covariables. h_0 es el valor *baseline – hazard*, el valor de $h(t)$ si todo x_i tuviera valor nulo, donde $\exp(0) = 1$.

Los valores $\exp(\beta_i)$ son los índices de riesgo o *hazard – ratios (HR)*, donde, si algún coeficiente de impacto $\beta_i > 0$, lo que implicaría $\exp(\beta_i) > 1$, el valor de la covariable β_i incrementa, provocando un incremento del riesgo de evento y por lo tanto disminuyendo la supervivencia. Por ende, $HR = 1$ significaría que el factor no causa efecto, $HR > 1$ aumenta el riesgo bajando la supervivencia, y $HR < 1$ disminuye el riesgo aumentando la supervivencia.

Aunque, como hayamos comentado, este método implementado por `coxph()` [49] asume que las covariables son independientes del tiempo, y en nuestros conjuntos de datos contemos con variables que claramente no lo son (como la edad de las pacientes), Cox PH puede generalizarlas estratificándolas y asumiendo el riesgo proporcional en cada nivel de la estratificación.

Otra de las suposiciones que hace este modelo, como su nombre indica, es la de la **proporcionalidad de riesgos**, lo que significa que el índice de riesgo de dos individuos es constante en el tiempo, es decir, son proporcionales. Esta relación constante para dos individuos x_1 y x_2 se expresa 20 y debe ser constante e independiente del tiempo.

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{\exp(\beta \cdot x_1)}{\exp(\beta \cdot x_2)} = \exp(\beta \cdot (x_1 - x_2)) \quad (20)$$

Evaluaremos la suposición usando la función `cox.zph()`, también del paquete [49], que evalúa la proporcionalidad en el tiempo de la forma 21. Comienza basándose en la hipótesis nula de que se cumple la proporcionalidad, la cual se acepta a no ser que el $p-value$ sea menor que el umbral definido (0.05).

$$HR = \exp\left(\sum_{j=1}^p \beta_j (X_j^* - X_j)\right) \quad (21)$$

Siendo X^* y X las especificaciones de las variables, y donde el riesgo de un individuo debe ser proporcional al del otro.

2.2.3. Índice de concordancia

El Índice de concordancia o $C - index$ es la métrica estándar para evaluar a los modelos de supervivencia. Se define como la probabilidad de concordancia de dos observaciones elegidas de forma aleatoria, donde concordancia significa que la observación con el tiempo de supervivencia menor tendrá la puntuación de riesgo más alta, y viceversa. Son comparables todos los pares de observaciones donde no haya ninguna que se esté observando después de su tiempo de censura.

Nos indica como de bueno es el ajuste para salidas binarias en una regresión logística. En nuestro caso, indica la probabilidad de que un paciente elegido aleatoriamente que haya experimentado el evento tenga mayor puntuación de riesgo que uno que no lo haya experimentado. Su valor es el área bajo la curva *Receiver Operating Characteristic (ROC)*.

Un mayor índice de concordancia indica una mayor precisión de la predicción de la supervivencia. Su rango de valores se encuentra en el intervalo $[0, 5, 1]$, siendo 0.5 un modelo muy malo, 0.8 un modelo fuerte, y 1 un modelo que predice perfectamente el grupo de pacientes que experimentarán el evento y aquellos que no.

Una de las grandes desventajas de este índice es que únicamente mide el riesgo relativo en lugar de la diferencia real entre el resultado del predictor y el real. Además, dificulta las comparaciones con métodos de aprendizaje automático debido a que estos emplean una salida de predictor distinta.

Lo usaremos en la sección 5.3.2 para evaluar los métodos de supervivencia.

3. Contexto Aprendizaje Computacional

3.1. Aprendizaje no supervisado

El aprendizaje no supervisado, *unsupervised learning* o *self – organization* es una de las ramas principales del aprendizaje computacional. Su característica diferenciativa es que los datos a aprender no están clasificados o etiquetados. De esta manera, en lugar de aprender en base a corrección de un error generado por la comparación con el valor real (Figura 5), estos algoritmos identifican patrones desconocidos y cosas en común en los datos, definiendo así grupos diferenciados 6.

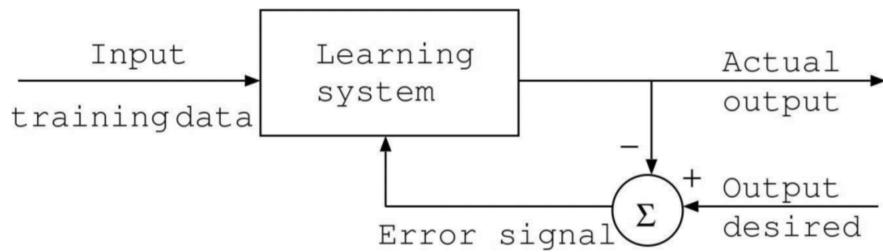


Figura 5: Esquema de aprendizaje de un algoritmo supervisado. [36]

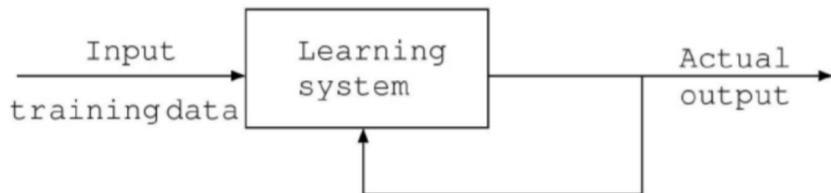


Figura 6: Esquema de aprendizaje de un algoritmo no supervisado. [36]

Uno de las técnicas del aprendizaje no supervisado es el *Clustering*, dentro de la que encontramos dos de sus métodos más conocidos, *K – means* 3.1.2 y *Hierarchical – Clustering* 3.1.3.

En este proyecto hemos hecho uso estos dos últimos métodos mencionados para crear estratificaciones de las variables numéricas de los pacientes (edad, ganglios afectados, ganglios extraídos, nodos afectados, tamaño del tumor,...), estratificaciones de las que evaluaremos su mayor significancia respecto a las agrupaciones médicas estandar (5.1.1).

3.1.1. Métricas clusters

Como hemos comentado, en el aprendizaje supervisado es muy sencillo saber como de preciso es el modelo ya que solo hay que comparar la salida de este con la real para saber que error ha cometido, sin embargo, en el aprendizaje no supervisado no tenemos ningún valor con el que comparar la salida generada. En su lugar, para evaluar los clusters generados nos basamos en medidas de validación interna. El objetivo ideal de un algoritmo de clustering es identificar conjuntos de datos que sean compactos entre ellos y con poca varianza entre los miembros de un mismo conjunto (cohesión), a la vez que estos están lo más separados posible de los miembros de otros clusters (separación).

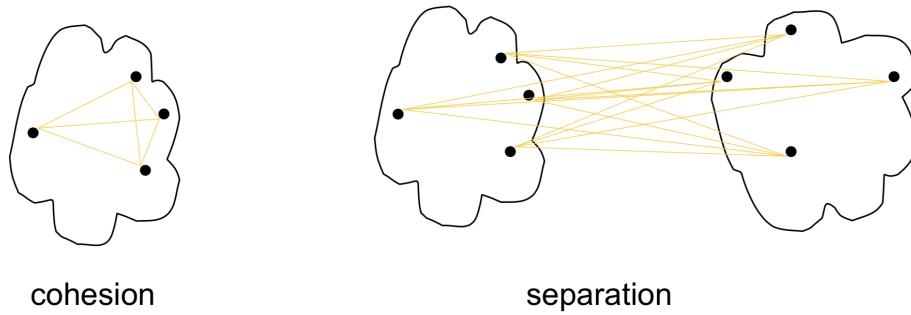


Figura 7: Cohesión y separación clusters de forma gráfica. [36]

Otra de las características de estos algoritmos es que requieren de que se les indique de forma previa el número de grupos a realizar, definido como k . Dado que no sabemos el número de agrupaciones a realizar de forma previa, usaremos las siguientes medidas de validación interna para saber que k es más adecuada.

- **Elbow method:** Analiza el porcentaje de variación según el número de clusters. Se basa en que el número más adecuado es aquel que al añadir otro los cambios en el conjunto dejan de ser significativos. Los primeros clusters añaden muchos cambios y varianza, mientras cae la ganancia marginal hasta el “codo” (*elbow*), punto de estabilización en el que más clusters pueden añadir complejidad innecesaria.

Se basa en la suma total de cuadrados TSS 22, que representa la suma del cuadrado de las diferencias de cada observación con la media.

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (22)$$

- **Silhouette Index:** Mide la similitud de cada objeto con su propio clusters (cohesión).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \in [-1, 1] \quad (23)$$

Donde $a(i)$ es la distancia media de i a los puntos de su cluster y $b(i)$ la mínima $a(i)$. Un valor $+1$ indica que el *match* de esa paciente en ese cluster es perfecto, 0 que se encuentra en al frontera natural de dos clusters, y -1 que debería pertenecer a otro cluster.

- **Dunn index:** Es el cociente la la distancia intercluster mínima y el tamaño máximo de cluster. Un mayor índice de *Dunn* (DI) indica una mejor clusterización, ya que mayores distancias inter-cluster conlleva mejor separación y tamaños de cluster más pequeños indican cluster más compactos.

$$DI = \frac{\min_{1 \leq i \leq j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k} \quad (24)$$

- **Connectivity Index:** Indica el grado de conectividad entre clusters determinado por el algoritmo $k - nearest - neighbours$. Esta conectividad representa como de extensas están las pacientes situadas en el mismo cluster como sus vecinos más cercanos en el espacio de datos. Su valor se comprende en $[0, \infty)$ y debe ser minimizado.

3.1.2. K-means clustering

Uno de los algoritmos de aprendizaje no supervisado más conocidos y simples. Este algoritmo identifica k centroides que representan los centros de los clusters, y sitúa cada punto (paciente) en el cluster al que pertenezca su centroide más cercano. Es un algoritmo iterativo. Cada vez que tiene todos los puntos situados en su cluster correspondiente, optimiza la posición de los centroides como punto medio de todos los puntos que ahora pertenecen al cluster, y vuelve a repetir el paso anterior, iterativamente. El algoritmo para

cualquier momento al recalcular los centroides estos no cambian respecto a su valor en la iteración anterior.

Para un dataset D , con k clusters y d dimensiones, siendo C_i el cluster i th,

1. Partición inicial de $D = \{C_1, C_2, \dots, C_k\}$
2. Calcular la distancia entre el punto i y el cluster j (d_{ij}).
3. $n_i = \arg \min_{1 \leq j \leq k} d_{ij}$.
4. Asignar punto i a cluster n_i .
5. Recalcular los centroides de los clusters.
6. Repetir desde los puntos 2.–5. hasta que haya dos iteraciones seguidas con el mismo valor de centroides.

Algunas desventajas de este algoritmo: depende de su inicialización, es sensible a los *outliers*, solo es eficiente con distribuciones esféricas simétricas.

3.1.3. Hierarchical clustering

Este algoritmo, en lugar de dividir los datos en particiones individuales como es el caso de $k - means$, divide los datos en una secuencia de particiones anidadas.

Podemos diferenciar entre clustering jerárquico aglomerativo y divisivo. En el primer caso se inicializa formando un cluster de cada objeto, y repite el procedimiento uniendo los pares de clusters más cercanos acorde a un criterio de similitud hasta que llega a un solo cluster. El segundo tipo lo hace de forma inversa, empieza con todos los objetos en un solo cluster y en cada iteración parte los clusters más lejanos.

Las desventajas de este algoritmo: Los clusters mal agrupados en un comienzo no tienen la posibilidad de ser realocados, distintas medidas de similitud darán distintos resultados.

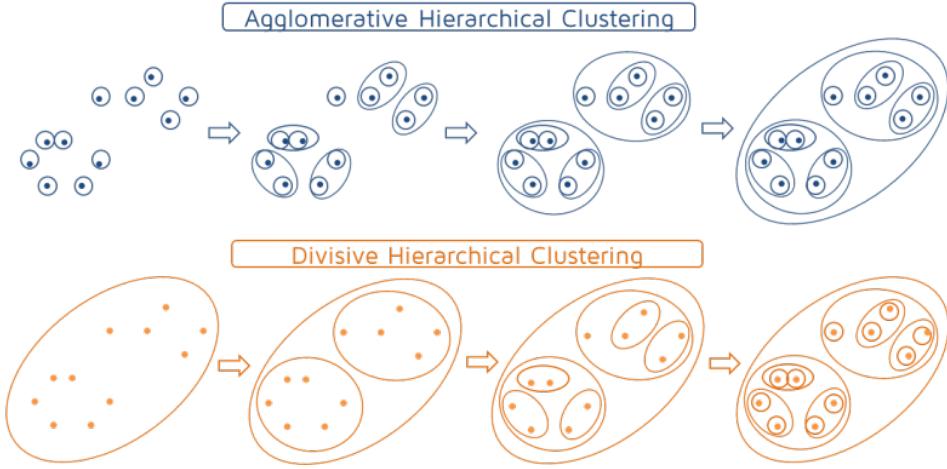


Figura 8: Funcionamiento de *Hierarchical Clustering* aglomerativo y divisivo. [37]

Las medidas en las que se pueden basar estos algoritmos para unir o separar los clusters, según la modalidad son:

- *Single – Link*: $d_{min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$, Distancia de los puntos más cercanos.
- *Complete – Link*: $d_{min}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$, Distancia de los puntos más lejos.
- *Average – Link*: $d_{min}(C_i, C_j) = \text{avg}_{p \in C_i, q \in C_j} d(p, q)$, Distancia media entre pares de elementos.

3.2. Modelos de Clasificación Probabilística

Estos modelos pertenecen al subtipo de aprendizaje computacional Aprendizaje supervisado. Como hemos comentado en la sección 3.1 y visto en la Figura 5, a diferencia de los no supervisados, estos tienen la variable a predecir etiquetada (en nuestro caso el evento), a partir de la cual pueden corregir su error de salida para ajustarse y aprender a partir de este. Este algoritmo mapea los *inputs* (variables clínicas) con el *output* (evento), de esta forma y a partir del error generado se entrena hasta ser capaz de generalizar y predecir una respuesta acertada cuando reciba datos nuevos (una paciente con la que no se le ha entrenado).

Dentro de los modelos de aprendizaje supervisado, podemos distinguir entre dos grandes categorías: Regresión (predicen valor continuo) y Clasificación (predicen categoría). Dado que nuestra variable de salida es categórica (**evento** con los niveles 1 o 0), pero queremos estimar la supervivencia o, más específicamente, la probabilidad para cada tiempo t de que el evento sea 0, hemos elegido, además de modelos de supervivencia, Clasificadores probabilísticos. Estos clasificadores, dada una observación, pueden predecir una distribución de probabilidad sobre un conjunto de niveles a demás de a cuál de ellos es más probable que pertenezca.

3.2.1. Métricas para modelos de clasificación

La matriz de confusión es una tabla bidimensional que representa los cuatro casos posibles de una predicción y en la que se basan la mayoría de las medidas más conocidas:

		Predicted class	
		<i>P</i>	<i>N</i>
Actual Class	<i>P</i>	True Positives (TP)	False Negatives (FN)
	<i>N</i>	False Positives (FP)	True Negatives (TN)

Figura 9: Matriz de confusión.

Minimizar Falsos Negativos y positivos y maximizar Verdaderos positivos y negativos.

- Accuracy (ACC) mide el número de predicciones correctas entre todas las realizadas. En nuestro caso, por estar la variable a predecir algo desbalanceada, hemos observado tambien valor medio por clase además de la global. Debe ser maximizado.

$$\frac{TP + TN}{TP + FP + FN + TN}$$

- Precisión: Nos indica que proporción de los pacientes que ha predicho sufrir el evento realmente han sufrido el evento. Debe ser maximizado.

$$\frac{TP}{TP + FP}$$

- Sensibilidad o recall (TPR): Nos indica que proporción de los pacientes que han sufrido el evento han sido predichos como tal. Debe ser maximizado.

$$\frac{TP}{TP + FN}$$

- Especificidad: Que proporción de pacientes que no han sufrido el evento han sido predichos como no sufridores del evento. Debe ser maximizado.

$$\frac{TN}{TN + FP} \text{ o } 1 - FPR$$

- Area Under the Roc Curve (AUC): ROC es una curva de probabilidad obtenida de graficar el índice de verdaderos positivos (TPR) frente al índice de falsos negativos (FPR). El área de esta nos indica cómo de bien es capaz nuestro modelo de distinguir entre clases. Su valor se encuentra $\in [0,5, 1]$ siendo 0.5 un clasificador que se comporta como la aleatoriedad, y 1 un clasificador perfecto.
- Missclassification rate ($MMCE$): Indica el porcentaje de instancias mal clasificadas. Es el inverso de ACC . Debe ser minimizado.

$$\frac{FP + FN}{TP + FP + FN + TN}$$

- Mean Squared Error (MSE): Representa la media cuadrada de la diferencia entre los valores estimados y los valores reales a través de la media del cuadrado de los errores. Siempre es estrictamente positiva y debe ser minimizada.

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

3.2.2. Deep Learning (DL)

Se considera aprendizaje profundo a aquellas redes neuronales artificiales (*ANN*) con más de una capa interna. Una red neuronal artificial es un conjunto de nodos (neuronas artificiales) conectados para transmitirse señales que simulan el comportamiento de las neuronas como si de un cerebro biológico se tratase. La forma de corregir el error y aprender es ajustar los pesos sinápticos (conexiones) entre nodos.

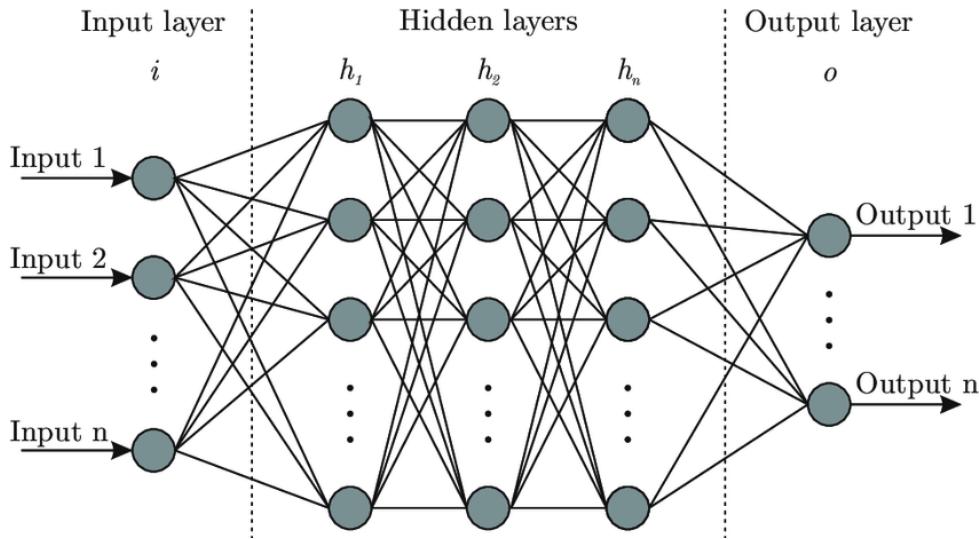


Figura 10: Arquitectura de una red neuronal artificial. [39]

Cada neurona en la capa de entrada representa una variable del conjunto de datos. Las capas internas (*hidden layers*) representan las relaciones no lineales entre las variables de entrada y la salida. Cada neurona adquiere la suma de los pesos de los *outputs* anteriores, la transfiere a través de una función de activación, y le da el valor a su siguiente neurona. Algunas de las funciones de activación existentes podemos verlas en la Figura 12.

