# Modelos estadísticos para el estudio longitudinal y transversal de eventos no-SIDA

Grado en Estadística Empresarial
Universidad Miguel Hernández
Abril 2020, Elche




Marco Luis Lievers Ruiz

# Índice

# 1 Introducción

## 1.1 VIH: definición e historia.

El Virus de la Inmunodeficiencia Humana (VIH) es un virus que ataca el sistema inmunitario, el cual se encarga de defender el cuerpo contra microorganismos infecciosos. Especialmente ataca un tipo de células del sistema llamadas CD4, las cuales se encargan de la creación de anticuerpos, por tanto, si esta infección no es tratada, se puede llegar al extremo de perder la capacidad del cuerpo de defenderse contra infecciones y otro tipo de enfermedades. Existe una serie de enfermedades, denominadas enfermedades oportunistas, las cuales se desarrollan cuando el sistema inmunitario está deteriorado, y principalmente afectan al cerebro, hígado, piel, ojos y pulmones, entre otros órganos. Las formas más comunes de transmisión del VIH son las relaciones sexuales sin la debida protección, así como compartir agujas con una persona infectada, debido al contacto con ciertos líquidos corporales. Las madres infectadas por el virus también pueden infectar a sus hijos durante el embarazo, el parto o la lactancia, hoy en día existen medicamentos para reducir considerablemente el riesgo de contagio madre-hijo. El virus en si no tiene sintomas, lo que dificulta su detección. La única forma de saber si estás infectado es mediante una prueba, la cual está disponible de forma gratuita en todos los hospitales y centros de salud públicos.

El año 1981, en California, se detectaron los primeros casos de esta infección entre jóvenes homosexuales y más tarde ese mismo año, en consumidores de drogas inyectables, pero no fue hasta dos años más tarde, en 1983, cuando se descubrió el virus y se le puso nombre, se encontró también la relación entre el VIH y el SIDA. En 1987 se presentó el primer antirretroviral para luchar contra la infección, el cual tenia un elevado precio y fuertes efectos secundarios, pero no fue hasta 1996 cuando presentaron nuevos medicamentos más económicos, potentes y con menos efectos secundarios. En 2002 se aprobaron y comercializaron los primeros tests rápidos, con una exactitud del 99,6% en 20 minutos y en 2014 fue planteada la estrategia 90-90-90 por el Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA (ONUSIDA), la cual consistía en, para 2020, conseguir que el 90% de las personas infectada fuesen detectadas, el 90% de estas, en tratamiento, y el 90% de pacientes en tratamiento, tuviesen carga viral indetectable. En 2012, se consiguió el primer paciente ("El paciente Berlín") que perdió por completo la presencia del virus en la sangre, y no fue hasta 2019 cuando se consiguió el segundo paciente ("El paciente Londres") curado, el cual no presentaba restos del VIH en su organismo tras 19 meses.

La etapa final de la infección es conocida como Síndrome de la Inmunodeficiencia Adquirida (SIDA), y se produce cuando el organismo pierde por completo, o en gran parte, la capacidad de protegerse. Únicamente en 2018, 1.7 millones de personas al rededor de todo el mundo fueron infectadas con VIH, lo que hace un total de 74.9 millones de personas infectadas desde el inicio de la epidemia. En total, han muerto 32 millones de personas por enfermedades relacionadas con el SIDA, 770.000 solo en el año 2018, lo que es un descenso del 56% respecto a 2004, cuando en un año, murieron un total de 1.7 millones de personas por enfermedades relacionadas con el SIDA.

En algunos países subdesarrollados, la gravedad del virus es mucho mayor por la falta de educación sexual, medidas de prevención y recursos sanitarios disponibles, como tests, medicación, etc. A continuación (Fig. 1) observamos un mapa que representa el porcentaje de personas infectadas con VIH que reciben tratamiento contra el VIH al rededor del mundo. Se observa que en centro Europa este porcentaje es bastante alto, con la mayoría de los países por encima del 75%, mientras que, en otros lugares, como Rusia, la cantidad de infectados que reciben tratamiento está por debajo del 50%, y en algunas zonas de África, por debajo del 25%. En Sudáfrica, por ejemplo, en 2017, se concentraron el 28,46% de las muertes por VIH/SIDA a nivel mundial, y en Rusia, se produjeron al rededor de 139.000 nuevos contagios por VIH.