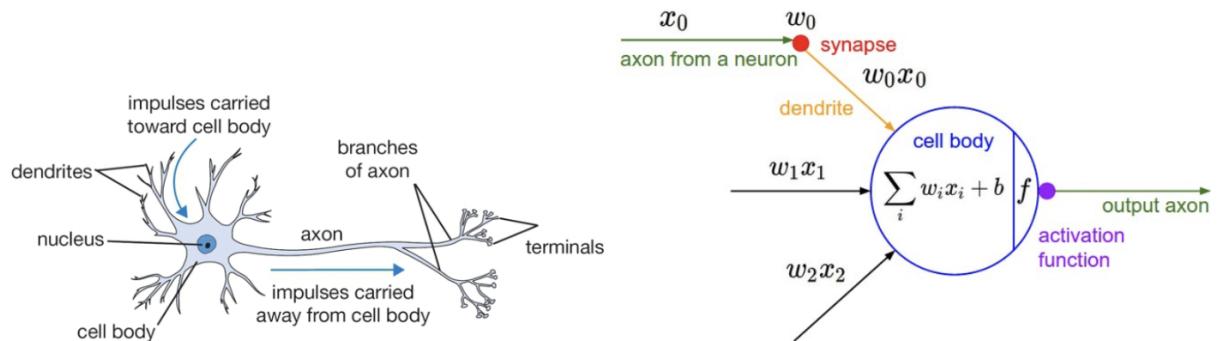


Figura 11: Neurona biológica (izquierda) y su modelo matemático (derecha). [40]

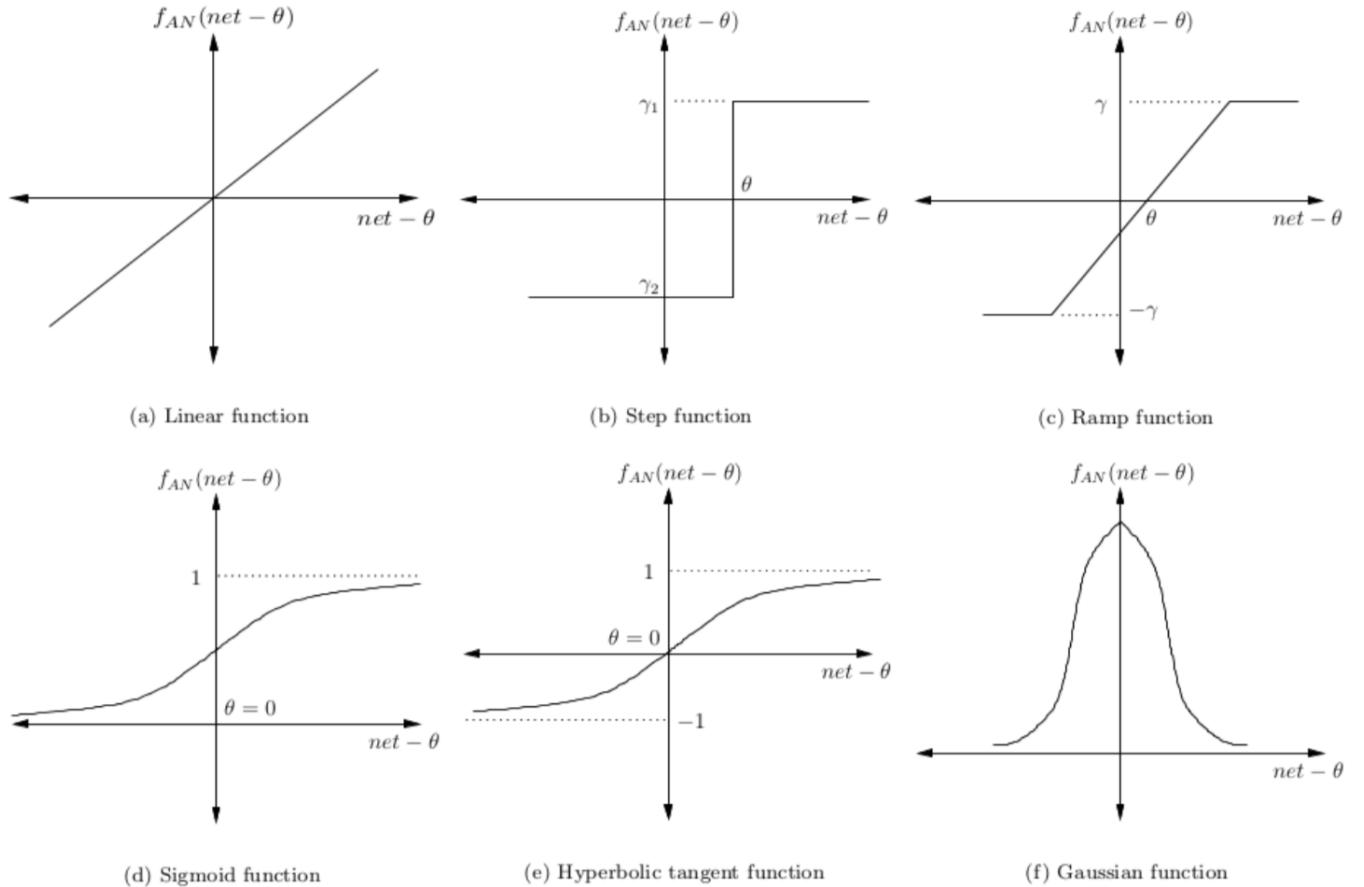


Figura 12: Tipos de funciones de activación. [36]

El objetivo de una red neuronal es minimizar su función de pérdida. El procedimiento para conseguirlo es el siguiente:

1. Inicializar los pesos de forma aleatoria.
2. Propagar los *inputs* a través de la red para calcular los *outputs*.
3. Usar un optimizador para minimizar la función de perdida respecto a cada peso w_{ij} y después de cada iteración actualizar ese peso al nuevo valor del peso a $w_{ij} + \Delta w_{ij}$, donde Δw_{ij} está determinado por el optimizador.
4. Repetir los pasos 2.-3. hasta que se cumpla la condición de parada.

3.2.3. Random Forest (RF)

Un *Random Forest* lleva a cabo el cómputo de n Árboles de decisión (*Decision Trees*) dando como salida la media (en caso de ser una regresión) o la moda (en caso de ser una clasificación) de la salida de estos, lo que evita el *overfitting* que suele generarse en los árboles de decisión.

Un Árbol de Decisión (*DT*) es un grafo de decisiones. El nodo raíz representa a toda la población de pacientes. Este nodo se va dividiendo en otros dos iterativamente hasta que cada uno de los dos nodos resultantes es homogéneo. Cada nodo interno o nodo de decisión representa un *test* para el atributo, que definirá por cuál de sus nodos hijos continua. Cada nodo hoja o nodo terminal representa un nivel de la variable a predecir. Podemos ver esta nomenclatura representada en la Figura 13.

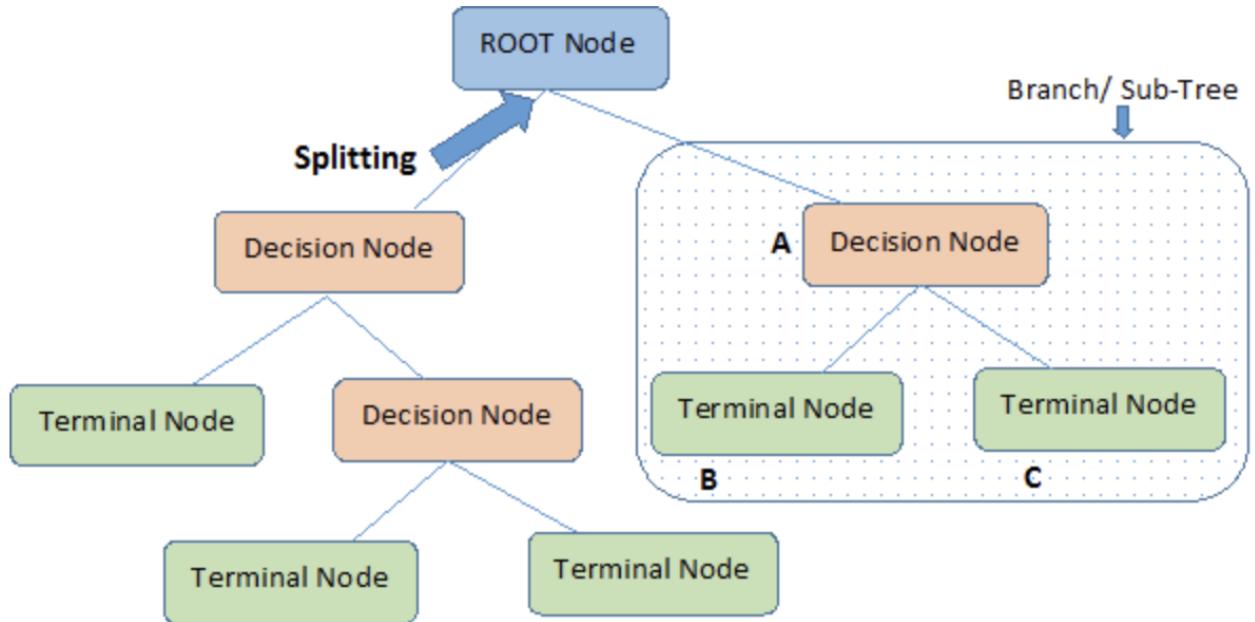


Figura 13: Arquitectura de un árbol de decisión. [41]

Para calcular la homogenidad de cada nodo con el objetivo de decidir si se vuelve a dividir o no, existen varios algoritmos: *ID3*, *Gini Index*, *Chi-Square* y *Reduction in Variance*.

3.2.4. Gradient Boosting Machine (GBM)

Boosting es una técnica para convertir modelos de aprendizaje débiles en modelos de aprendizaje robustos. *Gradient Boosting Machine* entrena gradualmente varios modelos de forma aditiva y secuencial. Comienza entrenando un árbol de decisión donde a cada observación se le asigna el mismo peso, este árbol se evalúa y se incrementan los pesos de aquellas observaciones difíciles de clasificar. Así, cada árbol es la suma de los arboles anteriores con predicción mejorada. Este proceso se repite n iteraciones.

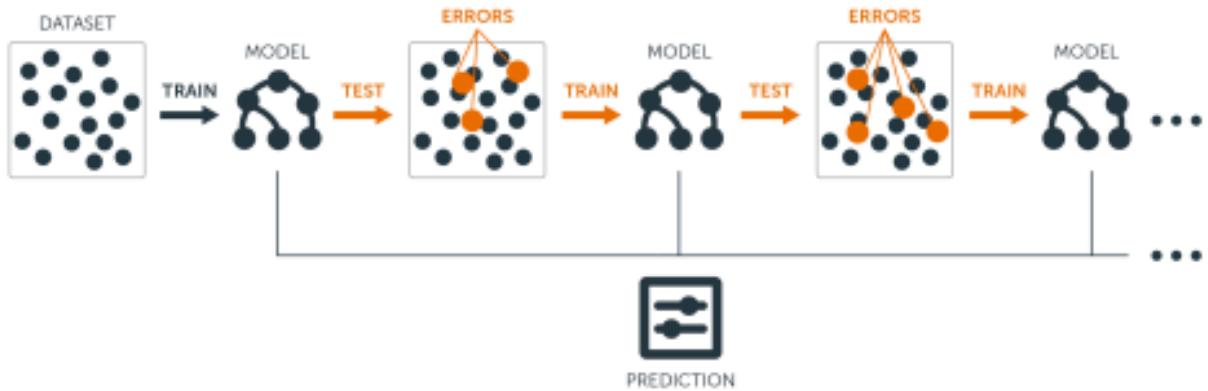


Figura 14: Esquema de iteraciones de una máquina Gradient Boosting. [42]

3.2.5. Extreme gradient boosting (XGBoost)

Este algoritmo sigue el principio de “*llover al extremo los límites de computación de una máquina para conseguir un modelo escalable, portable y preciso.*”[43]

Se basa en un modelo *Gradient Boosting Machine*, con la diferencia de que es un modelo más regularizado y controla mejor el *overfitting*, dando mejores resultados.

3.2.6. Generalized Linear Models (GLM)

Un modelo de regresión lineal general define la relación lineal entre la variable dependiente (respuesta) Y y las independientes (variables predictoras) X_i a través de coeficientes de regresión b_i :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

El modelo, usando la función de coste, trata de encontrar los coeficientes b_i que minimicen la función de coste, es decir, aquellos que hagan que el error entre el valor real y el predicho sea mínimo.

$$\text{Función de coste} = \frac{1}{n} \sum_{i=1}^n (pred_i - y_i)^2 \quad (25)$$

El modelo linear generalizado, como su nombre indica, se trata de una regresión lineal generalizada. La diferencia respecto a una regresión lineal común es que acepta variables cuya distribución de error no siga una distribución normal. Esto se consigue asociando el modelo lineal a la variable respuesta a través de una *link function* la cual la transforma.

3.3. K-Fold Cross Validation

K-Fold Cross Validation es la estrategia de *Resampling* más popular. Las técnicas de *resampling* consisten en, repetidamente, dividir el conjunto de datos completo en conjuntos de entrenamiento con los que se entrena el modelo y de validación con los que se validan las predicciones y se calcula la medida correspondiente de precisión. En cada iteración la división de los conjuntos es distinta. Después de las iteraciones, las medidas de precisión correspondiente a cada una de ella es agregada, normalmente con la media. De esta manera podemos asegurar que la medida del comportamiento del modelo sobre datos nunca vistos no es debida al azar de una división de datos concreta.

El modelo de *Cross – Validation*, en cada una de las K iteraciones divide el conjunto de datos en k *folds* con igual número de filas en cada una. Una vez divididos los datos, una de las cajas es usada como conjunto de validación y el resto como conjunto de entrenamiento. Después del entrenamiento se calcula la medida de precisión correspondiente. Realizadas las K iteraciones, se calcula la media de las K medidas calculadas. Podemos verlo ilustrado en la Figura 15.

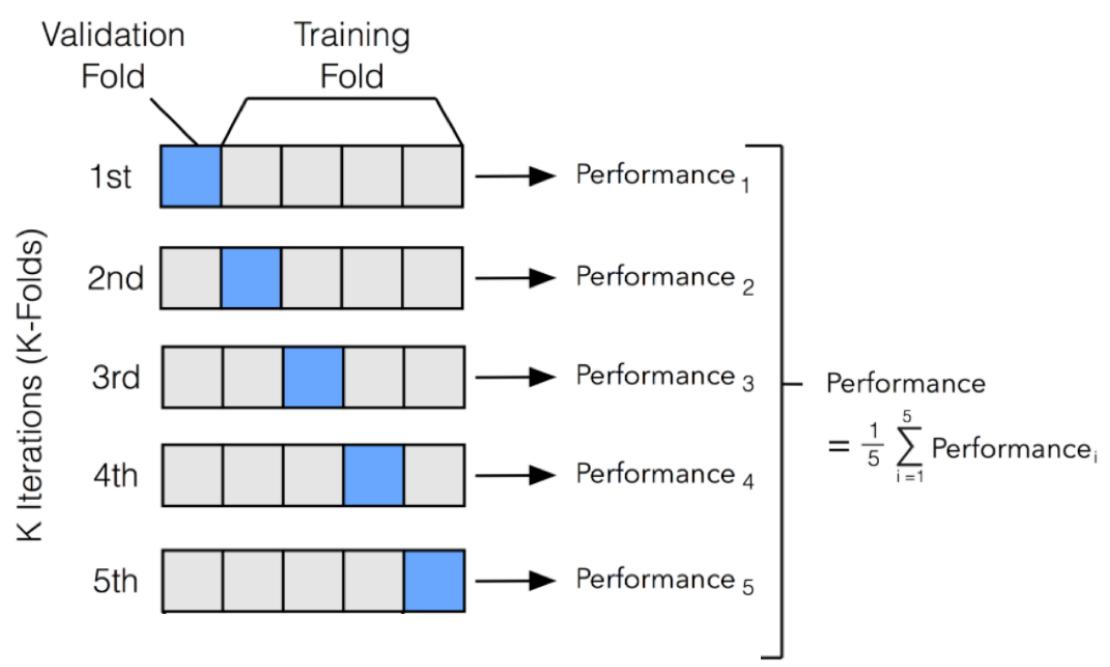


Figura 15: Cross validation 5 Folds. [45]

4. Conjuntos de datos

En el desarrollo de este proyecto hemos trabajado paralelamente con dos conjuntos de datos de supervivencia (4.1 y 4.2). Antes de empezar a trabajar con nuestros datos, los analizaremos y limpiaremos como explicaremos a continuación.

4.1. Hospital

Se trata de un conjunto de datos de pacientes con cáncer de mama y receptores hormonales positivos tratadas con hormonoterapia adyuvante, libres de enfermedad a 5 años del primer diagnóstico. Se han extraído del hospital Virgen de la Victoria (Málaga). De estos analizaremos su probabilidad de recurrencia.

4.1.1. Atributos

- **Variables numéricas:** Edad (en años), ki67 (en porcentaje), número de ganglios afectados, gnúmero de angilos extraídos, tamaño del tumor (en mm) y tiempo de seguimiento (en años). Ver distribución en Figuras 16 y 17.
- **Variables categóricas:** Estado menopausico (*Premenopausica, Postmenopausica*), grado (*I, II, III*), receptores hormonales (de estrógeno, progesterona y her2), subtipo de cáncer (*Luminal A, Luminal B*), riesgo (*Bajo, Medio, Alto*), hormonoterapia (*Inhibidores Aromatasa (IA), Tamoxifeno, IA+Tamoxifeno*), estado último control (*muerta con enfermedad, muerta sin enfermedad, viva con enfermedad, viva sin enfermedad*), receptor hormonal (creada a partir de los receptores hormonales: *Positiva en estrógeno y progesterona, Positiva en estrógeno*), y evento (en este caso significa recurrencia, con los valores 0, 1). Ver frecuencias en Figura 18.

4.1.2. Distribución y valores atípicos

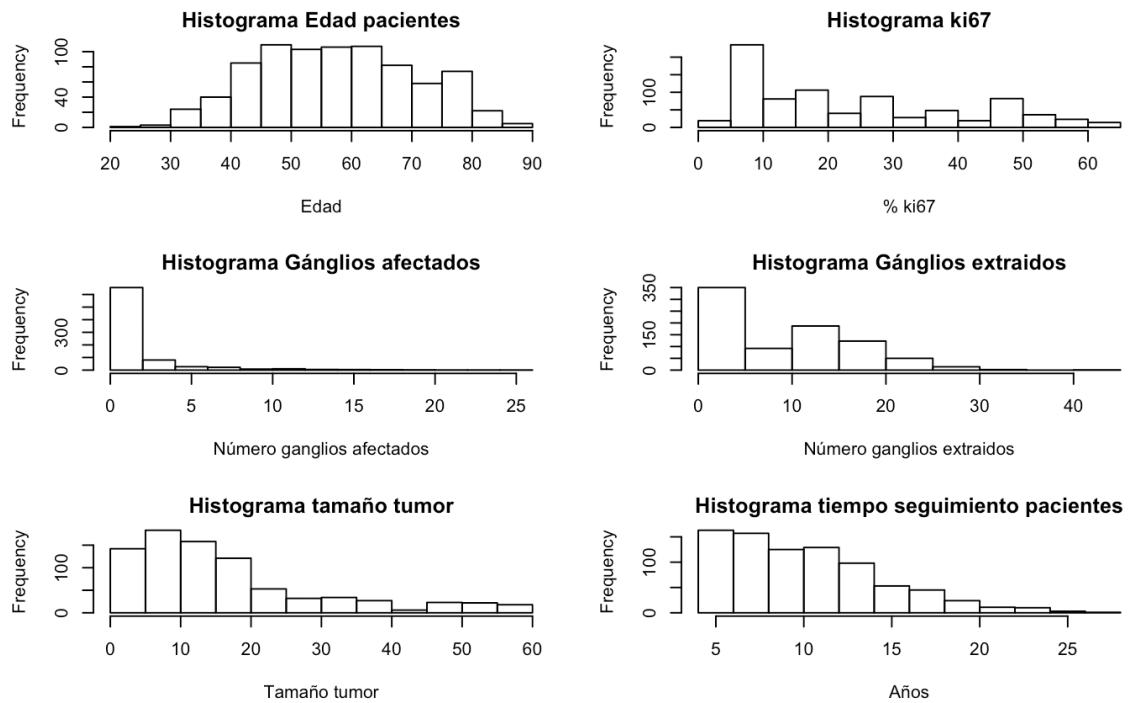


Figura 16: Histogramas distribución variables numéricas datos Hospital.

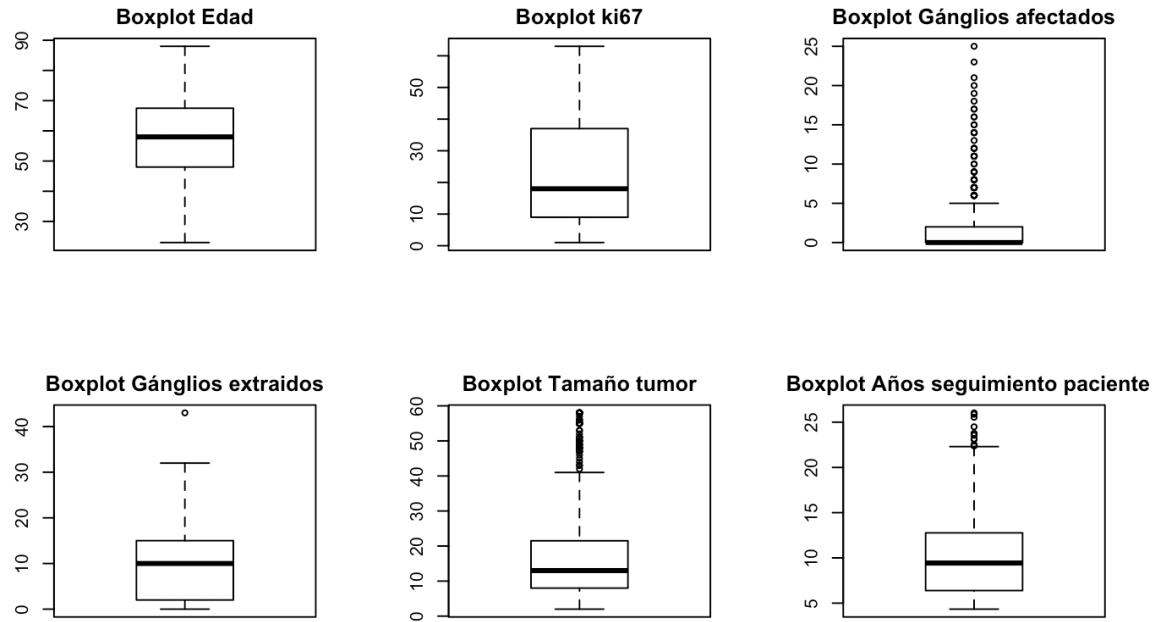


Figura 17: Diagramas de cajas variables numéricas datos Hospital.

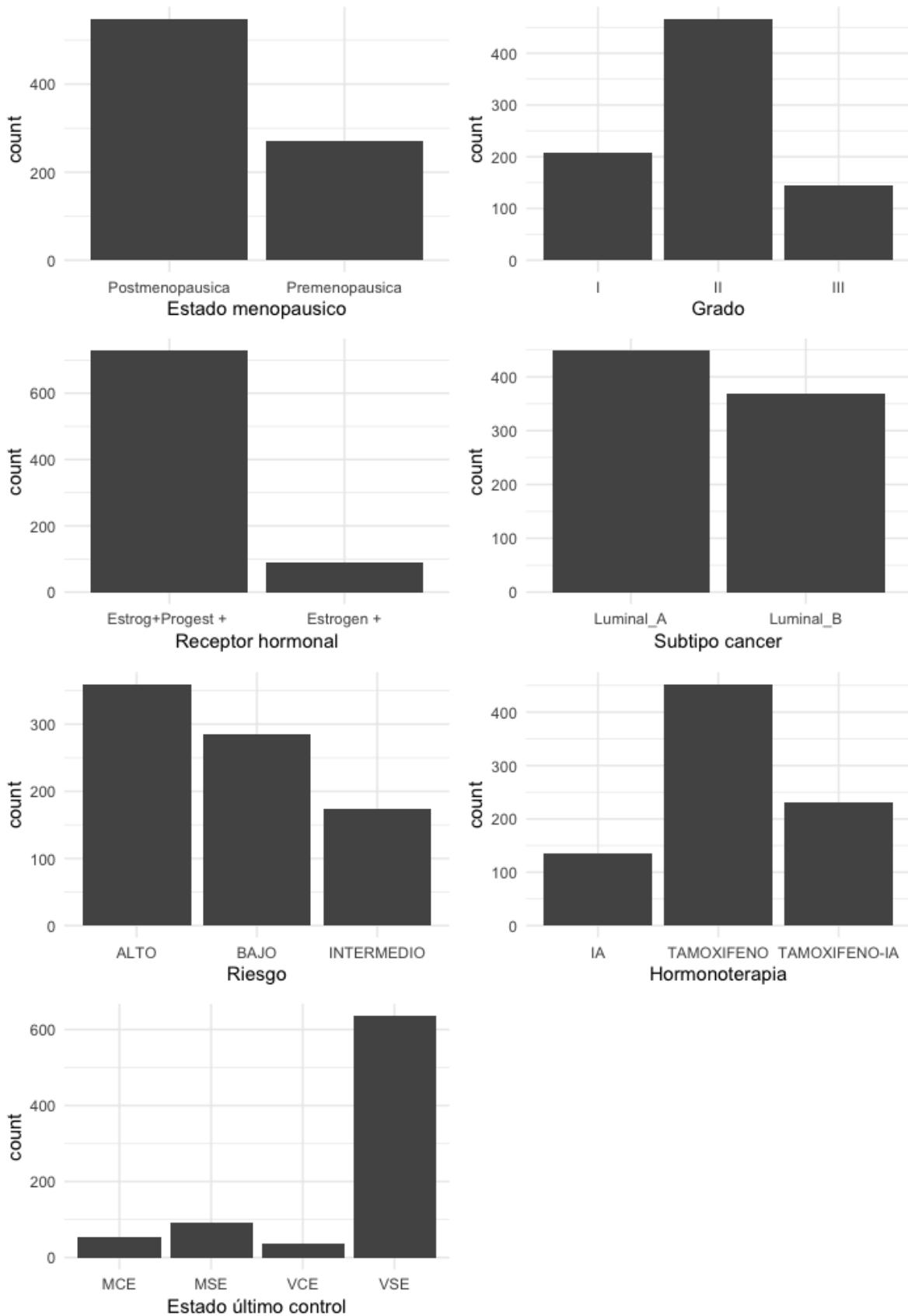


Figura 18: Frecuencia variables categóricas datos Hospital.

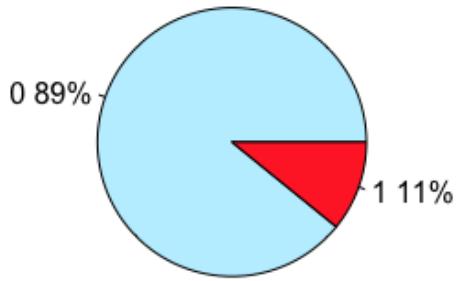


Figura 19: Diagrama de tarta variable *evento* datos Hospital.

Exceptuando la edad que sigue una distribución normal, el resto de atributos se distribuyen de forma *right-skewed*, especialmente el tiempo de seguimiento (habiendo menos observaciones a mayor tiempo transcurrido).

En los diagramas de cajas de la Figura 17 podemos ver qué los valores atípicos son aceptables en todas las variables excepto en los ganglios afectados. Como se ve en la Figura 16 estos valores atípicos no son más que un valor con una frecuencia muy distinta al resto, por lo que, en lugar de eliminarlos haremos un solo nivel con estos datos.

Respecto a las variables categoricas, exceptuando los receptores hormonales y el estado del último control las variables no están desbalanceadas de forma significativa. La variable que define el estado del último control la hemos eliminado debido a que contiene la información de la variable *evento*, la cual queremos predecir. La media y mediana de edad de las pacientes es 58 años, habiendo así más postmenopausicas.

En cuanto a la variable *evento* podemos ver en la Figura 19 que un 89 % de los datos son censurados, como ya estimábamos en secciones anteriores.

4.1.3. Valores eliminados

Inicialmente, el conjunto de datos *Hospital* contaba con 1195 registros y 20 variables. Además de los registros individuales con valores perdidos (*NA*), hemos eliminado las filas correspondientes a los niveles que contenían valores indefinidos, sin valor estadístico, muy desbalanceados o con su mayoría perdidos . Las variables y registros eliminados han sido:

- Variable **Fecha de progresión a distancia** por tener perdidos el 92 % de sus datos.
- Variable **Sexo** por contener solamente el valor **mujer**.

- Variables Fecha diagnóstico y Fecha último control por haberlas sustituido por Años de seguimiento.
- Variables indicadoras de receptores de estrógeno, progesterona y her2 siendo sustituidas por la variable Receptor hormonal.
- Variable Estado ultimo control por la razón explicada previamente.
- En hormonoterapia: eliminamos "análogos" con un solo registro, "análogos-IA" con 3 registros y "tamoxifeno-análogos-IA" con 3 registros.
- En receptor de progesterona: "ND" (no definido) con 5 registros.
- En her2: "2+" con 6 registros.
- En subtipo de cáncer: "Luminal ND" (No definido) con 11 registros.
- En receptores hormonales: aquellos receptivos positivos en progesterona y no en estrógeno (error de laboratorio) con 28 registros.

**Los registros descritos no tienen que ser independientes, pudiendo coincidir más de un caso para el mismo.*

Finalmente tenemos un conjunto de 827 registros y 13 variables. Hemos perdido un 31 % de los datos en la limpieza.

4.2. TCGA

Este conjunto de datos, también de pacientes de cáncer de mama lo hemos obtenido de *The Cancer Genome Atlas* [46], una librería publica con datos de 33 tipos de cánceres. Para ello hemos hecho uso de la librería `TCGAretriever` en R [50]. De estos analizaremos su probabilidad de muerte.

4.2.1. Atributos

De entre todas las variables disponibles en la librería, hemos intentado encontrar un conjunto lo más similar posible al de *Hospital 4.1.1*, además de otras variables de interés.

El conjunto final elegido se compone por:

- **Variables numéricas:** Edad (en años), ganglios afectados y tiempo de seguimiento (en años).
- **Variables categóricas:** Estado menopausico (*Premenopausica*, *Postmenopausica*), receptores hormonales (de estrógeno, progesterona y her2), Etapa tumoral (*I*, *II*, *III*, *IV*), Estadio tumoral (*T1*, *T2*, *T3*, *T4*), receptor hormonal (creada a partir de los receptores hormonales: *Positiva en estrógeno y progesterona*, *Positiva en estrógeno*, *Triple negativa*), subtipo de cáncer (creado a partir del receptor hormonal: *Luminal A*, *Luminal B*, *Triple negativo*, *HER2-enriched*, *Normal-like*), y evento (en este caso significa muerte, con los valores 0, 1).

4.2.2. Distribución y valores atípicos

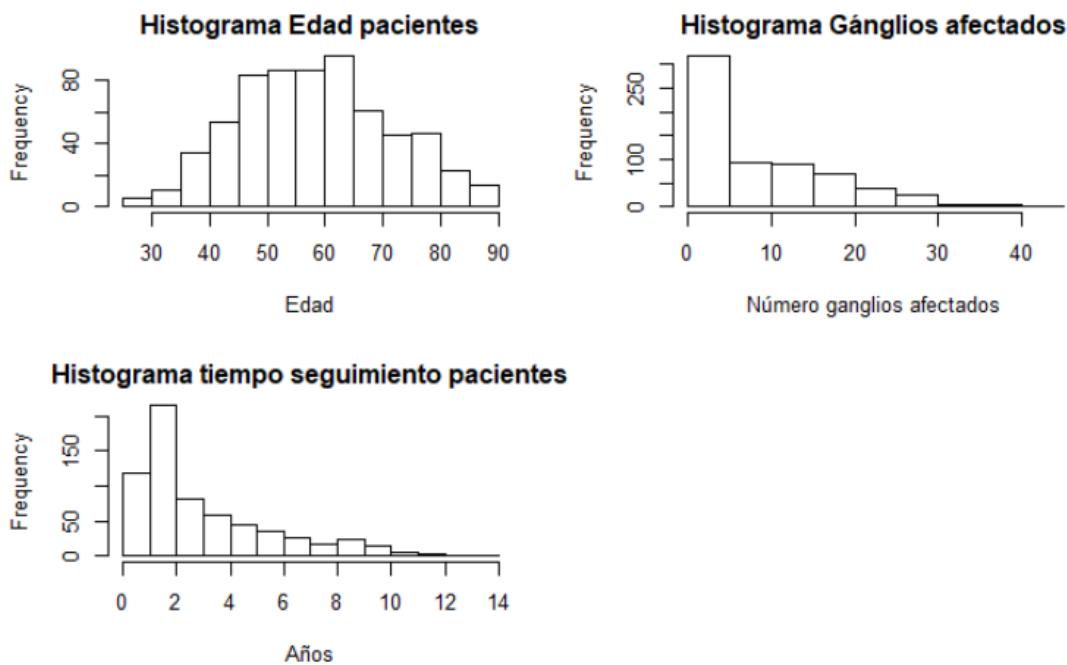


Figura 20: Histogramas distribución variables numéricas datos TCGA.

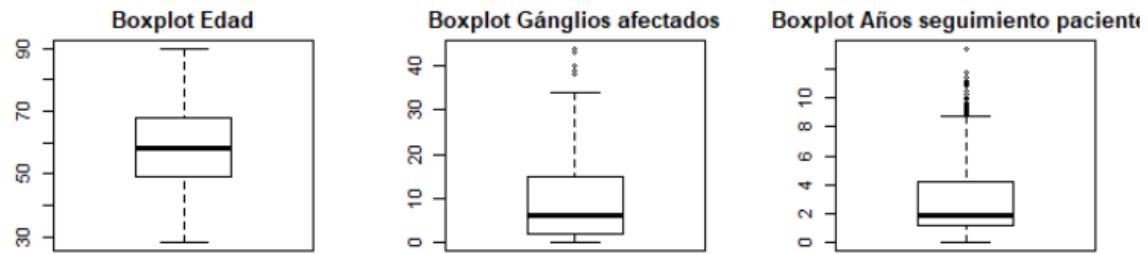


Figura 21: Diagramas de cajas variables numéricas datos TCGA.

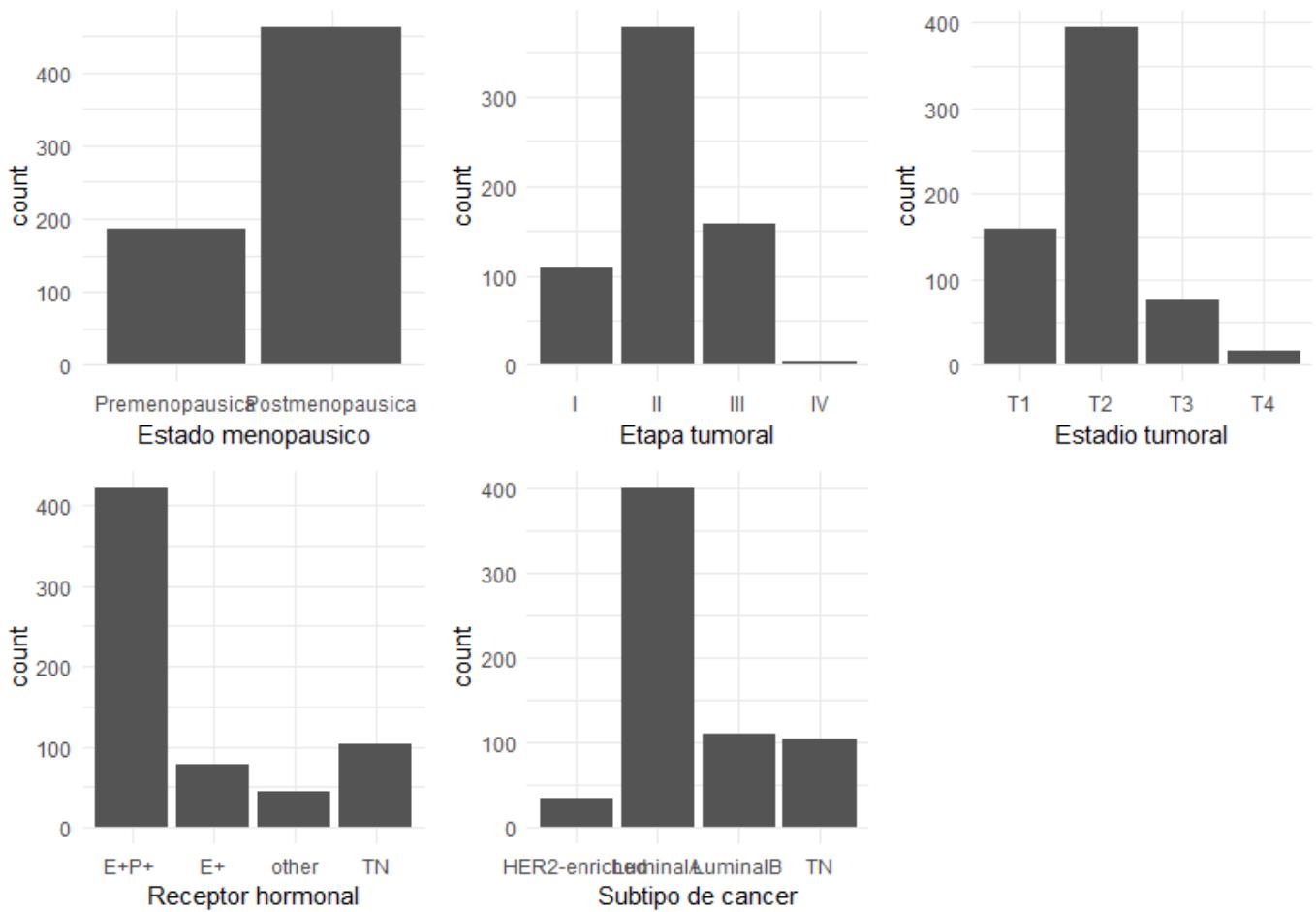


Figura 22: Frecuencia variables categóricas dataset TCGA.

Al igual que ocurre en el conjunto anterior, la edad tiene una distribución normal y el resto de variables numéricas (ganglios afectados y años de seguimiento) tienen sesgoamiento derecho (Figura 20.) Los valores atípicos son aceptables 21.

Respecto a las frecuencias, las pacientes con receptores hormonales de valor estrógeno y progesterona positivos son la mayoría, al igual que en el conjunto anterior y la población clínica estandar. También como en *Hospital*, hay más pacientes postmenopausicas que premenopausicas con la misma media y mediana de edad. El subtipo Luminal A también tiene más concentración de pacientes que el resto, representando bien a la población global (un 30 %-70 % de casos [47].)

En la Figura 23 vemos que el 90 % de los datos son censurados (un 1 % más que en el conjunto *Hospital*).

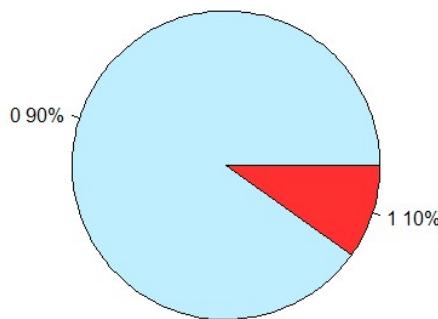


Figura 23: Diagrama de tarta variable *evento* datos TCGA.

4.2.3. Valores eliminados

Inicialmente, este conjunto de datos contaba con 1108 registros y 10 variables. Hemos eliminado:

- Variables indicadoras de receptores de estrógeno, progesterona y her2 siendo sustituidas por la variable **Receptor hormonal**.
- En etapa tumoral: registros nulos e indefinidos ("X"), sumando 28 registros entre los dos casos.
- En estadio tumoral: registros nulos e indefinidos ("TX"), sumando 7 registros entre los dos.
- En receptores hormoanales (rp, re y her2): registros nulos e indefinidos ("Indeterminate"), sumando 192 en her2, 58 en progesterona y 55 en estrógeno.
- En estado menopausico: registros con valor indeterminado ("Indeterminate"), valor presente en 34 registros.

**Los registros descritos no tienen por que ser independientes, pudiendo coincidir más de un caso para el mismo.*

Finalmente tenemos un conjunto de 648 registros y 9 variables. Hemos perdido un 42 % de los datos en la limpieza.

Cuadro 1: Resumen conjuntos de datos

Conjunto de datos	Número de registros	Número de variables predictoras	Número de registros censurados	Tiempo de evento máximo	Significado evento
Hospital	874	12	738 (89 %)	26 años	Recurrencia
TCGA	648	8	584 (90 %)	13 años	Muerte

5. Métodos

El objetivo es encontrar predictores de la supervivencia para cada conjunto de datos , y poder visualizar las curvas predichas a partir de variables de entrada en una interfaz. El primer paso es preparar los datos, para ello, usando métodos estadísticos evaluaremos que estratificación es más adecuada. Intentaremos encontrar un conjunto final óptimo de datos significativos respecto a la variable a predecir y que no estén correlacionados entre ellos. Una vez tenemos un conjunto de datos de calidad y haberlos replicado, evaluaremos si para nuestros datos funcionan mejor predictores de tipo *survival* o de tipo clasificación probabilística. Una vez lo sepamos, exploraremos los diferentes algoritmos correspondientes y sus hiperparámetros, validando los predictores por la media del conjunto de validación en una validación cruzada. El mejor modelo encontrado para cada conjunto será usado como predictor de la interfaz.

5.1. Ingeniería de características

Es el siguiente paso a la limpieza de datos y fundamental para el buen funcionamiento de los algoritmos de aprendizaje automático. Nos permite alcanzar modelos más simples y flexibles. En este proceso se crean variables a partir de las ya existentes, definiendo un conjunto final de datos para entrenar el modelo.

En primer lugar, definiremos diferentes agrupaciones para las variables numéricas de ambos conjuntos, agrupaciones ya definidas y agrupaciones encontradas por aprendizaje automático no supervisado. Una vez creadas las variables, las evaluaremos para ver su calidad y significancia pudiendo decidir así cuales conservar para nuestros conjuntos finales con los que entrenaremos el modelo. Una vez elegidas, aplicaremos análisis de correlación y confirmaremos esta por pruebas de hipótesis por si fuera necesario eliminar alguna de este conjunto final.

5.1.1. Creación de variables

Las agrupaciones que construiremos y evaluaremos, además de las definidas por los algoritmos de *clustering* serán, en las variables que lo permitan, las definidas por el Paper que replicamos como primera tarea de este proyecto [48] y las comúnmente usadas en la práctica médica. Podemos verlas en el Cuadro 2.

Cuadro 2: Agregaciones variables numéricas

Variable	Agrupación Estandar	Agrupación Paper	Aplicable y clusterizable en dataset
Edad	<=50, >50	Disgregada	Hospital y TCGA
Estado nodal (ganglios afectados)	0, 1, 2-3, 4-9, +9	0, 1, 2-3, 4-9, +9	Hospital y TCGA
Etapa tumoral (tamaño tumor en mm)	0, 0-20, 20-50, >50	<10, 10-20, 20-30, >30	Hospital
Concentración ki67	<=14, >14	<10, 10-20, >20	Hospital
Ganglios extraídos	-	-	Hospital

Sabemos que el buen funcionamiento de un algoritmo de clusterización depende en gran parte de la distribución espacial de los datos. En la Figura 24 para el dataset *Hospital* y 25 para el dataset *TCGA* podemos ver que la distribución de las variables a clusterizar está dentro de la normalidad y no tenemos que usar ningún algoritmo específico que se adapte a las mismas. Para estas distribuciones usaremos los algoritmos de propósito general *K – means* 3.1.2 y *Hierarchical* 3.1.3, evaluándolos previamente para ver cual se adapta mejor en cada caso.

Para decidir el número de clusters y algoritmo usado en cada variable seguiremos el siguiente procedimiento:

1. Visualizar los valores de *Silhouette index* y *Elbow method* para cada $k \in (1, 10]$.
2. Ver, para cada valor de k (en un conjunto más reducido al previo) y algoritmo, los valores de *Silhouette index*, *Dunn index* y *Conenctivity*. Aquí veremos, según

cada índice, el número de clusters y método más adecuado para la variable en cuestión.

3. En caso de que la elección sea *hierarchical*, evaluamos el *silhouette index* para las diferentes medidas *average*, *complete* y *single*.
4. Decidido el número de clusters y algoritmo, realizamos las particiones, comprobamos que no haya niveles excesivamente desbalanceados y las visualizamos en el histograma de distribución.

Todas las métricas mencionadas se explican en el apartado 3.1.1.