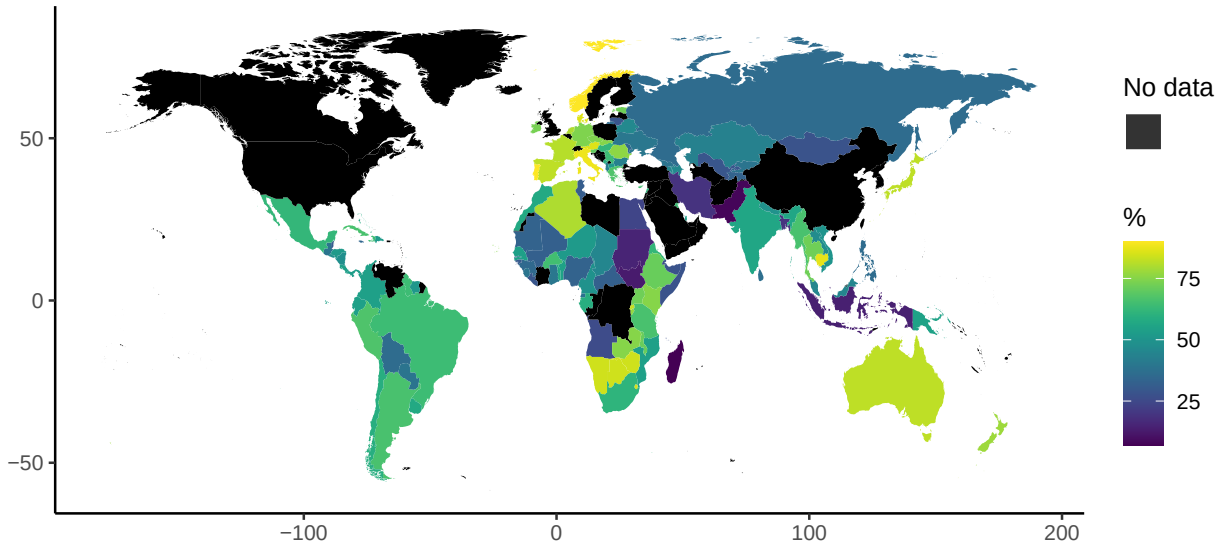


Figure 1: Mapa de porcentaje de infectados con VIH que reciben tratamiento.

Actualmente no existe una cura para dicho virus, pero, en los últimos años, con la introducción de las inmunoterapias CART, se ha observado en varios estudios que ha aumentado considerablemente la supervivencia de los pacientes infectados con VIH, así como ha disminuido la probabilidad de transmisión de este. En junio de 2019, 24.5 millones de personas al rededor del mundo estaban siendo tratadas con inmunoterapias, crecimiento considerable en comparación con 2010, cuando solo 7.7 millones de personas estaban siendo tratadas con inmunoterapias. Gracias al gran avance en cuanto a técnicas de detección temprana del virus, los fármacos antirretrovirales (ARVs) se inician en los estados iniciales de la infección, lo que es crucial para evitar que el paciente infectado acabe desarrollando SIDA. Como se puede observar a continuación (Fig. 2), en los últimos años han caído considerablemente las muertes por VIH, mientras que ha aumentado la cantidad de gente que vive una vida normal con el virus.

Los medicamentos antirretrovirales (ARVs) que son usados para combatir el VIH, reducen la carga viral en el organismo. Estos medicamentos son parte de las terapias CART, el conjunto de medicamentos ARVs que toma el paciente se le denomina cocktail anti-VIH. Cuando se combinan dichos medicamentos, el ritmo con el que el virus se reproduce en el organismo se desacelera, lo que permite al sistema inmunitario mantenerse sano. En algunas personas, este cockatail anti-VIH puede tener efectos secundarios y causarle dolor abdominal, vómitos o cansancio, y es muy importante seguir las indicaciones en cuanto a la ingesta de los medicamentos, ya que el VIH podría volverse resistente a los medicamentos si no se siguen.

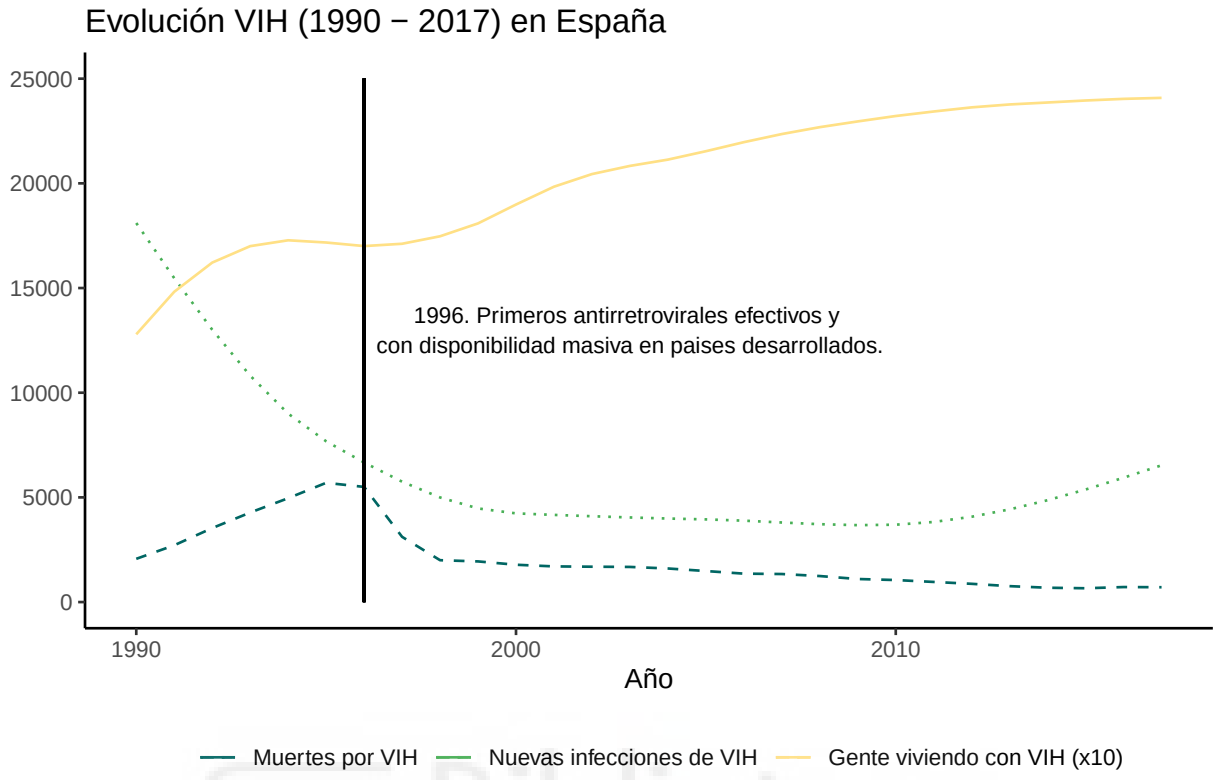


Figure 2: Evolución de VIH en España desde 1990 hasta 2017.

## 1.2 Eventos No-SIDA (ENOS).

En los últimos años han cambiado los patrones de morbilidad y mortalidad en pacientes infectados VIH, pasando de ser principalmente enfermedades relacionadas con el SIDA a los denominados eventos No SIDA (ENOS), enfermedades no relacionadas directamente con el VIH, según el Centro para el Control y la Prevención de Enfermedades (CDC). Las mas comunes son enfermedades de riñón, cardiovasculares o infecciones severas. A pesar de los grandes avances en el ámbito del VIH, a día de hoy, la gente infectada tiene, de media, una esperanza de vida menor y con mayores complicaciones que una persona no infectada. La aparición de estos eventos se suele relacionar con el aumento de la esperanza de vida de los pacientes infectados con VIH, la toxicidad de los propios fármacos o el débil sistema inmunitario de los infectados, entre otros.

Como podemos observar en el siguiente diagrama (Fig. 3), los eventos ENOS están relacionados directamente con el VIH por la disminución de las células CD4 que causa el virus, e indirectamente a causa de la activación inmunitaria. El efecto directo del VIH puede ayudar a fallos renales y a aumentar el riesgo de falos de riñón. De la misma forma puede propiciar la aparición de tumores, bloqueando los genes que suprimen los tumores. Principalmente causan inmunodeficiencia, a causa de la disminución de células CD4.El virus también puede causar disfunción en ciertos órganos por su efecto directo en las células hepáticas.