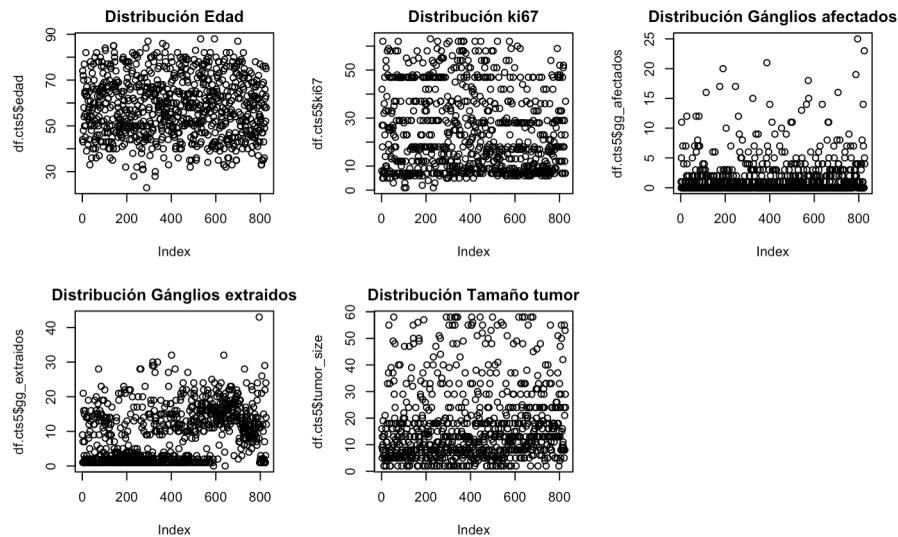


Figura 24: Grafico dispersión variables a clusterizar Hospital.

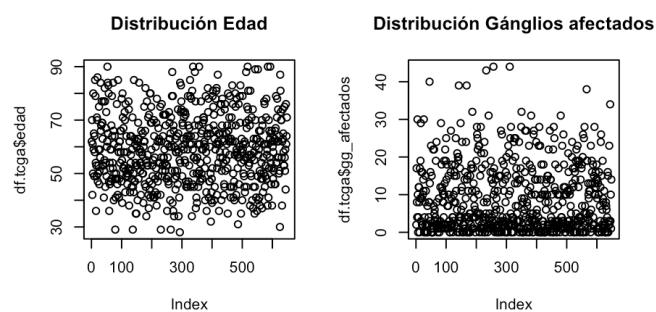


Figura 25: Grafico dispersión variables a clusterizar TCGA.

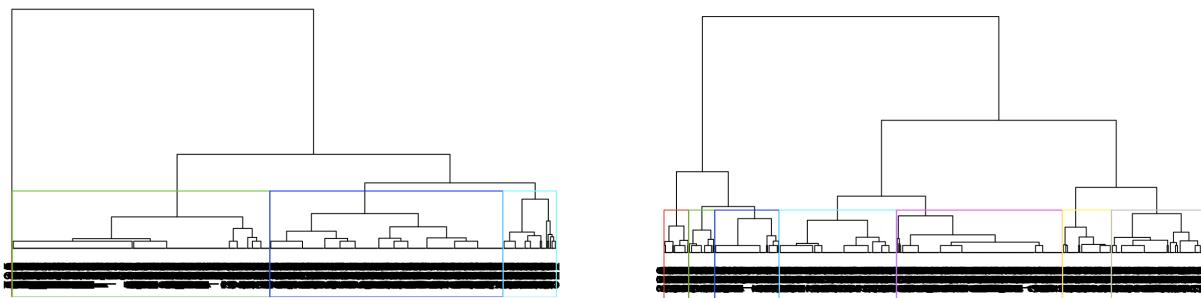
Clusters **Hospital**

- Edad: *K-means* con $k = 4$; $(0, 44]$, $(44, 56]$, $(56, 69]$, $(69, 100]$ contienen 134, 249, 266 y 178 registros respectivamente.
- Ganglios afectados: *K-means* con $k = 2$; $[0, 4]$, $(4, 25]$ contienen 743 y 84 registros respectivamente. *Hierarchical* con 2 clusters además de separar el valor 0 y medida *single*; 0 , $(0, 10]$, $(10, 25]$ contienen 415, 349 y 27 registros respectivamente.
- Ganglios extraídos: *Hierarchical* con 4 clusters y medida *average*; $[0, 1]$, $(1, 11]$, $(11, 17]$, $(17, 32]$ contienen 183, 290, 222 y 131 registros respectivamente.
- Tamaño tumor: *Hierarchical* con 2 clusters y medida *average*; $(0, 27]$, $(27, 60]$ contienen 672 y 155 registros respectivamente. *K-means* con $k = 3$; $(0, 12]$, $(12, 31]$, $(31, 58]$ contienen 373, 328 y 126 registros respectivamente.
- Ki67: *Hierarchical* con 7 clusters y medida *average*; $(0, 11]$, $(11, 20]$, $(20, 31]$, $(31, 40]$, $(40, 49]$, $(49, 56]$, $(56, 63]$ contienen 262, 182, 131, 76, 100, 39 y 37 registros respectivamente.

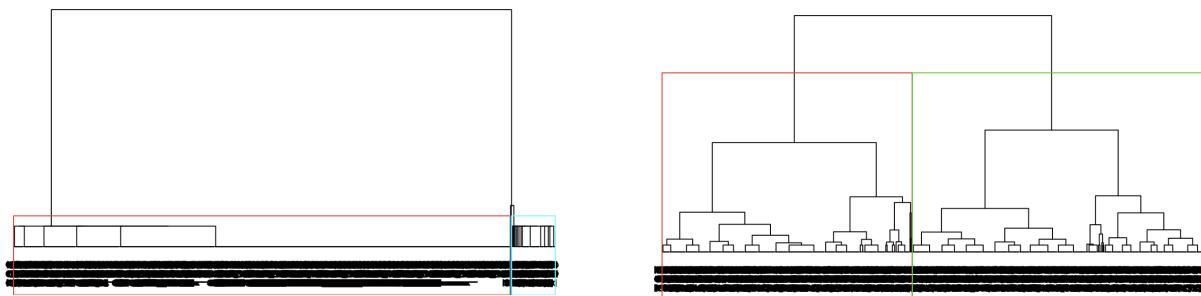
Clusters **TCGA**

- Edad: *K-means* con $k = 4$; $(0, 45]$, $(45, 57]$, $(57, 70]$, $(70, 100]$ contienen 105, 202, 211 y 130 registros respectivamente. *K-means* con $k = 3$; $(0, 54]$, $(54, 70]$, $(70, 100]$ contienen 263, 255 y 130 registros respectivamente.
- Ganglios afectados: *K-means* con $k = 4$; $[0, 5]$, $(5, 12]$, $(12, 21]$, $(21, 44]$ contienen 319, 126, 130 y 73 registros respectivamente. *K-means* con $k = 3$ además del valor 0; 0 , $(0, 8]$, $(8, 19]$, $(19, 44]$ contienen 95, 273, 196 y 84 registros respectivamente.

Para la variable que indica el número de ganglios afectados, debido a la alta concentración del valor 0 en ambos conjuntos (un 55 % de los datos en *Hospital* y un 15 % en *TCGA*), hemos creado un conjunto solo con este valor además de los indicados por el algoritmo.

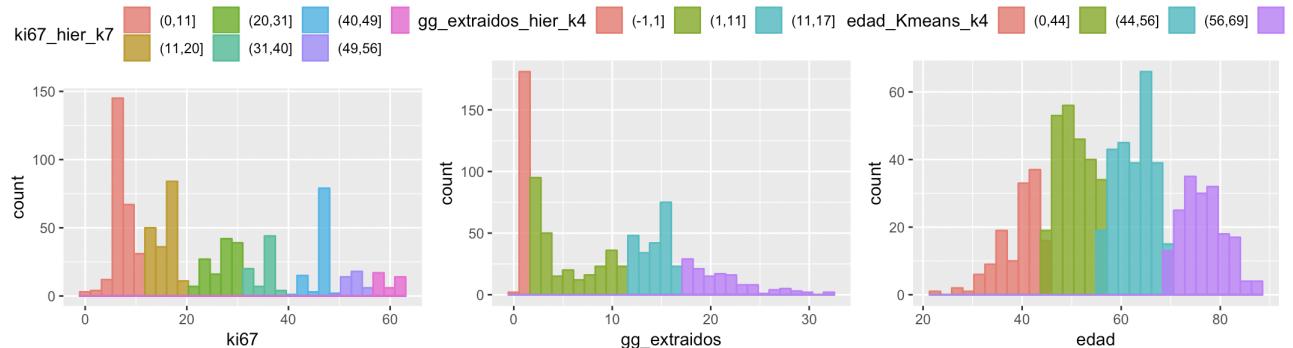


(a) Ganglios extraídos (izquierda) y Ki67 (derecha)

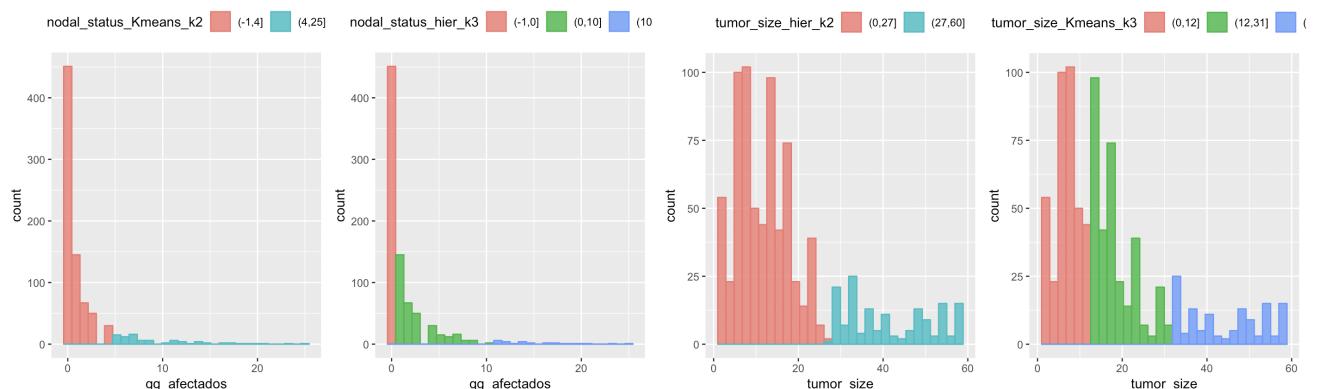


(b) Ganglios afectados (izquierda) y Tamaño tumoral (derecha)

Figura 26: Dendogramas *hierarchical clustering* Hospital

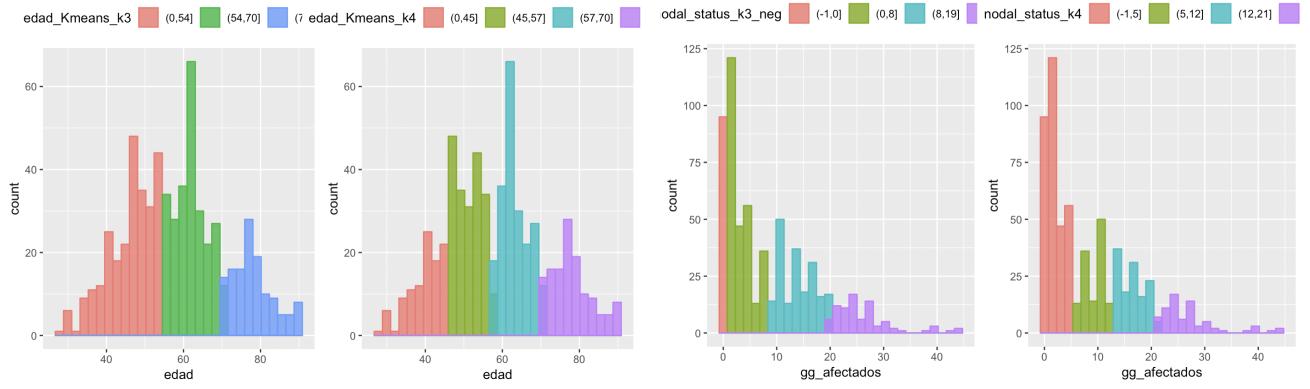


(a) 1 agrupación: Ki67, Ganglios extraídos, Edad



(b) 2 agrupaciones: Ganglios afectados (izquierda) y tamaño tumoral (derecha)

Figura 27: Histogramas clusters Hospital



(a) 2 agrupaciones: Edad (izquierda) y Ganglios aceptados (derecha)

Figura 28: Histogramas clusters TCGA

5.1.2. Análisis y elección de variables

Siguiendo a las 3 distribuciones distintas de variables numéricas, hemos definido a partir de cada conjunto de datos inicial (Hospital y TCGA) y teniendo en común las variables categóricas, 3 conjuntos de datos distintos:

- *Estandar*: contiene las variables numéricas correspondientes a la agrupación estándar (definidas en el cuadro 2).
- Solo para el conjunto *Hospital; Paper*, con las agrupaciones numéricas definidas en el paper y también en el cuadro anterior.
- *Combinado*: Para cada variable numérica, elegiremos una estratificación de entre estándar, paper y definidas por algoritmos de cluster en base al $p-value$ y 95%CI dado por *Kaplan – Meier* y *Cox PH*.

Formados los diferentes datasets referentes a los dos conjuntos de datos originales, evaluaremos si cumplen el test de riesgos proporcionales de Cox y los test de Cox de significancia global de variables, pudiendo así elegir la combinación de variables más adecuada para cada conjunto. Todas las medidas y test estadísticos mencionadas se explican en la sección 2.1.4. Los métodos estadísticos (Cox PH y Kaplan Meier) en 2.2.

Conjunto *Hospital*

1. Análisis estadístico de estratificaciones

Para todos los casos, aunque las de menor $p - value$ según ambos estimadores son las disgregadas, sus intervalos de confianza son muy amplios, por lo que las descartaremos como opciones de calidad. A continuación vemos las estratificaciones de cada variable ordenadas de mayor a menor calidad aportada al conjunto, siendo la primera la elegida para formar parte del dataset *Combinado*. Para esta clasificación nos basamos en los valores del Cuadro 3.

- Edad: *kmeans* k=4, estándar.
- Ganglios afectados: *hierarchical* k=3, *paper* y estándar, *kmeans* k=2
- Ganglios extraídos: *hierarchical* k=4.
- Tamaño tumor: *paper*, *kmeans* k=3, estándar, *hierarchical* k=2.
- Ki67: *paper*, estándar, *hierarchical* k=7.

2. Test de riesgos proporcionales de Cox

Como veíamos en el apartado 2.2.2, una de las suposiciones que hace este modelo es la proporcionalidad de riesgos. Comprobaremos si nuestros conjuntos la cumplen teniendo que descartarlos en caso contrario. Nos basamos desde un principio en la hipótesis nula de que se cumple la proporcionalidad. En caso de que el $p - value$ global sea menor a 0.05, tendremos que descartar esta hipótesis y aceptar que no se cumple.

En el cuadro 4 podemos ver que la proporcionalidad se cumple para los tres conjuntos. También podemos ver que según el estimador *Chi*, ninguno rechaza la hipótesis de independencia, aunque el más propenso a hacerlo sería *Paper*.

3. Test de significancia global de Cox

Los resultados son los mismos para los test de *LogRank* y *Likelihood*. Encontramos ligera diferencia en el test de *Wald* a favor del conjunto *Combinado*, indicando que este conjunto de variables es más significativo. Podemos verlo en el Cuadro 5.

Por esto, la combinación de variables elegida para representar el conjunto *Hospital*, es la de *Combinado*.

Cuadro 3: Análisis calidad variables Hospital
a partir de $p - value$ y 95 %CI

Variable	p-value Kaplan Meier	95 % CI Kaplan Meier	p-value Cox PH
Edad disgragada	< 0,0001	0.576-0.999	0.4
Edad estandar	0,86	0.854-0.933	0.9
Edad kmeans k=4	0,37	0.831-0.943	0.4
Ganglios afectados disgragada	< 0,0001	0.725-0.96	0.005
Ganglios afectados estandar y paper	< 0,0001	0.771-0.942	< 0,0001
Ganglios afectados kmeans k=2	0,0063	0.843-0.929	0.01
Ganglios afectados hierarchical k=3	< 0,0001	0.811-0.927	< 0,0001
Ganglios extraidos disgragada	0,0005	0.642-0.987	0.005
Ganglios hierarchical k=4	0,074	0.825-0.937	0.06
Tamaño tumor disgragada	< 0,0001	0.648-0.982	0.1
Tamaño tumor estandar	0,07	0.838-0.935	0.08
Tamaño tumor paper	0,027	0.823-0.939	0.03
Tamaño tumor hierarchical k=20	0,2	0.852-0.932	0.2
Tamaño tumor kmeans k=3	0,07	0.837-0.936	0.07
Ki67 disgragada	< 0,0001	0.642-0.988	0.5
Ki67 estandar	0,46	0.854-0.933	0.5
Ki67 paper	0,44	0.841-0.943	0.5
Ki67 hierarchical k=7	0,68	-	0.7
Receptor hormonal	0,79	0.855-0.931	0.8
Estado menopausico	0,71	0.855-0.9331	0.7
Grado	0,43	0.842-0.935	0.4
Subtipo de cáncer	0,29	0.854-0.933	0.3
Riesgo	0,0001	0.825-0.919	< 0,0001
Hormonoterapia	0,011	0.836-0.943	0,02

Cuadro 4: Test de riesgos proporcionales de Cox conjuntos Hospital.

Conjunto	Estimador Rho Global	Estimador Chi Global	p-value Global
Estandar	0.037	0.54	1
Paper	NA	0.38	1
Combinado	NA	0.5	1

Conjunto *TCGA*

1. Análisis estadístico de estratificaciones

- Edad: *kmeans* k=3, *kmeans* k=4, estández.
- Ganglios afectados: *kmeans* k=3, estández, *kmeans* k=4.

Cuadro 5: Tests significancia global variables Cox PH model conjuntos Hospital

Conjunto	Test Likelihood Global	Test Wald Global	Test Logrank Global
Estandar	$< 2e - 16$	0.9	$< 2e - 16$
Paper	$< 2e - 16$	0.9	$< 2e - 16$
Combinado	$< 2e - 16$	1	$< 2e - 16$

Vemos los detalles de esta clasificación en el Cuadro 6.

2. Test de riesgos proporcionales de Cox PH

Después de hacer el test una primera vez hemos tenido que eliminar la variable *Estadio* debido a que no cumplía la independencia de tiempo. Además, tenía un *p – value* muy bajo que desequilibraba la media haciendo que la global tuviera valor 1. Una vez eliminada esta variable, aunque el *p – value* sea igual para los dos conjuntos, el estimador *Chi* nos indica que en *Estandar* se rechaza la hipótesis nula de independencia de variables. *Combinado* en cambio sí tiene un conjunto de variables que aportan información distinta. Cuadro 7.

3. Test de significancia global de Cox PH

En el test de *Wald* (frente a una coincidencia de valores en *Logrank* y *Likelihood* confirmamos que *Combinado* tiene más significancia global que *Estandar*. Cuadro 8.

Por esto, la combinación de variables elegida para representar el conjunto *TCGA*, es la de *Combinado*.

Cuadro 6: Análisis calidad variables TCGA
a partir de $p - value$ y 95 %CI

Variable	p-value Kaplan Meier	95 % CI Kaplan Meier	p-value Cox PH
Edad disgragada	< 0,0001	0.489-1	< 0,0001
Edad estandar	0,0014	0.82-0.93	0.0004
Edad kmeans k=3	< 0,0001	0.789-0.95	< 0,0001
Edad kmeans k=4	< 0,0001	0.777-0.955	< 0,0001
Ganglios afectados disgragada	0,54	0.612-0.998	0.4
Ganglios afectados estandar y paper	0,006	0.788-0.965	0,02
Ganglios afectados kmeans k=3 + neg	0,001	0.785-0.967	0.004
Ganglios afectados kmeans k=4	0,61	0.796-0.962	0,6
Estado menopausico	< 0,0001	0.816-0.924	< 0,0001
Etapa tumoral	0,00029	0.809-0.942	0,001
Estadio tumoral	< 0,0001	0.788-0.96	0,01
Receptor hormonal	0,44	0.803-0.949	0,5
Subtipo de cáncer	0,18	0.798-0.958	0,2

Cuadro 7: Test de riesgos proporcionales de Cox conjuntos TCGA.

Conjunto	Estimador Rho Global	Estimador Chi Global	p-value Global
Estandar	NA	0,009	1
Combinado	NA	0.591	1

Cuadro 8: Tests significancia global variables Cox PH model conjuntos TCGA

Conjunto	Test Likelihood Global	Test Wald Global	Test Logrank Global
Estandar	< 2e - 16	0,6	< 2e - 16
Combinado	< 2e - 16	0,9	< 2e - 16

5.1.3. Correlación e hipótesis

Usando el coeficiente de correlación de *Spearman* 2.1.4 veremos la fuerza de asociación de las variables para comprobar si es conveniente eliminar alguna que esté muy correlacionada con otra del conjunto final. Estudiaremos las hipótesis de las posibles correlaciones observadas por el índice de *Spearman*. Los test de hipótesis se realizan mediante la función `survdiff()` [49] que evalúa la diferencia entre curvas de supervivencia. Entendemos H_0 como independencia de variables.

Conjunto *Hospital*

Vemos las correlaciones observadas en la Figura 29 y correspondiente resultado del los test de hipótesis en el Cuadro 13. El test de independencia no se cumple para tamaño tumoral y riesgo, aún así, decidimos no eliminarlas debido a que el índice de correlación no es demasiado alto (0.51) y ambas tienen algo de correlación con la variable evento. No observamos ninguna correlación inversa significativa.