Las coinfecciones con VHB y VHC están directamente relacionadas con los eventos en el hígado, mientras que coinfecciones y translocaciones microbianas impulsan la activación inmune. Aproximadamente el 10% de pacientes infectados con VIH, padecen de Hepatitis B (VHB) y el 27% padecen Hepatitis C. Otras comorbilidades, como el hecho de que el paciente sea fumador o los riesgos cardiovasculares, están relacionadas directamente con los eventos ENOS. El hecho de que un paciente esté infectado con VIH supone mayores riesgos para desarrollar un evento cardiovascular que factores de riesgos más clásicos, como el colesterol o que el paciente sea fumador.

La inflamación o inmunoactivación es una de las causas principales de enfermedades como la progresión tumoral o la fibrosis hepática en la población general. Dicha inflamación, normalmente, se detecta por altos niveles de algunos biomarcadores como el IL-6 o el DD. En pacientes infectados los niveles de estos biomarcadores suelen ser bastante más altos que en controles no infectados, y, aunque sean tratados, los niveles se mantienen más altos que los controles no infectados. Finalmente, la toxicidad de los propios medicamentos (ARVs) contribuye a la aparición de eventos ENOS. A pesar de ello, la medicación es esencial para frenar la duplicación del virus y permitir que el sistema inmune se recupere.

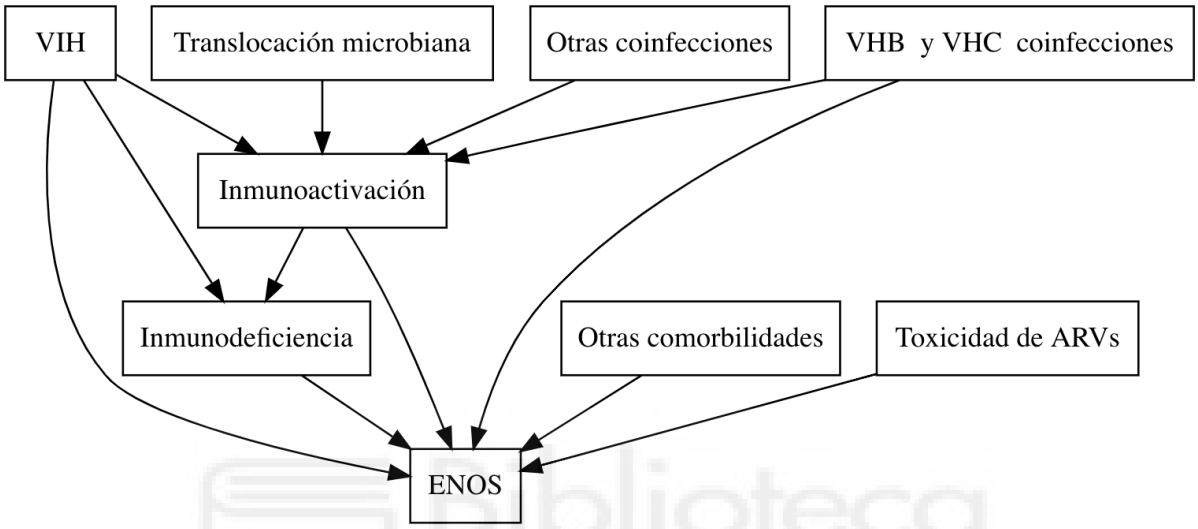


Figure 3: Patogénesis de ENOS

Las investigaciones disponibles en el campo de los eventos ENOS indican que algunos factores que pueden reducir el riesgo de eventos ENOS son comenzar el tratamiento con anterioridad a que el paciente desarrolle una inmunodeficiencia avanzada, dejar de fumar, la reducción de los factores de riesgo cardiovasculares y tratamiento contra el Virus de la Hepatitis C (VHC)