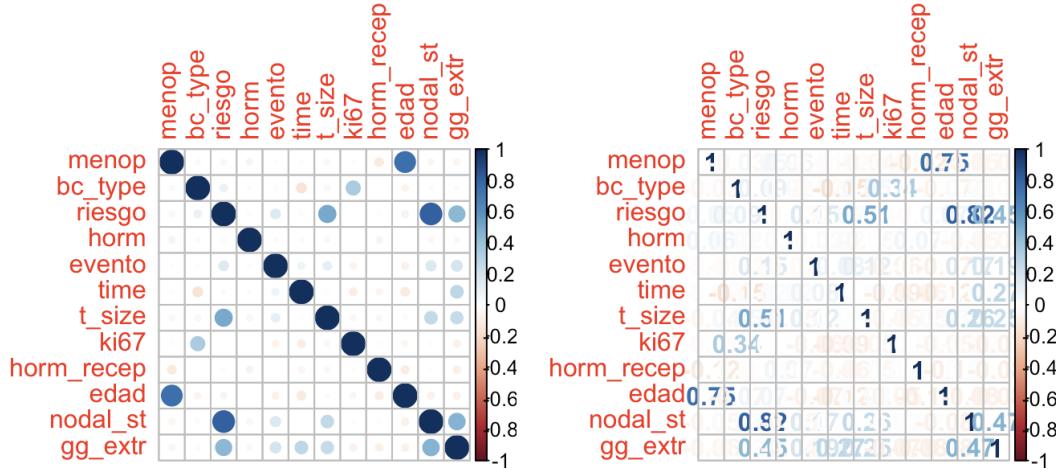


Figura 29: Valores de correlación *Spearman* a dataset final Hospital.

Cuadro 9: Pruebas de hipótesis correlaciones Hospital

Correlación observada	Índice correlación	p-value test	Hipótesis
Estado nodal y riesgo	0.82	0,06	H_0
Edad y estado menopausico	0.75	0,3	H_0
Tamaño tumoral y riesgo	0.51	0,004	H_1

Conjunto *TCGA*

Vemos las correlaciones observadas en la Figura 30 y correspondiente resultado del los test de hipótesis en el Cuadro 13. Por el test de hipótesis aceptamos la correlación entre menopausia y edad, pero al igual que antes, al no ser una correlación con un índice demasiado alto, y en este caso contar con número de variables total muy reducido no eliminaremos ninguna de ellas.

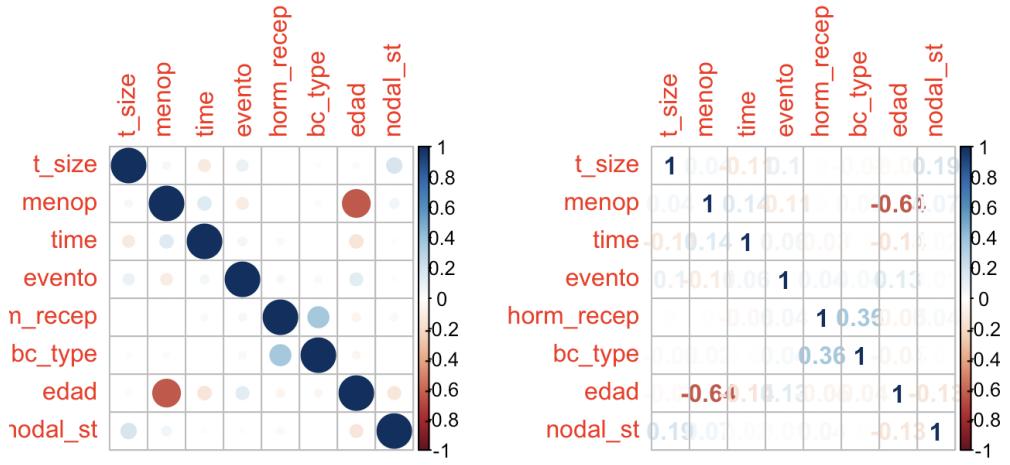


Figura 30: Valores de correlación *Spearman* a dataset final TCGA.

Cuadro 10: Pruebas de hipótesis correlaciones TCGA

Correlación observada	Índice correlación	p-value test	Hipótesis
Subtipo cáncer y receptor hormonal	0.43	0,6	H_0
Menopausia y edad	-0.67	0,002	H_1

5.1.4. Conjuntos finales

Conjunto *Hospital*

Cuadros 13 y 11. La mediana de supervivencia es nula debido a que la supervivencia nunca alcanza el valor 0.5 (50 %), manteniéndose siempre por encima.

Cuadro 11: Ajuste supervivencia Kaplan Meier del conjunto Hospital.

n	eventos	media	SE media	mediana	95 % CI
826	88	21.82	0.42	NA	NA NA

En ninguno de los conjuntos destaca la significancia de ninguna variable.

Conjunto *TCGA*

Cuadros 14 y 12.

Cuadro 12: Ajuste supervivencia Kaplan Meier del conjunto TCGA.

n	eventos	media	SE media	mediana	95 % CI
648	64	10.03	0.41	10.23	9.51 NA

Cuadro 13: Índice de impacto, error estándar, índice de riesgo, p-value e intervalos de confianza 95 % usando el modelo de Cox en conjunto Hospital.

Variable	Coeficiente impacto	Hazard Ratio	Error estandar	p-value	95 % CI HR
Postmenopausica	-3.54	0.28	6.08	0.56	1.9e-07 4370
Luminal B	-0.76	0.46	3.14	0.81	9.8e+04 222
riesgo INTERMEDIO	-0.59	0.55	0.89	0.88	1.6e+04 222
riesgo ALTO	-2.15	1.97	6.52	0.74	3.2e-07 1850
TAMOXIFENO	-2.35	0.12	3.39	0.49	1.2e-04 4.15e+04
TAMOXIFENO-IA	-1.73	0.18	4.01	0.66	6.8e-05 73.3
Tumor size (10,20]	-2.09	0.12	3.81	0.58	7.05e-5 218
Tumor size (20,30]	0.22	1.25	5.01	0.96	7.05e-05 2.29e+04
Tumor size (30,100]	-2.18	0.11	4.38	0.62	6.81e-05 602
Ki67 Borderline	2.05	7.78	4.92	0.68	4.98e-04 1.21e+05
Ki67 High	2.11	8.23	4.04	0.60	0.003 2.27e+04
hormone receptor EP+RP	1.48	4.39	4.00	0.71	1730 1.11e+04
Edad (44,56]	2.16	8.66	5.63	0.70	1.37e-04 5.45e+05
Edad (56,69]	3.48	32.35	6.41	0.58	1.13e-04 9.23e+06
Edad (69,100]	3.76	43.11	8.40	0.65	3.03e-06 6.13e+08
Nodal status (0,10]	0.88	2.41	4.49	0.84	3.66e-04 1.58e+04
Nodal status (10,25]	1.35	3.87	6.50	0.83	1.13e-05 1.31e+06
gg extraidos (1,11]	1.05	2.84	4.42	0.81	4.86e-04 1.66e+04
gg extraidos (11,17]	1.27	3.57	4.57	0.78	4.60e-04 2.77e+04
gg extraidos (17,32]	8.17	1.08	5.61	0.98	1.82e-05 6.46e+04

Cuadro 14: Índice de impacto, error estándar, índice de riesgo, p-value e intervalos de confianza 95 % usando el modelo de Cox en conjunto TCGA.

Variable	Coeficiente impacto	Hazard Ratio	Error estandar	p-value	95 % CI HR
Premenopausica	1.35	3.8	3.168	0.66	7e-03 1.930e+03
Luminal A	-1.05	0.34	4709	0.99	0 Inf
Luminal B	1.64	0.19	4709	0.99	0 Inf
Tumor stage II	-0.97	0.37	2.89	0.73	1.294e-03 1.105e+02
Tumor stage III	-0.16	0.85	2.54	0.94	5.852e-03 1.238e+02
Tumor stage IV	-7.27	6.9e-04	2.0e+04	0.99	0 Inf
Hormone receptor EP	-1.18	0.30	2.36	0.62	3.011e-03 3.132e+01
Hormone receptor Other	-2.94	0.05	4709	0.99	0 Inf
Hormone receptor TN	-2.67	0.07	4709	0.99	0 Inf
Edad (54,70]	-0.48	0.61	3.75	0.89	3.871e-04 9.709e+02
Edad (70,100]	0.08	1.08	3.29	0.97	1.693e-03 7.004e+02
Nodal status (0,8]	1.43	4.19	2.30	0.53	4.58e+02 3.83e+02
Nodal status (19,44]	1.30	3.68	2.38	0.58	3.446e-026 3.931e+02
Nodal status (8,19]	0.35	1.42	2.2	0.87	1.798e-02 1.132e+02

5.2. Estudio de los efectos de la quimioterapia y hormonoterapia

La primera tarea que realizamos como parte de este proyecto fue un análisis de la supervivencia y la utilidad de rendimiento de la hormonoterapia y quimioterapia solicitado por la doctora que nos proporcionó los datos del conjunto *Hospital*.

Por ello, esta tarea se realiza únicamente con las variables originales (y limpiadas) de este conjunto y las agrupaciones correspondientes del paper con el que se comparan los resultados [48].

Medidas de valoración de utilidad o rendimiento:

- RR: Riesgo relativo de muerte de pacientes que reciben el tratamiento relativo al grupo de control. $RR = \frac{\text{expuestos al tratamiento}}{\text{no expuestos al tratamiento}}$.
- RRR: Reducción relativa del riesgo. El tanto por ciento que el tratamiento reduce el riesgo de muerte. $RRR = (1 - RR) * 100$.
- RRI: Aumento relativo del riesgo. $RRI = \frac{ARI}{\text{tasa grupo control}}$.
- RAR: Reducción Absoluta del riesgo, tanto por ciento de personas en las que se puede evitar la muerte por aplicar el tratamiento. $RAR = (\text{no expuestos al tratamiento} - \text{expuestos al tratamiento}) * 100$.
- NNT: Número necesario de pacientes a tratar para reducir un evento (recurrencia).
$$NNT = \frac{1}{RAR}$$
- ARI: Aumento absoluto del riesgo. $ARI = \text{tasa experimental} - \text{tasa grupo control}$.
- NNH: Número de pacientes que es necesario tratar para que un paciente sufra un evento adverso. $NNH = \frac{1}{ARI}$.

Rendimiento quimioterapia:

Como vemos en el Cuadro 15 y como se observaba en el paper, la exposición a quimioterapia aumenta la tasa de muerte de forma notable, más a mayor gravedad de la paciente. En el cuadro 16 comprobamos que la quimioterapia no reduce el riesgo de muerte sino todo lo contrario, por esa razón no tiene sentido calcular el número necesario de pacientes a exponer a quimioterapia para reducir un evento (NNT). En un test de hipótesis con el

Cuadro 15: Tasa de muerte por grupo riesgo según exposición a quimioterapia conjunto Hospital

Grupo Riesgo	Expuestos quimioterapia	No expuestos quimioterapia
Global	14 %	5.9 %
Bajo	6 %	5.4 %
Intermedio	7.7 %	5.5 %
Alto	19 %	7.4 %

Cuadro 16: Valoración utilidad exposición a quimioterapia

Grupo Riesgo	RRR	NNT
Bajo	-10.9 %	-167
Intermedio	-36.5 %	-49
Alto	-157.3 %	-9

método *Kaplan Meier* rechazamos la hipótesis de que el tratamiento de quimioterapia no causa efectos sobre la supervivencia con un $p - value = 3e - 04$. Con el estimador de Cox vemos que la variable quimioterapia tiene un coeficiente de impacto de 0,84, un índice de riesgo de 2,33 con un error estándar de 0,24 y un $p - value$ 0,0004.

Rendimiento hormonoterapia:

Cuadro 17: Aceptación de tratamientos hormonoterapia

Tratamiento hormonoterapia	Tasa grupo control	Tasa grupo experimental	NNT	NNH
IA	0.106	0.105	19	-19
Tamoxifeno	0.101	0.121	-53	53
IA+Tamoxifeno	0.108	0.045	30	-30

Con el número positivo mínimo de pacientes a tratar para evitar un caso de recurrencia, tenemos a los Inhibidores de Aromatasa (*IA*) como mejor tratamiento, seguido por este combinado con *Tamoxifeno*. Cuadro 17.

Por recomendación de la doctora, dejamos de trabajar con la variable quimioterapia desde el principio.

5.3. Propuesta de tratamiento hormonal basado en la supervivencia

5.3.1. Replicación datos

Los algoritmos de aprendizaje automático aprenden entendiendo la relación entre las variables predictores y la variable a predecir para poder crear un patrón. Además, sea cual sea el tamaño de los datos, es una pequeña fracción de la población global, y en la naturaleza siempre hay fluctuaciones que difieren de un patrón estricto. Nuestro objetivo es encontrar el patrón de toda la población con una muestra de tamaño N , y, a mayor N , menores las diferencias habrá con la población global.

Nuestros dos conjuntos de 827 y 648 registros son muy pequeños e insuficientes para representar a la población global, por lo que es necesario replicarlos antes de poder usarlos para entrenar un modelo.

El algoritmo de replicación de datos diseñado, consiste en, para cada registro:

Se almacena su `evento` y `tiempo_seguimiento` (en meses). Se asigna un tiempo de inicio y un intervalo acorde al `tiempo_seguimiento`: si es mayor que 10 meses, el tiempo de inicio será 10 meses anteriores a ese mes, si no, será 0. El intervalo será una décima parte de la diferencia entre el tiempo final y `tiempo_seguimiento`.

- Desde su tiempo de inicio definido hasta `tiempo_seguimiento` y con intervalo 4 si es el dataset *Hospital* o 3 si es el dataset *TCGA* se replicarán el resto de variables con el evento no observado para cada tiempo.
- Si el evento ha tenido lugar (`evento=1`), desde `tiempo_seguimiento` hasta el tiempo máximo de ese conjunto y con el intervalo calculado, se replican el resto de variables.

Pseudocódigo del algoritmo de replicación:

```
last.time<-max(df$seguimiento_months)
step_censura=ifelse(df.hospital, 4, 3)

for (patient in pacientes) {
    status<-df$evento[patient]
    time<-df$seguimiento_months[patient]
    start_time=ifelse(time>10,time-10, 0)
    step_evento=(last.time-time)/10

    #alive before time
    for(month in seq(start_time,time,by=step_censura)) {
        new.row<-c(valores variables, 0, month)
        replicated.df<-rbind(replicated.df, new.row)
    }

    #not censored: if dead, still dead after time
    if(status==1) {
        for(month in seq(time,last.time,by=step_evento)) {
            new.row<-c(valores variables, 1, month)
            replicated.df<-rbind(replicated.df, new.row)
        }
    }
}
```

La idea principal es que un paciente con un tiempo $t \in [0, T]$, no había sufrido el evento en un tiempo $t_i \leq t$, y si lo ha sufrido en el tiempo t , también lo habrá sufrido para un tiempo $t_i \geq t$. En los registros que replican en tiempo atrás, solo lo hacen hasta un tiempo proporcional, y los que van al final, lo hacen también con un intervalo proporcional a su posición para que se mantengan las distribuciones de los valores de las variables. Además, hemos hecho el intervalo que replica los datos con `evento=1` menor para reducir el posiblemente problemático sesgamiento inicial de la variable `evento`.

Como podemos ver en los gráficos contenidos en los documentos anexos del proyecto y no expuestos aquí por ser demasiados, las distribuciones de frecuencia de todas las

variables se mantienen casi a la perfección para ambos conjuntos. La distribución más importante de mantener es la de eventos en el tiempo. Podemos verla en las Figuras 31 y 32 de los conjuntos *Hospital* y *TCGA* respectivamente. En el cuadro 31 podemos ver como manteniendo las distribuciones, hemos conseguido reducir los datos censurados en un 17% y 12%.

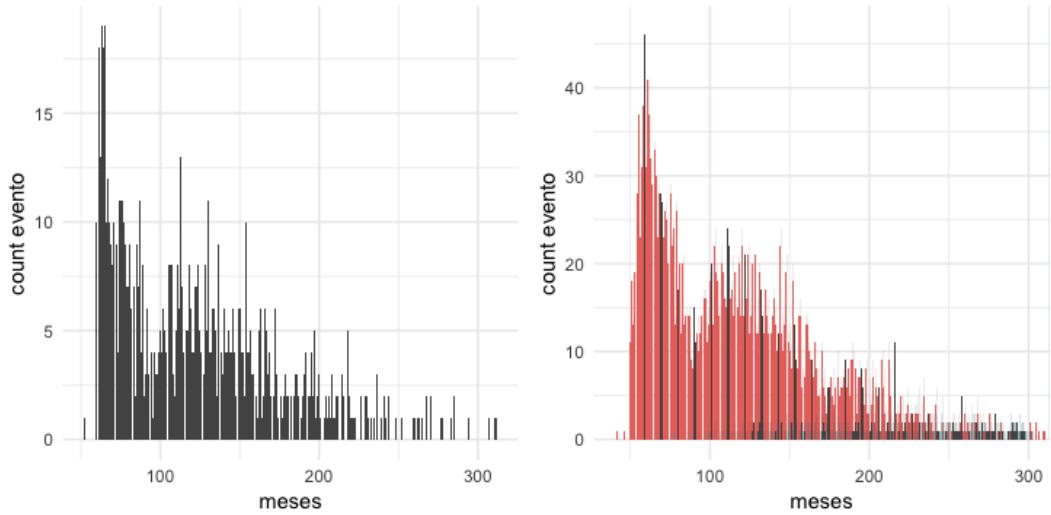


Figura 31: Distribución evento en el tiempo en conjunto *Hospital*

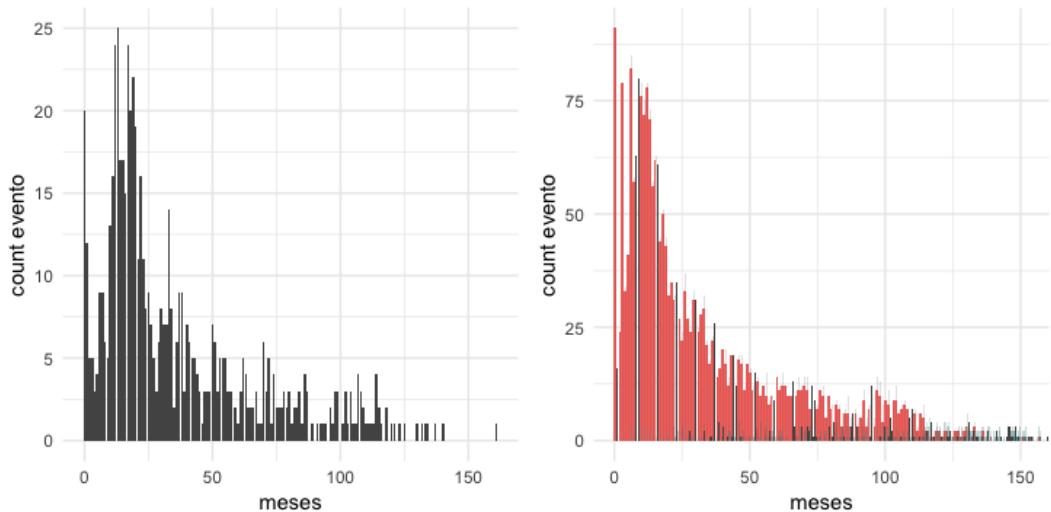


Figura 32: Distribución evento en el tiempo en conjunto *TCGA*.

Cuadro 18: Estado conjuntos antes y después de la replicación.

Conjunto	Tamaño incicial	Tamaño tras replicación	Censurados inicialmente	Censurados tras replicación
Hospital	827	3422	738 (89 %)	2454 (72 %)
TCGA	648	3137	584 (90 %)	2433 (78 %)

5.3.2. Usando *mlr*: Elección del tipo de predictor

Teniendo los datos preparados para usar un modelo de predicción, haremos una exploración para saber que tipo de predictor es más adecuado. Exploraremos el funcionamiento de predictores de tipo *Survival* (los usados de forma estándar en este tipo de estudios) y predictores de Clasificación probabilística. Estos últimos, para cada nivel de la variable a predecir (0 u 1) calcularán su probabilidad y nos devolverán el más probable como predicción.

Para esta exploración haremos uso del paquete *mlr* para R [51]. Este paquete (acrónimo de *machine learning in r*) proporciona una infraestructura que, usando métodos de *resampling*, evalúa los algoritmos indicados haciendo un ajuste interno para encontrar los mejores hiperparámetros en cada caso.

Para usar esta infraestructura, introducimos los conceptos *Task* y *Learner*. Los *Tasks* encapsulan los datos e información de estos necesaria para el algoritmo de aprendizaje automático. Los crearemos con las funciones `makeSurvTask()` indicando la variable de tiempo de seguimiento y el evento, y `makeClassifTask()` indicándole la variable evento y el nivel que define la supervivencia. Los *Learners* contienen las propiedades de los métodos, los creamos con la función `makeLearner()` y les indicaremos el algoritmo y tipo de predicción en cada caso. También definiremos las características de la estrategia de *resampling* con la función `makeResampleDesc()`. En nuestro caso hemos elegido *Cross – Validation* con $k = 5$ (técnica explicada en el apartado 3.3). Finalmente, con la función `resample(learner, task, resampling)` indicándole los parámetros creados de forma previa obtendremos las medidas de validación medias del conjunto de validación en las iteraciones del algoritmo de *resampling*. Para ambos conjuntos probaremos sus algoritmos disponibles en cada caso.

Predictores *Survival*:

Lo evaluamos con la medida $C - index$ 2.2.3. Como hemos visto en la explicación de esta métrica, un valor menor a 0,7 define al modelo como muy débil e inconsistente. Ninguno de los algoritmos da buenos resultados, y los que son algo mejores tienen un tiempo de ejecución alto. El conjunto *TCGA* funciona mejor en este tipo de predictores que el conjunto *Hospital*. Podemos ver los resultados en los cuadros 19 y 20.

Cuadro 19: Resultados clasificadores *Survival* usando *mlr Hospital*.

Algoritmo	C-Index	Tiempo de ejecución
<i>CV CoxBoost</i>	0.58	8.6 min
<i>CoxBoost</i>	0.58	1.8 min
<i>CoxPH</i>	0.57	1 sec
<i>CV Glmnet</i>	0.5	6 secs

Cuadro 20: Resultados clasificadores *Survival* usando *mlr TCGA*.

Algoritmo	C-Index	Tiempo de ejecución
<i>CV CoxBoost</i>	0.70	7.2 min
<i>CoxBoost</i>	0.69	26 secs
<i>CoxPH</i>	0.68	0.2 secs
<i>CV Glmnet</i>	0.62	4 secs

Predictores Clasificación probabilística:

Los evaluamos con las medidas *ACC* y *MMCE* 3.2.1. En estos predictores también coincidimos en el mejor funcionamiento sobre el conjunto *TCGA*, aunque para ambos conjuntos los resultados son notablemente mejores que los de los predictores *Survival*, y presentan tiempos de ejecución más estables entre los diferentes algoritmos. podemos ver los resultados en los cuadros 21 y 22. Destacamos el buen funcionamiento del algoritmo *Bart Machine (Bayesian Additive Regression Trees)* sobre el resto, por lo que nos centraremos en explorar los algoritmos derivados de árboles de decisión (*Random Forest* y *Gradient Boosting*).

Cuadro 21: Resultados clasificadores *Clasificación* usando *mlr Hospital*.

Algoritmo	ACC	MMCE	Tiempo de ejecución
<i>Bart Machine</i>	0.88	0.12	1.3 mins
<i>AdaBoost machine</i>	0.83	0.17	0.32 secs
<i>Binomial</i>	0.85	0.15	0.2 secs

Cuadro 22: Resultados clasificadores *Clasificación* usando *mlr TCGA*.

Algoritmo	ACC	MMCE	Tiempo de ejecución
<i>Bart Machine</i>	0.99	0.10	62 secs
<i>Binomial</i>	0.89	0.11	0.3 secs
<i>AdaBoost machine</i>	0.82	0.14	0.7 secs

5.3.3. Usando *h2o*: Análisis exploratorio de modelos clasificación

Una vez sabemos por el paso anterior que vamos a usar predictores de tipo clasificación probabilística, usaremos la librería *h2o* en R [52]. Esta plataforma abierta ofrece implementaciones paralelizadas de algoritmos de aprendizaje computacional. La usaremos para automatizar el proceso de entrenamiento de los algoritmos explicados en la sección 3.2 (*Deep Learning*, *Random Forest*, *Gradient Boosting Machine* y *Extreme gradient boosting machine*). En el apartado anterior hemos visto el buen funcionamiento de los algoritmos derivados de los árboles de decisión, pero no dejaremos de comprobar también el comportamiento del tan conocido algoritmo de aprendizaje profundo.

Al igual que en el análisis del apartado anterior, también evaluaremos las medidas medias del conjunto de validación en una validación cruzada $k = 5$. Estas medidas de calcularán a partir de la diferencia entre la predicción devuelta (0 o 1) con el valor real de **evento**. Una vez hayamos validado un modelo, por ser este de tipo probabilístico, podremos acceder a la probabilidad de que la salida sea 0, es decir, la probabilidad de sobrevivir. Estas probabilidades para los distintos tiempos definirán una curva de supervivencia.

Dado que es destacable el desbalanceo en los niveles en la variable a predecir como hemos visto en el Cuadro 18, probaremos los modelos con la variable **evento** balanceada y sin balancear para ambos conjuntos.

5.3.4. Decisión de hormonoterapia en base a la supervivencia predicha

Usando la extensión Shiny para R [53], crearemos una interfaz. En esta interfaz se podrán introducir los datos clínicos de la paciente disponibles en cada conjunto en la forma de las agrupaciones finalmente elegidas (ver en 5.1.4), además de un tiempo fijo. Haciendo uso de los modelos finalmente elegidos y validados para cada conjunto (ver en sección 6) predeciremos para los datos introducidos la probabilidad de sobrevivir (de que `evento` tenga el valor 0) para cada tiempo, devolviendo una curva de supervivencia. En el caso del conjunto *Hospital*, en el cual hemos contado con la variable `hormonoterapia` con los valores `IA`, `Tamoxifeno`, `Tamoxifeno+IA` para el entrenamiento, podremos ver curvas diferenciadas de supervivencia para cada uno de los tratamientos, pudiendo así saber cual es el que hace máxima la supervivencia en cada caso particular. Se devolverán también los valores exactos de supervivencia para el tiempo fijo introducido.

Curva supervivencia paciente

Introducir los datos clínicos de la paciente en el dataset deseado, esperar para visualizar la curva de supervivencia completa así como el valor de probabilidad de supervivencia exacto para el tiempo fijo.

Datos clínicos CTS5	
Año fijo de supervivencia	<input type="text" value="10"/>
Estado menopausico	<input type="text" value="Postmenopausica"/>
Subtipo de Cancer	<input type="text" value="Luminal_A"/>
Riesgo	<input type="text" value="ALTO"/>
Grupo tumoral	<input type="text" value="(10,20]"/>
Estado ki67	<input type="text" value="Borderline (>=10% y <=20%)"/>
Receptores hormonales	<input type="text" value="Estrogen-and-Progesterone-receptor-positive"/>
Grupo Edad	<input type="text" value="(0,44]"/>
Estado nodal	<input type="text" value="(0,10]"/>
Ganglios extraidos	

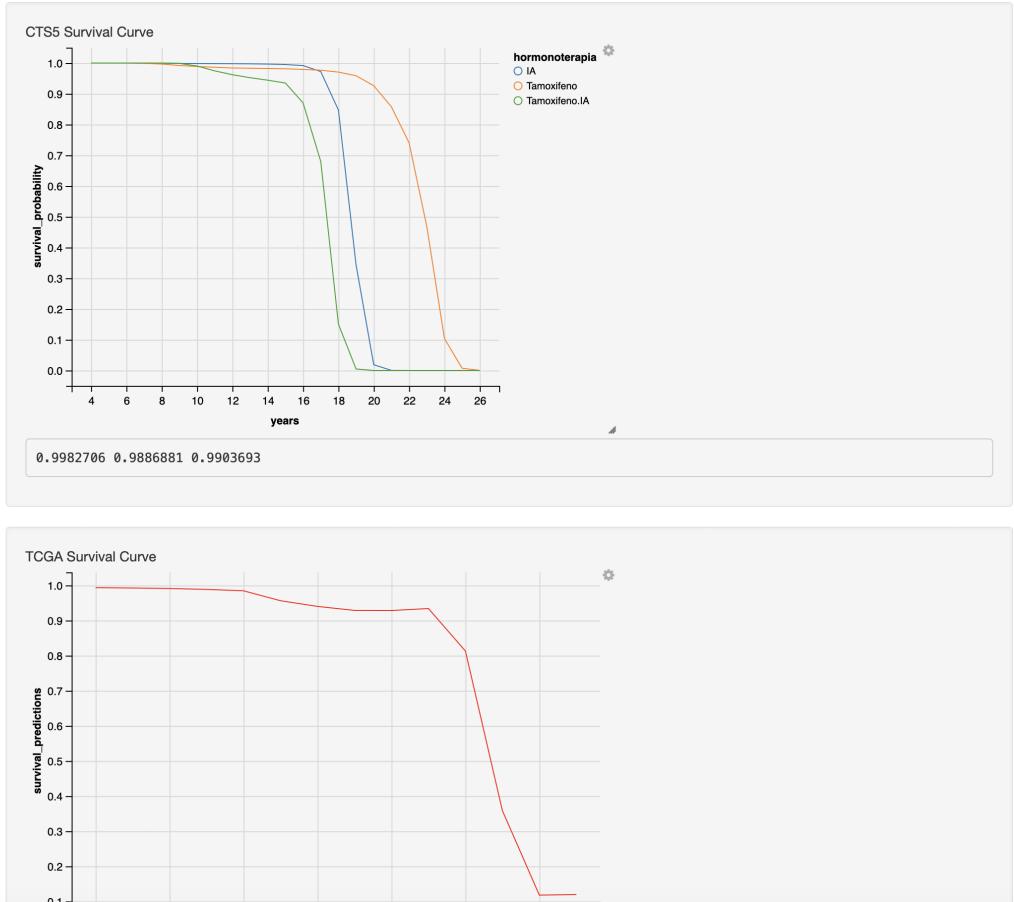


Figura 33: Previsualización interfaz.

*CTS5 hace referencia al conjunto Hospital.

6. Resultados

Resultados medios del conjunto validación con predictores de clasificación.

Conjunto *Hospital*

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.628 ± 0.044	0.543 ± 0.095	0.658 ± 0.032	0.146 ± 0.025	0.342 ± 0.032
<i>XGBoost</i>	0.692 ± 0.040	0.789 ± 0.112	0.650 ± 0.022	0.094 ± 0.013	0.349 ± 0.022
<i>GBM</i>	0.658 ± 0.043	0.767 ± 0.054	0.661 ± 0.056	0.106 ± 0.016	0.339 ± 0.056
<i>RF</i>	0.654 ± 0.039	0.763 ± 0.104	0.649 ± 0.027	0.097 ± 0.015	0.351 ± 0.027

Cuadro 23: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *Hospital* con variables estandar y sin replicar.

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.974 ± 0.002	0.926 ± 0.006	0.911 ± 0.005	0.059 ± 0.002	0.089 ± 0.005
<i>XGBoost</i>	0.964 ± 0.003	0.915 ± 0.006	0.888 ± 0.005	0.065 ± 0.002	0.112 ± 0.005
<i>GBM</i>	0.963 ± 0.003	0.915 ± 0.005	0.882 ± 0.005	0.066 ± 0.002	0.118 ± 0.005
<i>RF</i>	0.961 ± 0.003	0.916 ± 0.005	0.881 ± 0.005	0.069 ± 0.002	0.119 ± 0.005

Cuadro 24: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *Hospital* replicado y sin balancear .

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.974 ± 0.003	0.931 ± 0.005	0.917 ± 0.005	0.010 ± 0.003	0.082 ± 0.010
<i>XGBoost</i>	0.966 ± 0.005	0.925 ± 0.005	0.898 ± 0.010	0.061 ± 0.005	0.102 ± 0.010
<i>GBM</i>	0.965 ± 0.006	0.921 ± 0.011	0.889 ± 0.005	0.010 ± 0.006	0.111 ± 0.010
<i>RF</i>	0.965 ± 0.004	0.922 ± 0.007	0.902 ± 0.008	0.067 ± 0.006	0.098 ± 0.008

Cuadro 25: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *Hospital* replicado y balanceado.

Capa	Función de activación	Unidades	Índice de Dropout
Entrada	-	-	15 %
Oculto 1	<i>RectifierDropout</i>	500	50 %
Salida	<i>Softmax</i>	2	-

Cuadro 26: Modelo de mayor precisión para conjunto *Hospital*: *Deep Learning* con los datos balanceados, hiperparámetros. *Learning rate=0.005, 2611 epochs*.

Conjunto *TCGA*

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.591 ± 0.019	0.706 ± 0.076	0.639 ± 0.024	0.147 ± 0.032	0.361 ± 0.024
<i>XGBoost</i>	0.631 ± 0.029	0.695 ± 0.101	0.634 ± 0.038	0.091 ± 0.012	0.366 ± 0.037
<i>GBM</i>	0.647 ± 0.046	0.826 ± 0.038	0.639 ± 0.035	0.109 ± 0.017	0.361 ± 0.035
<i>RF</i>	0.642 ± 0.046	0.806 ± 0.081	0.629 ± 0.036	0.090 ± 0.014	0.371 ± 0.036

Cuadro 27: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *TCGA* con variables estándar y sin replicar.

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.961 ± 0.009	0.925 ± 0.004	0.883 ± 0.010	0.064 ± 0.005	0.117 ± 0.010
<i>XGBoost</i>	0.962 ± 0.005	0.929 ± 0.004	0.897 ± 0.009	0.055 ± 0.003	0.102 ± 0.009
<i>GBM</i>	0.961 ± 0.005	0.929 ± 0.006	0.892 ± 0.012	0.056 ± 0.002	0.108 ± 0.012
<i>RF</i>	0.959 ± 0.005	0.926 ± 0.004	0.882 ± 0.003	0.061 ± 0.002	0.118 ± 0.003
<i>GLM</i>	0.933 ± 0.007	0.890 ± 0.018	0.859 ± 0.005	0.109 ± 0.006	0.143 ± 0.013

Cuadro 28: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *TCGA* replicado y sin balancear.

Algoritmo	AUC	ACC	ACC medio por clase	MSE	MSE medio por clase
<i>DL</i>	0.961 ± 0.009	0.925 ± 0.004	0.882 ± 0.010	0.064 ± 0.005	0.117 ± 0.010
<i>XGBoost</i>	0.961 ± 0.005	0.929 ± 0.004	0.897 ± 0.009	0.056 ± 0.003	0.108 ± 0.012
<i>GBM</i>	0.962 ± 0.005	0.929 ± 0.006	0.892 ± 0.012	0.056 ± 0.002	0.108 ± 0.012
<i>RF</i>	0.959 ± 0.005	0.926 ± 0.004	0.882 ± 0.003	0.061 ± 0.002	0.118 ± 0.003
<i>GLM</i>	0.933 ± 0.007	0.890 ± 0.010	0.859 ± 0.005	0.019 ± 0.006	0.143 ± 0.013

Cuadro 29: Resultados con desviación estándar de predictores *Clasificación* usando *h2o* en *TCGA* replicado y balanceado.

Número de árboles	Profundidad máxima	Learning rate
153	15	0.05

Cuadro 30: Modelo de mayor precisión para conjunto *TCGA*: *Extreme Gradient Boosting* con los datos sin balancear, hiperparámetros. 160 epochs.

7. Conclusión

Nuestro primer objetivo consistía en analizar la utilidad y rendimiento de la hormonoterapia y quimioterapia como tratamientos en nuestras pacientes con terapia adyuvante y receptores hormonales positivos. Hemos visto que la hormonoterapia (especialmente la que contiene inhibidores de aromatasa) tiene una buena aceptación pudiendo mejorar la supervivencia de forma notable. En cambio, la quimioterapia, lejos de ser efectiva y reducir el riesgo de recaer en la enfermedad, lo aumenta en aquella pacientes que son expuestas a este tratamiento. Además, el riesgo de recaer ha resultado ser mayor a mayor gravedad de la paciente.

El segundo objetivo consistía en descubrir si, para las variables numéricas de nuestros conjuntos, las estratificaciones creadas por aprendizaje automático no supervisado que se adaptan a los datos explican mejor el riesgo de muerte y recurrencia que las usadas de forma estándar en la práctica clínica. Habiendo analizado para cada caso la mejor opción de clusterización, y analizado estas con los métodos estadísticos semi-paramétrico de Cox y no-paramétrico Kaplan-Meier, hemos obtenido los resultados esperados. En el caso del conjunto *TCGA* con solo dos variables numéricas, ambas han dado su mayor significancia en las estratificaciones definidas por el algoritmo *Kmeans*. En el caso del conjunto *Hospital*, las estratificaciones más significativas se dividen entre las usadas en el paper y las obtenidas por los clusters. Para ambos conjuntos se cumple que las estratificaciones estándar en ningún caso son las que mejor explican el riesgo.

Una vez definidos los conjuntos con las variables elegidas, queríamos encontrar un predictor de supervivencia preciso usando algoritmos de aprendizaje automático y comprobar su mejora respecto a los métodos estadísticos tradicionalmente usados.

Habíamos estimado que obtendríamos mejores resultados con métodos de aprendizaje computacional que con métodos estadísticos. También habíamos estimado que nuestro tratamiento de datos mejoraría respecto a los resultados normalmente obtenidos en este tipo de problemas.

En el análisis exploratorio realizado con `mlr` hemos visto que, para los métodos estadísticos, el *C – Index* máximo alcanzado es 0,58 en el conjunto *Hospital* y 0,7 en el conjunto *TCGA*, sin embargo, las aproximaciones de aprendizaje computacional encuentran su valor mínimo en una precisión de 0,83 en el conjunto *Hospital* y 0,82 en el conjunto *TCGA*. Respecto a los predictores estadísticos encontrados en la literatura podemos ver en algu-

nos ejemplos que nuestros resultados no se alejan del paradigma real en casos de cáncer de mama. [54] obtiene un $C - index$ máximo de 0,629 en predecir el riesgo de recurrencia. También con un método estadístico, [55] obtiene un AUC máximo de 0,740, o [56] y [57], que obtienen AUC s máximos de 0,791 y 0,714 respectivamente. En cuanto los resultados con aproximaciones de aprendizaje computacional en cáncer de mama, nuestros resultados de validación para el conjunto *Hospital* encuentran un AUC máximo de 0,974 (con un error estándar 0.003) con DL para el conjunto *Hospital* y 0,962 (con un error estándar de 0.005) con XGBoost para el conjunto *TCGA*. En cuanto a los ejemplos encontrados en la literatura de los últimos años con esta aproximación encontramos un AUC máximo de 0,86 con predictores ML-RO [58], 0,930 con RF [60] o 0,936 con DL y RF en un estudio que usaba más de 200.000 casos y aseguraba ser el mejor resultado hasta la fecha en 2005 [59].

No solo podemos afirmar que, efectivamente, las aproximaciones de aprendizaje computacional superan de forma muy significativa la precisión que puede obtenerse con aproximaciones estadísticas en predicción de riesgo en pacientes de cáncer de mama, si no que, además, nuestros resultados obtenidos a partir de datos replicados con estratificaciones ajustadas con aprendizaje automático parecen superar a aquellos estudios en los que no se realiza este paso.

Dentro de nuestros propios resultados, podemos ver como, en ambos conjuntos, los resultados de precisión obtenidos a partir del mismo conjunto original, habiendo replicado y elegido las estratificaciones difieren de forma evidente a los obtenidos en la versión estándar y sin replicar. La precisión en algoritmos de aprendizaje automático aumenta entre un 20 % y 30 % gracias al tratamiento realizado. En el conjunto *Hospital* el algoritmo que mejor funciona es DL, probablemente por contar con más datos y poder sacar provecho de la complejidad de este (lo comprobamos en el conjunto sin replicar, donde DL cuenta con los peores resultados entre los algoritmos analizados). En el conjunto *TCGA*, con menos datos y resultados ligeramente inferiores, el mejor funcionamiento lo obtenemos con el algoritmo XGBoost con poca diferencia respecto a DL. También es destacable que los valores de precisión por clase son inferiores a los globales debido a la gran diferencia de frecuencia en los niveles de la variable a predecir. Debido a esta diferencia, hemos

ejecutado el mismo análisis balanceando los datos. En el conjunto *Hospital* podemos ver una mejora en los valores de precisión aunque sus rangos de error estándar por lo general los peores. En el conjunto *TCGA* las diferencias causadas por el balanceo de datos son casi imperceptibles.

Respecto al cómputo, no hemos encontrado ningún impedimento gracias a la paralización ofrecida por `h2o`. De no contar con ella, es probable que no hubiéramos llegado a resultados tan precisos, o no con el mismo esfuerzo.

7.1. Trabajos Futuros

La principal tarea a mejorar es el funcionamiento de la interfaz realizada a la que no se le ha podido dedicar demasiado tiempo. En la versión actual se ejecuta una predicción completa con el consiguiente tiempo de ejecución y espera cada vez que se cambia un parámetro. Para poder disfrutar de un uso práctico y efectivo, es conveniente mejorar el código y que solamente se realice una predicción cuando la combinación de parámetros final esté seleccionada. Otras mejoras a realizar en la interfaz serían poder almacenar las curvas previas para poder comparar como afectan a la supervivencia los distintos parámetros, y añadir la visualización de una barra de progreso que indique el estado de la predicción al usuario.

Respecto a los algoritmos de aprendizaje computacional, queda un amplio rango de modelos e hiperparámetros por explorar, esperando que los resultados pudieran mejorarse aún más.

Habiendo visto que la individualización de riesgo tiene un gran efecto en estos casos, y el buen resultado obtenido gracias a la clusterización a nivel de variable, para mayor precisión, se podría dividir el conjunto final en clusters y diseñar un modelo de predicción específico para cada uno de ellos.

Finalmente, cabe destacar que todos los objetivos del proyecto se han alcanzado exitosamente. De forma personal, he disfrutado enormemente su ejecución por tener la libertad de investigar y aplicar muchos de los conceptos aprendidos en los últimos años.

8. Recursos

8.1. Software

Todo el software implementado en este proyecto se ha realizado en R, usando *R Studio*. Los paquetes usados de forma principal se han mencionado a lo largo de la metodología (*survival*, *tcga*, *mlr*, *h2o*,...).

El software desarrollado está subido en el repositorio público https://github.com/nairachiclana/ML_and_Survival-BreastCancer en forma de archivos *.rmd* exceptuando el código correspondiente a la interfaz. Los datos del conjunto *TCGA* se extraen en el código, los datos del conjunto *Hospital* no pueden ser compartidos por protección de datos.

8.2. Hardware

He usado mi máquina personal, software *macOs Mojave* versión 10.14.5 con 8GB RAM, 1TB de disco duro y un procesador Intel Core i5 3,4GHz.

Dado que los algoritmos de alto cómputo han sido entrenados usando la paralelización ofrecida por *h2o* como se ha mencionado anteriormente, el equipo no ha supuesto ninguna limitación.

Anexo-Curvas de supervivencia

En todas las Figuras se presenta en la parte superior la curva de supervivencia del conjunto de datos sin replicar. En la parte inferior a la izquierda la del conjunto replicado y a la derecha la de eventos predichos a partir del conjunto replicado y el modelo predictor elegido para cada conjunto.