## 1.3 Métodos e indicadores para la detección.

Generalmente, se usa el conteo de células de CD4, CD8, la ratio CD4/CD8 o la carga viral, entre otros, para estudiar la aparición de dichos eventos ENOS. Como se ha comentado anteriormente, el CD4 son un tipo de células del sistema inmunitario, conocidas como linfocitos-T, dedicadas a la producción de anticuerpos cuando detecta algún patógeno o la replicación errónea de las células del organismo. Cuando el paciente es infectado con VIH, el virus se introduce en estas células, haciendo que en vez de producir anticuerpos o de replicarse, generen más copias del virus. Un paciente sano tiene un recuento de células CD4 de 500 - 1600 células/$mm^3$, el cual disminuye con la progresión de la enfermedad si no es correctamente tratada. Se ha establecido de forma general que por debajo de 200 células/$mm^3$ existe un riesgo superior de contraer algún tipo de enfermedad oportunista. Las células CD8 pertenecen a la misma familia que las células CD4, con una función muy similar en el sistema inmunitario, atacan a virus o otras infecciones anormales en el sistema, pero estas son afectadas directamente por el VIH. Ya que la función de las células CD8, a diferencia de las células CD4, no se ve alterada por la infección, estas comienzan a atacar a las células CD4 infectadas, haciendo que cada vez baje más rápido el nivel de estas, mientras sigue aumentando la carga vírica o cantidad de un virus determinado, en este caso VIH, en el sistema.

Habitualmente se recurre a la ratio CD4/CD8, el cual es un indicador de como de fuerte es el sistema inmunitario del paciente y ayuda a predecir como evolucionará la infección. Si el valor de la ratio es igual o mayor a 2, generalmente indica un sistema inmunitario fuerte y pocas probabilidades de estar infectado con VIH, aunque valores muy por encima de lo normal, puede indicar infecciones graves o algunos tipos de cáncer, mientras que, si el valor de la ratio es menor a 1, puede indicar VIH, SIDA o otras enfermedades como anemia o infecciones del sistema nervioso.

A continuación, podemos observar la evolución típica de los valores de CD4, CD8 y carga viral de un paciente que no ha sido tratado (Fig. 4). Comienza con una infección aguda de VIH en las primeras semanas, durante la infección primaria, disparando los valores de carga vírica y de CD8 para combatir la infección, mientras que el conteo de CD4 disminuye a causa del ataque del virus, seguido por una estabilización del virus y de los niveles de CD4 y CD8, a este periodo de tiempo se le suele denominar etapa asintomática, ya que el paciente no muestra indicios, si no es testado, de estar infectado, y tiene duración variable dependiendo de las condiciones del paciente. Finalmente, el virus comienza a duplicarse de forma agresiva, produciendo un gran efecto en el sistema inmunitario, haciendo que decaigan así los valores de CD4 y CD8, hasta el punto de, generalmente, desarrollar SIDA y acabar muriendo.

Las limitaciones de este tipo de valores son que suponen una espera de varios días para obtenerlos e incurren en un coste alto, incluso en algunos lugares con recursos limitados son muy difíciles de obtener. Por otro lado, los propios ARVs usados para tratar la infección, afectan a los niveles de CD4 y de carga viral, entre otros, lo que complica el poder estudiar la asociación entre estos y los eventos ENOS. Es por ello por lo que a menudo se recurre a biomarcadores para poder estudiar diferentes infecciones, como por ejemplo, en este caso, la infección por VIH.