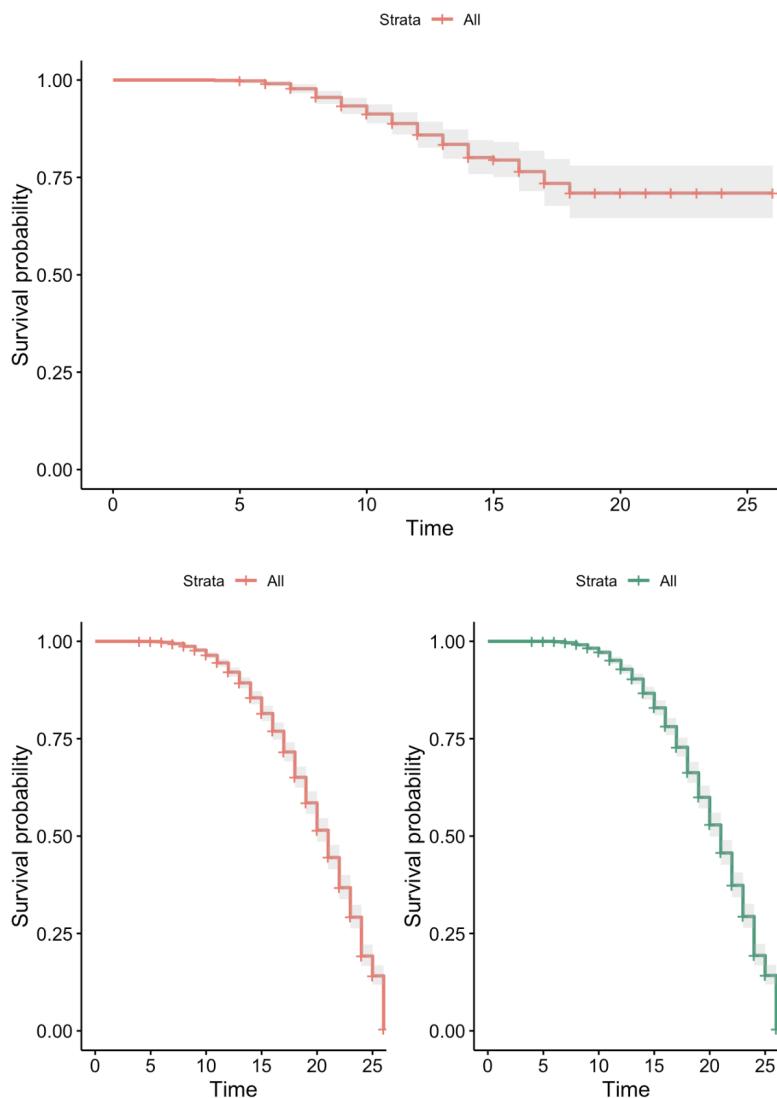


Figura 34: Curvas de supervivencia conjunto *Hospital* toda la población

Conjunto	n	eventos	median	95 % CI
Original	826	88	NA	NA-NA
Replicado	3422	968	21	20-21
Predicho	3422	939	21	20-21

Cuadro 31: Valores curvas de supervivencia conjuntos *Hospital*.

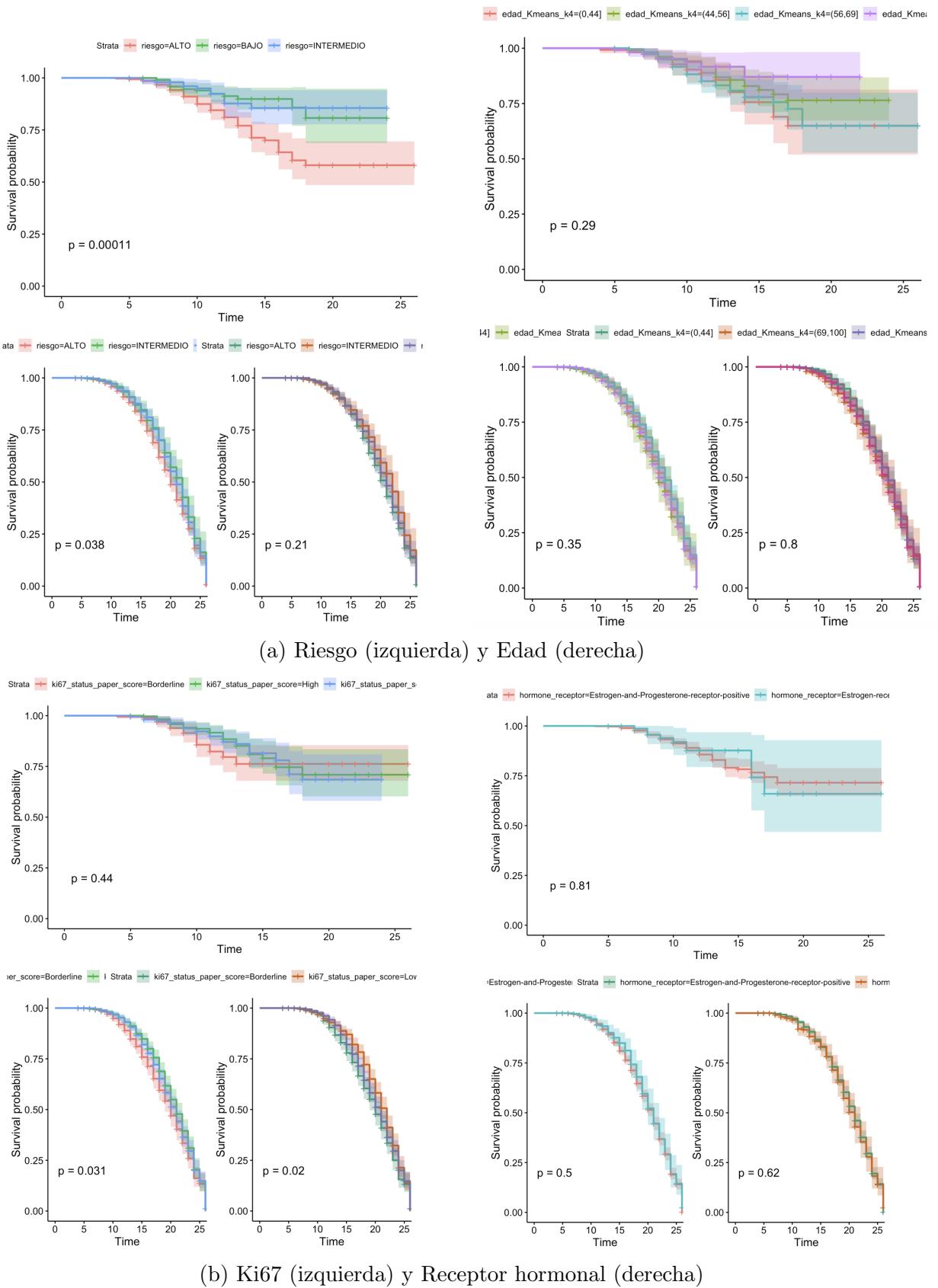
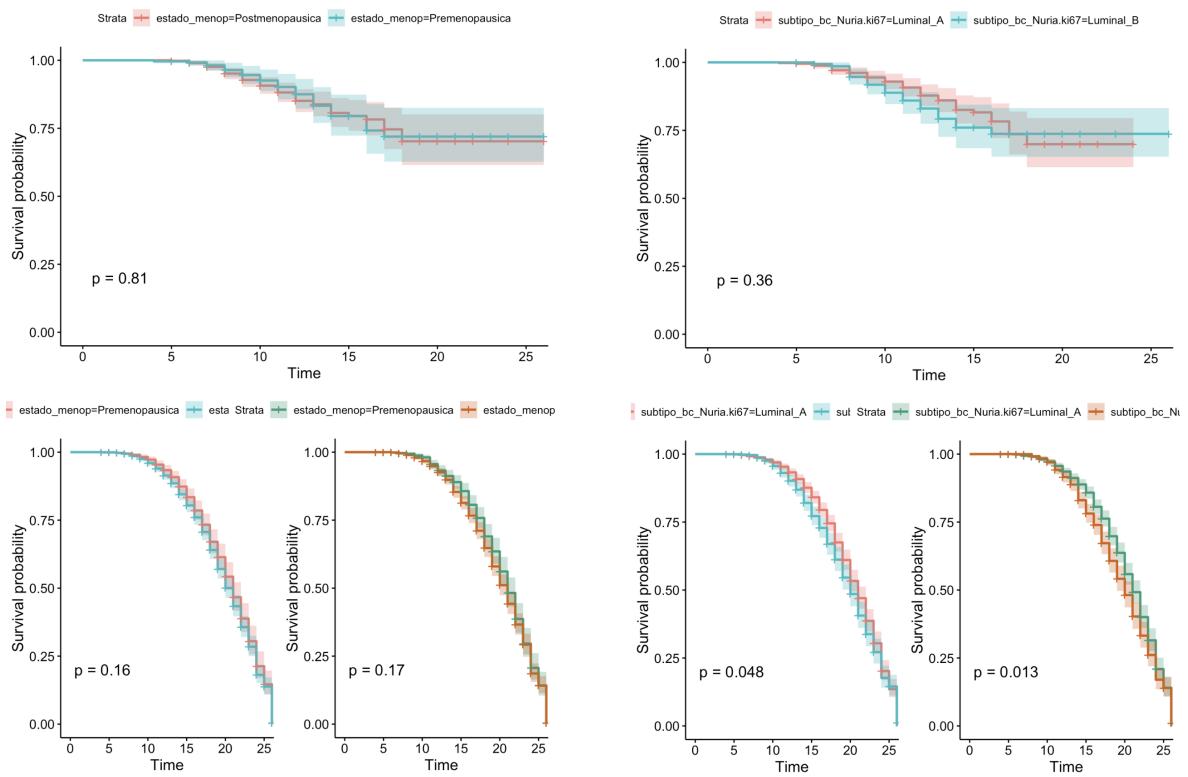
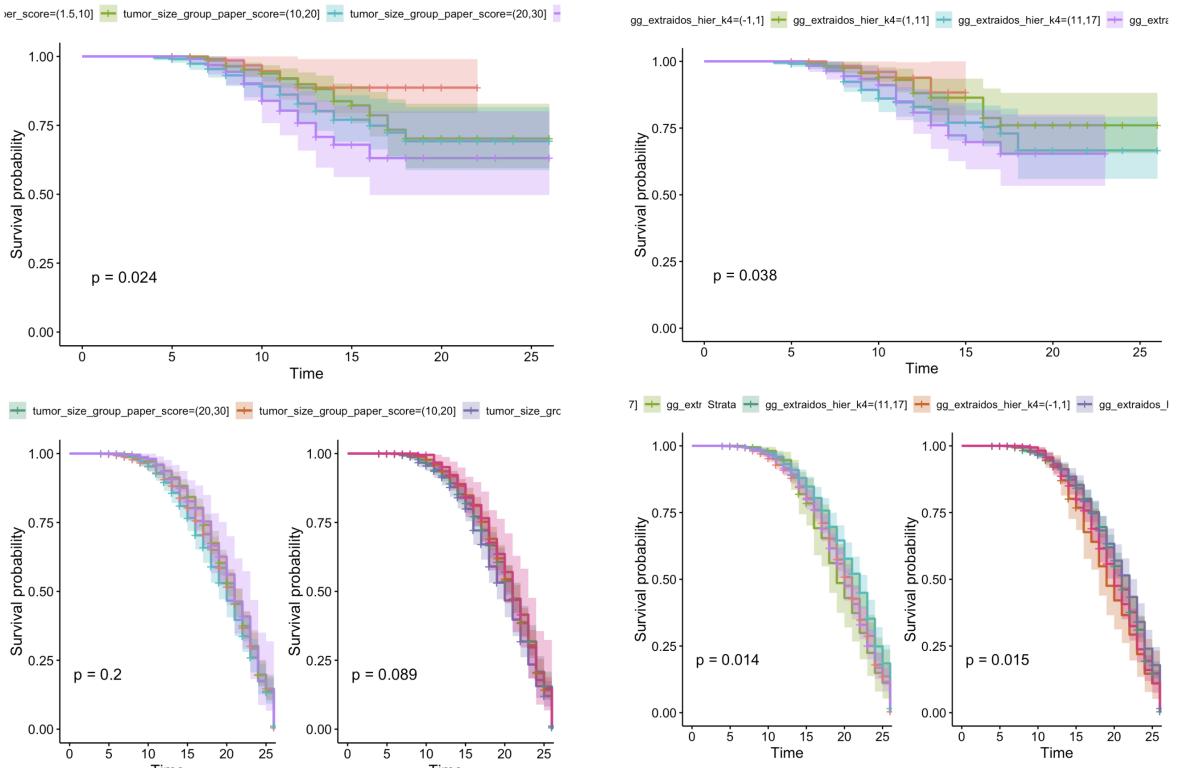


Figura 35: Curvas de supervivencia conjunto *Hospital (1)*.

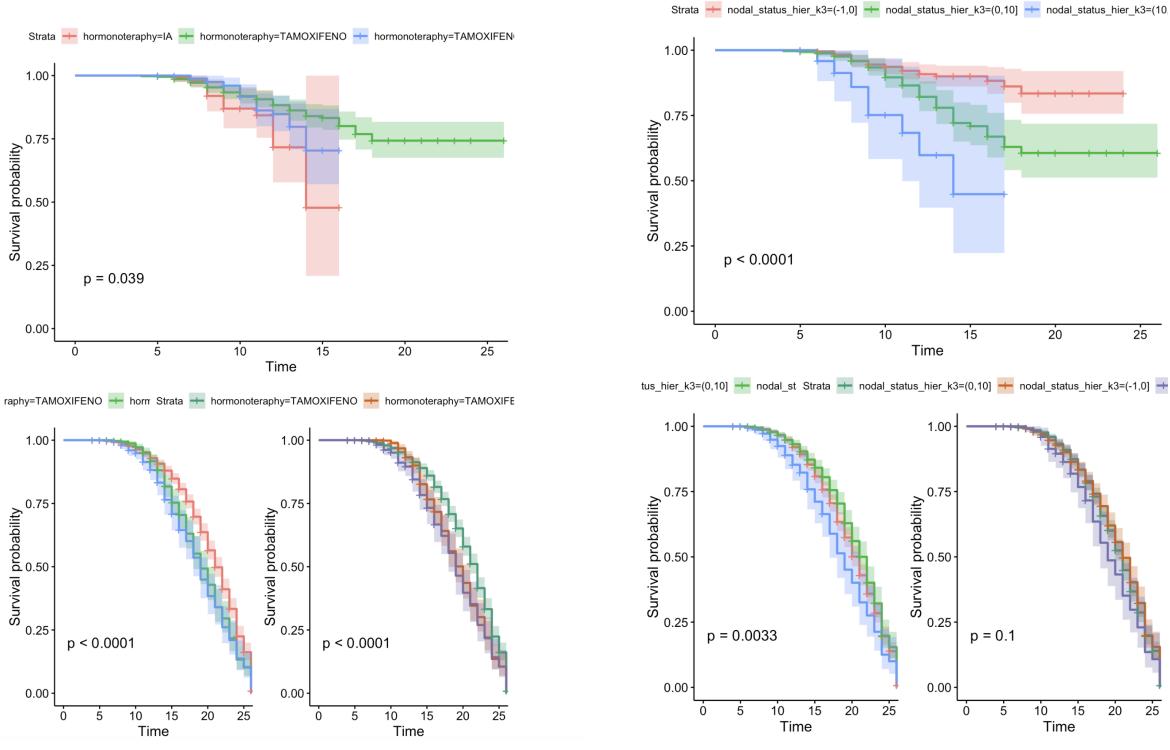


(a) Menopausia (izquierda) y Subtipo de cáncer (derecha)



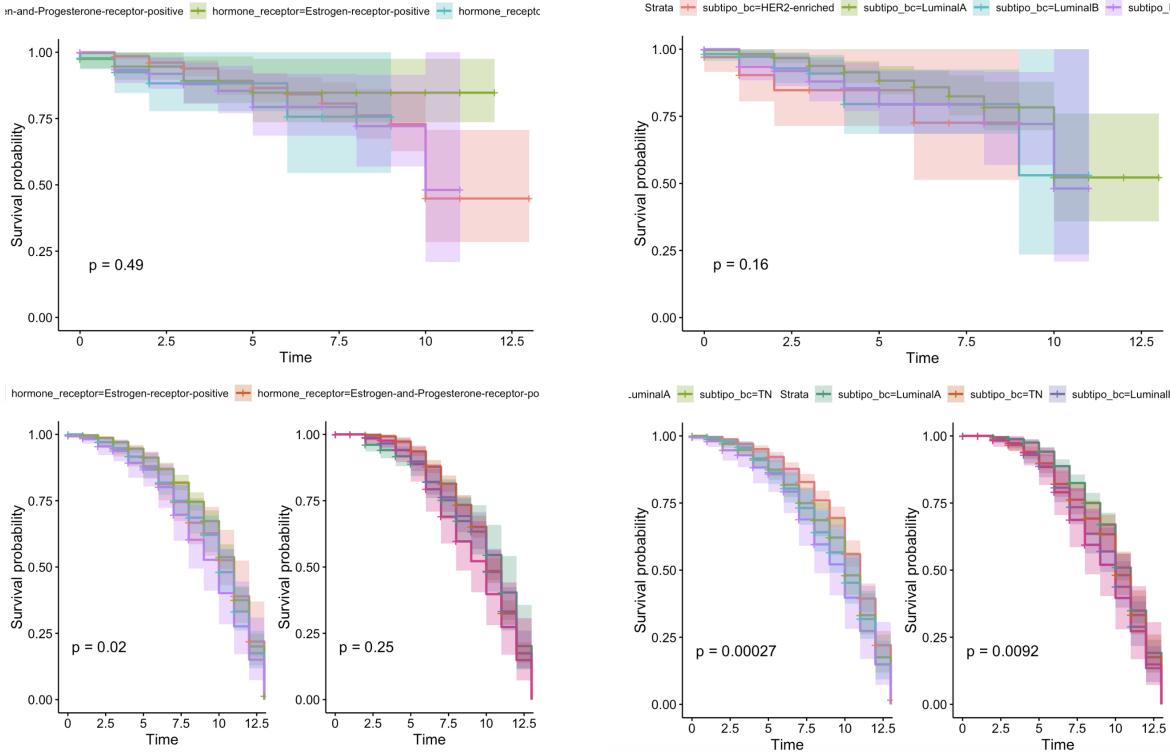
(b) Tamaño tumoral (izquierda) y Ganglios extraídos (derecha)

Figura 36: Curvas de supervivencia conjunto Hospital (2).



(a) Hormonoterapia (izquierda) y Estado nodal (derecha)

Figura 37: Curvas de supervivencia conjunto *Hospital (3)*.



(a) Menopausia (izquierda) y Subtipo de cáncer (derecha)

Figura 38: Curvas de supervivencia conjunto *TCGA (2)*.

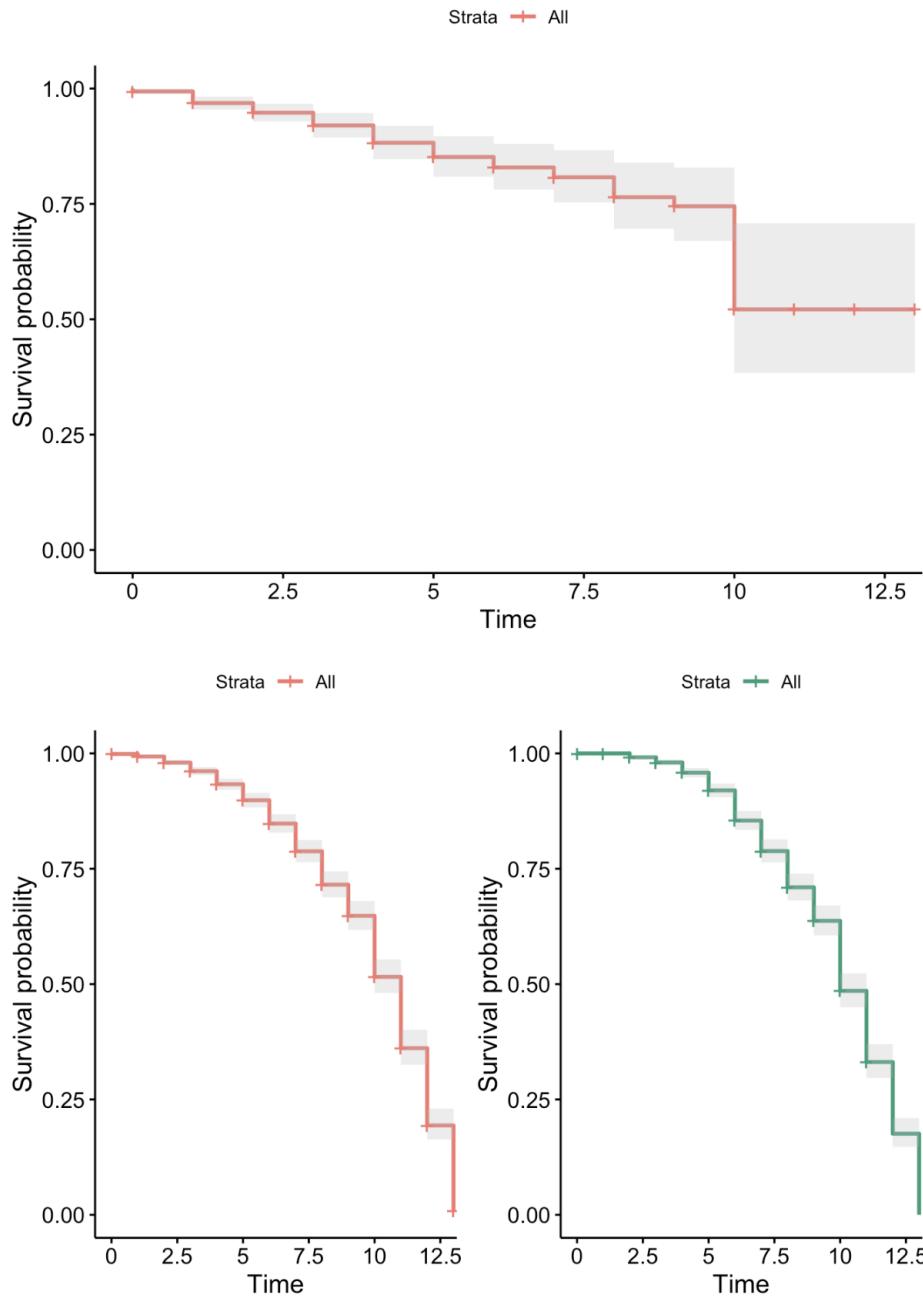
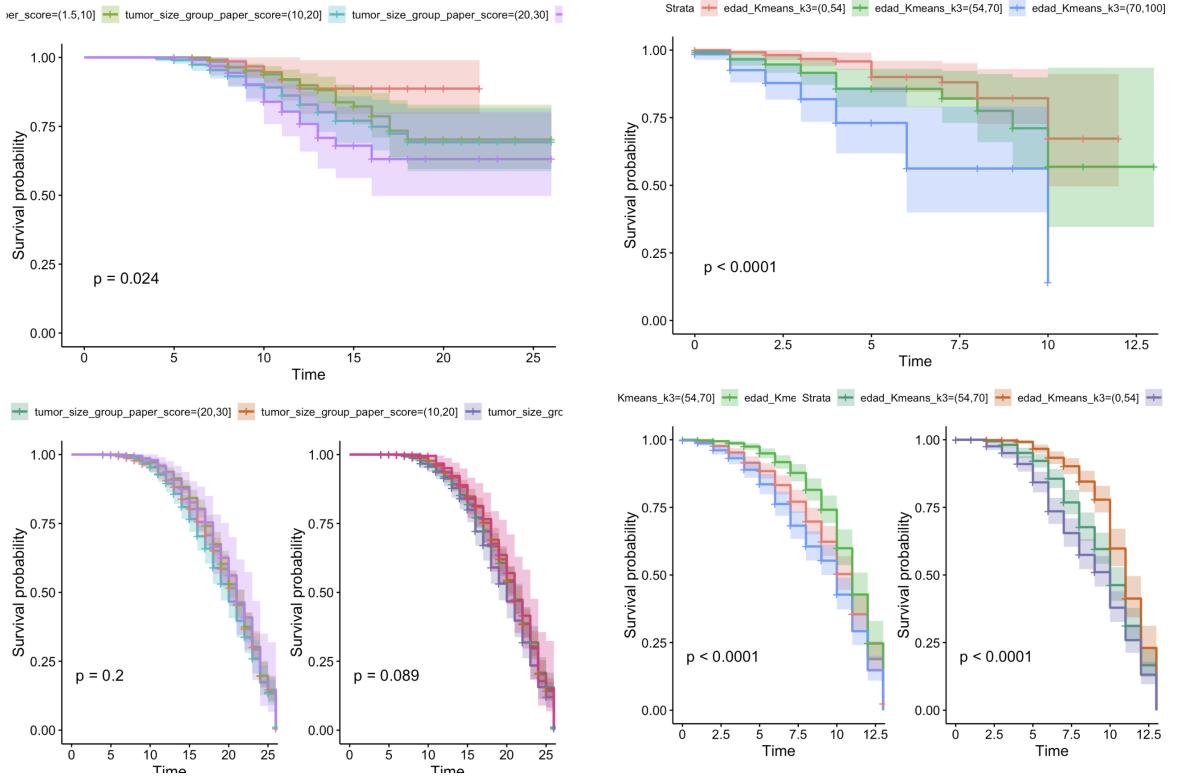


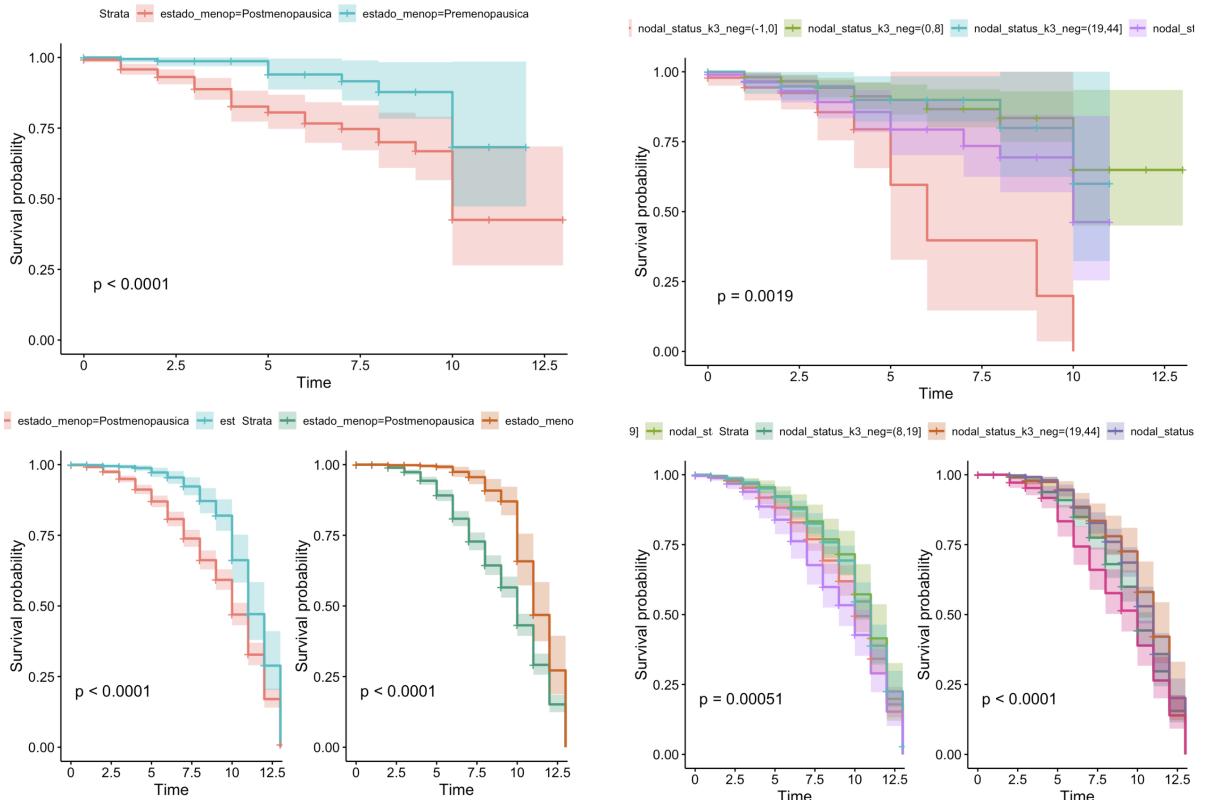
Figura 39: Curvas de supervivencia conjunto *TCGA* toda la población

Conjunto	n	eventos	median	95 % CI
Original	648	64	NA	10-NA
Replicado	3137	704	11	10-11
Predicho	3137	708	10	10-11

Cuadro 32: Valores curvas de supervivencia conjuntos *TCGA*.

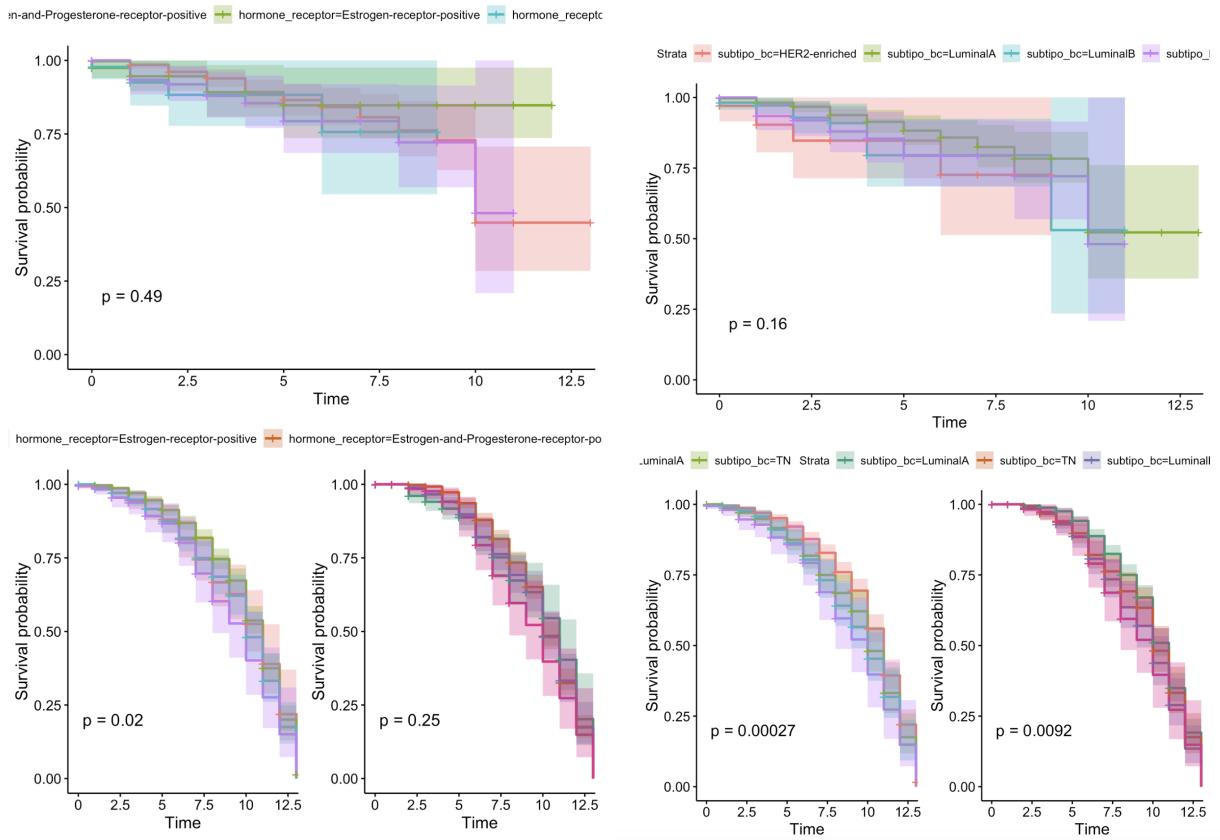


(a) Tamaño tumoral (izquierda) y Edad (derecha)

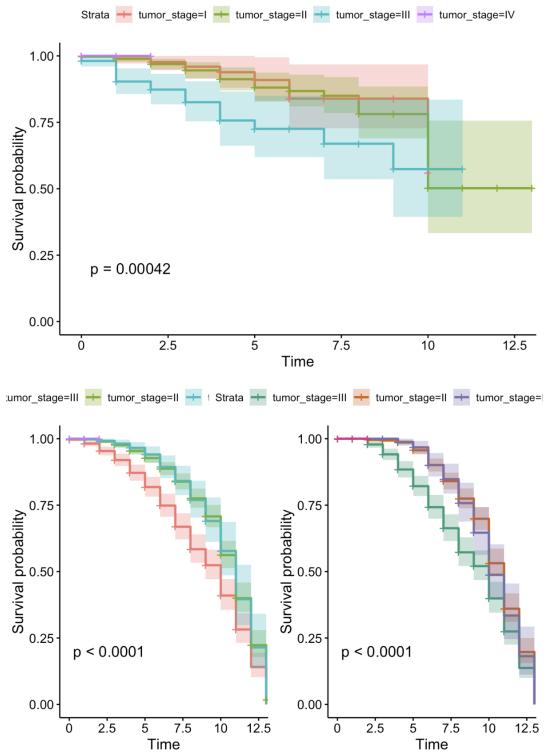


(b) Estado menopausico (izquierda) y Estado nodal (derecha)

Figura 40: Curvas de supervivencia conjunto *TCGA* (1).



(a) Menopausia (izquierda) y Subtipo de cáncer (derecha)



(b) Etapa tumoral

Figura 41: Curvas de supervivencia conjunto TCGA (2).

Referencias

- [1] National Cancer Institute *Cancer Statistics* <https://www.cancer.gov/about-cancer/understanding/statistics>
- [2] Cancer Research UK *Health economics: the cancer drugs cost conundrum* <https://www.cancerresearchuk.org/funding-for-researchers/research-features/2016-08-10-health-economics-the-cancer-drugs-cost-conundrum>
- [3] National Cancer Institute: *Cancer Statistics* <https://www.cancer.gov/about-cancer/understanding/statistics>
- [4] *Our World in Data: Cancer* <https://ourworldindata.org/cancer>
- [5] Science Direct *¿Sabemos qué causa el cáncer de mama?* <https://www.sciencedirect.com/science/article/pii/S0304501309726287?via%3Dihub>
- [6] The healthy Breast Program *Global Breast Cancer incidence 2018* <http://mammalive.net/research/global-breast-cancer-incidence-2018/>
- [7] KPMG *Investment in AI for healthcare soars* <https://home.kpmg/xx/en/home/insights/2018/11/investment-in-ai-for-healthcare-soars.html>
- [8] MaRS *Computers are already better than doctors at diagnosing some diseases* <https://www.marsdd.com/magazine/computers-are-already-better-than-doctors-at-diagnosing-some-diseases/>
- [9] nature *Skin cancer classification with deep learning.* <https://cs.stanford.edu/people/esteva/nature/>
- [10] Medical News Today *Artificial intelligence better than humans at spotting lung cancer* <https://www.medicalnewstoday.com/articles/325223.php>
- [11] TIME *Machines Treating Patients? It's Already Happening* <https://time.com/5556339/artificial-intelligence-robots-medicine/>
- [12] British Journal of General Practice *Artificial Intelligence in medicine: current trends and future possibilities.* <https://bjgp.org/content/68/668/143>
- [13] Exscientia <https://www.exscientia.co.uk/>

- [14] Science *Adversarial attacks on medical machine learning* <https://science.sciencemag.org/content/363/6433/1287.summary>
- [15] <https://www.gradiant.org/blog/analisis-supervivencia-industria-4-0/>
- [16] PubMed *Survival analysis in clinical trials: Basics and must know areas* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332/>
- [17] NIH *The Breast Cancer Risk Assessment Tool* <https://bcrisktool.cancer.gov/>
- [18] Bondy ML, Lustbader ED, Halabi S, et al. *Validation of a breast cancer risk assessment model in women with a positive family history.* <https://www.ncbi.nlm.nih.gov/pubmed/8003106?dopt=Abstract>
- [19] Spiegelman D, Colditz GA, Hunter D, Hertzmark E. *Validation of the Gail et al. model for predicting individual breast cancer risk.* J Natl Cancer Inst. <https://www.ncbi.nlm.nih.gov/pubmed/8145275?dopt=Abstract>
- [20] Rockhill B, Spiegelman D, Byrne C, et al. *Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention.* J Natl Cancer Inst. <https://www.ncbi.nlm.nih.gov/pubmed/11238697?dopt=Abstract>
- [21] Costantino JP, Gail MH, Pee D, et al. *Validation studies for models projecting the risk of invasive and total breast cancer incidence.* J Natl Cancer Inst. <https://www.ncbi.nlm.nih.gov/pubmed/10491430?dopt=Abstract>
- [22] Bernatsky S, Clarke A, Ramsey-Goldman R, et al. *Hormonal exposures and breast cancer in a sample of women with systemic lupus erythematosus.* Rheumatology. <https://www.ncbi.nlm.nih.gov/pubmed/15226516?dopt=Abstract>
- [23] Olson JE, Sellers TA, Iturria SJ, Hartmann LC. *Bilateral oophorectomy and breast cancer risk reduction among women with a family history.* Cancer Detect Prev. 2004;28:357–60. <https://www.ncbi.nlm.nih.gov/pubmed/15542261?dopt=Abstract>
- [24] Tice JA, Cummings SR, Smith-Bindman R, et al. *Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation*

- of a new predictive model.* Ann Intern Med. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674327/>
- [25] Schonfeld SJ, Pee D, Greenlee RT, et al. *Effect of changing breast cancer incidence rates on the calibration of the Gail model.* J Clin Oncol. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2881722/>
- [26] Tice JA, Cummings SR, Ziv E, Kerlikowske K., et al. *Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population.* <https://www.ncbi.nlm.nih.gov/pubmed/16261410?dopt=Abstract>
- [27] STUART J. SCHNITT. MD, JAMES L., *Pathologic Predictors of Early Local Recurrence in Stage I and I1 Breast Cancer Treated by Primary Radiation Therapy* <https://onlinelibrary.wiley.com/doi/epdf/10.1002/1097-0142%2819840301%2953%3A5%3C1049%3A%3AAID-CNCR2820530506%3E3.0.CO%3B2-0>
- [28] Martin Filipits, Margaretha Rudas, Raimund Jakesz., *A New Molecular Predictor of Distant Recurrence in ER-Positive, HER2-Negative Breast Cancer Adds Independent Information to Conventional Clinical Risk Factors* <https://clincancerres.aacrjournals.org/content/17/18/6012.short>
- [29] John Boyages, Abram Recht, James L. Connolly., *Early breast cancer: predictors of breast recurrence for patients treated with conservative surgery and radiation therapy* <https://www.sciencedirect.com/science/article/pii/016781409090163Q>
- [30] Ivana Sestak, Mitch Dowsett, Lila Zabaglo., *Factors Predicting Late Recurrence for Estrogen Receptor-Positive Breast Cancer* <https://www.sciencedirect.com/science/article/pii/016781409090163Q>
- [31] Ahmad Lg, Abbas Toloie Eshlaghy., *Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence* <https://www.sciencedirect.com/science/article/pii/016781409090163Q>
- [32] Dursun Delen, Glenn Walker, AmitKadam., *Predicting breast cancer survivability: a comparison of three data mining methods* <https://www.sciencedirect.com/science/article/pii/S0933365704001010>

- [33] *Vinayak Bhandari, Paul C.Boutros, Comparing continuous and discrete analyses of breast cancer survival information* <https://www.sciencedirect.com/science/article/pii/S0888754316300684>
- [34] *Fisterra: Análisis de supervivencia* <https://www.fisterra.com/mbe/investiga/supervivencia/supervivencia.asp>
- [35] *Sebastian Pölsterl, About Survival Analysis* <https://www.fisterra.com/mbe/investiga/supervivencia/supervivencia.asp>
- [36] *Temario asignatura Aprendizaje Computacional Ingeniería de la Salud (UMA)* <https://informatica.cv.uma.es/course/view.php?id=3636>
- [37] *Towards data science: Hierarchical clustering Clearly Explained* <https://towardsdatascience.com/https-towardsdatascience-com-hierarchical-clustering-6f3c98>
- [38] *Medium: Introduction to Confusion Matrix* <https://medium.com/tech-vision/introduction-to-confusion-matrix-classification-modeling-54d867169906>
- [39] *Facundo Bre, Juan M. Gimenez, Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks* https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051
- [40] *Isaac Changhau, Activation Functions in Neural Networks* https://isaacchanghau.github.io/post/activation_functions/
- [41] *My concepts with Data Science* <http://ramankulshrestha.blogspot.com/2019/>
- [42] *UC Business Analytics R Programming Guide: Gradient Boosting Machines* http://uc-r.github.io/gbm_regression
- [43] *Kan Nishida, Introduction to Extreme Gradient Boosting in Exploratory* <https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7>
- [44] *Resampling mlr* <https://mlr.mlr-org.com/articles/tutorial/resample.html>

- [45] *Machine Learning model selection* http://ethen8181.github.io/machine-learning/model_selection/model_selection.html
- [46] *The Cancer Genome Atlas Program* <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>
- [47] *Molecular subtypes of breast cancer* <https://ww5.komen.org/BreastCancer/SubtypesofBreastCancer.html>
- [48] *Mitch Dowsett, Ivana Sestak, Meredith M. Regan, Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients With Estrogen Receptor–Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5.*
- [49] *Package Survival, R* <https://cran.r-project.org/web/packages/survival/survival.pdf>
- [50] *Package TCGAretriever, R* <https://cran.r-project.org/web/packages/TCGAretriever/TCGAretriever.pdf>
- [51] *Package mlr, R* <https://mlr.mlr-org.com/>
- [52] *Package h2o, R* <https://cran.r-project.org/web/packages/h2o/h2o.pdf>
- [53] *Shiny, R* <https://shiny.rstudio.com>
- [54] *Pankratz VS, Degnim AC, Frank RD, Model for individualized prediction of breast cancer risk after a benign breast biopsy. (2015)* <https://www.ncbi.nlm.nih.gov/pubmed/25624442?dopt=Abstract>
- [55] *Tuomo J. Meretoja, Kenneth Geving Andersen, Clinical Prediction Model and Tool for Assessing Risk of Persistent Pain After Breast Cancer Surgery. (2017)* <https://ascopubs.org/doi/10.1200/JCO.2016.70.3413>
- [56] *Francisco J. Candido dos Reis, Gordon C. Wishart, Ed M. Dicks, An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. (2017)* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440946/>

- [57] Megan S Rice, Shelley S Tworoger, Susan E Hankinson, *Breast cancer risk prediction: An update to the Rosner-Colditz breast cancer incidence model.* (2017) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647223/>
- [58] Ferroni P, Zanzotto FM, Riondino S, *Breast Cancer Prognosis Using a Machine Learning Approach.* (2019) <https://www.ncbi.nlm.nih.gov/pubmed/30866535>
- [59] Delen D, Walker G, Kadam A, *Predicting breast cancer survivability: a comparison of three data mining methods.* (2005) <https://www.ncbi.nlm.nih.gov/pubmed/15894176>
- [60] Montazeri M, Beigzadeh A., *Machine learning models in breast cancer survival prediction.* (2016). <https://www.ncbi.nlm.nih.gov/pubmed/26409558>