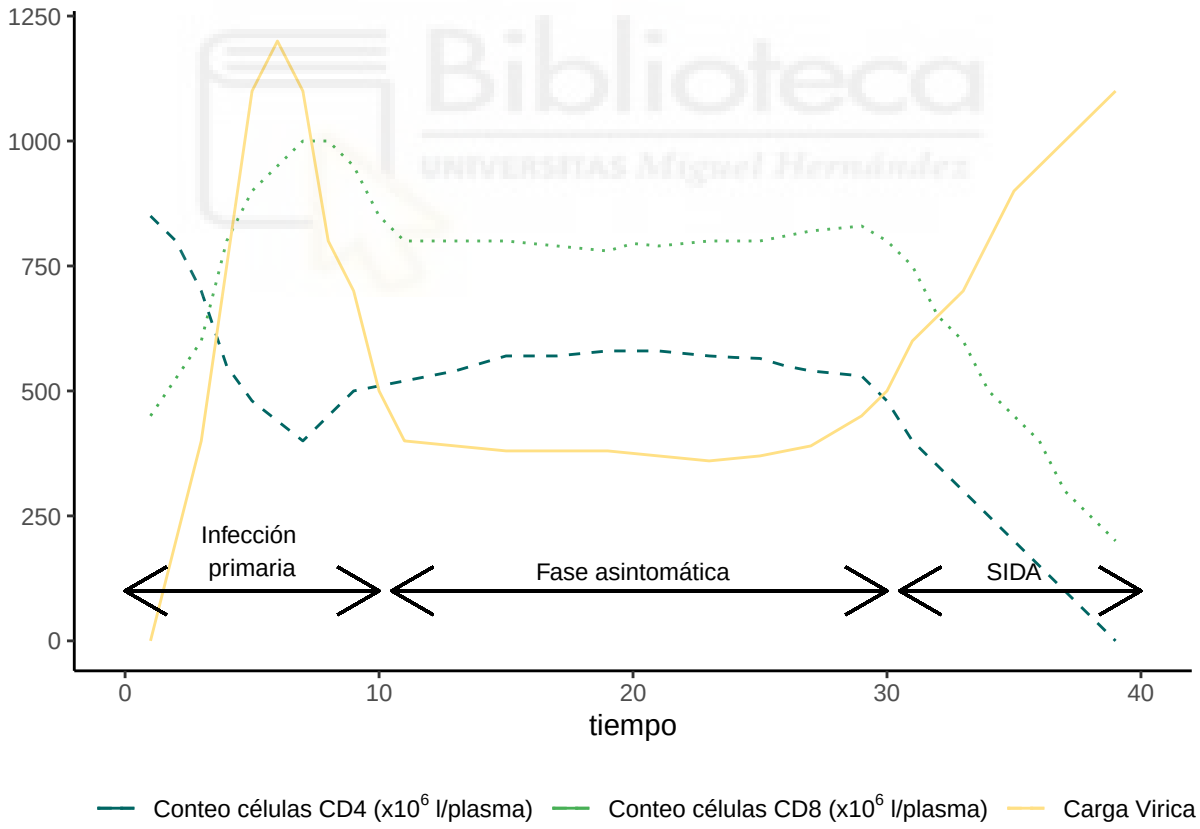


Figure 4: Evolución típica de un infectado con VIH que no recibe tratamiento.

La definición de biomarcador o marcador biológico, según el Instituto Nacional del Cáncer, es: "Molécula biológica que se encuentra en la sangre, otros líquidos o tejidos del cuerpo y es un signo de un proceso

normal o anormal, o de una afección o enfermedad". Los biomarcadores se suelen usar para detectar posibles patologías que pueda sufrir el paciente, en este caso se estudian unicamente los que pueden mantener una relación con los eventos ENOS. Los biomarcadores usados en este estudio se emplean normalmente para la detección de inflamaciones, infecciones y algunos tipos de cáncer. Algunos de ellos son los siguientes:

- PCR: La PCR es un marcador de respuesta inflamatoria aguda, y el aumento de su concentración se ha descrito en muchas enfermedades, incluyendo las enfermedades cardiovasculares, como por ejemplo en la insuficiencia cardiaca. [15]
- TNF: Proteína elaborada por los glóbulos blancos en respuesta a un antígeno (sustancia que hace que el sistema inmunitario brinde una respuesta inmunitaria específica) o a una infección. [7]
- IL-6 y IL-8: Dentro del grupo de las citoquinas, la IL-6 e IL-8 son citoquinas proinflamatorias que favorecen la carcinogénesis al promover la proliferación, generación de radicales libres, la supervivencia celular y la angiogénesis. [24]
- Neopterina: La neopterina es una sustancia liberada por los macrófagos durante el proceso de activación del sistema inmune; por tanto, sus niveles se incrementan en las infecciones. [1]
- VCAM: Molécula de adhesión implicada en la migración linfocitaria y en el reclutamiento de linfocitos al lugar de la inflamación. También se denomina CD106. [20]
- DD: El dímero D (DD) es un producto de degradación de la fibrina formado durante la lisis de un trombo. [3]
- Isoprotanos: Serie de compuestos semejantes a las Prostaglandinas, que son producidos por el ataque de Radicales Libres a Ácidos Grasos insaturados, especialmente el Ácido Araquidónico, de las Membranas celulares. [27]
- Malondialdehído (MDA): El MDA es un acetaldehído fisiológico producido por descomposición de los lípidos insaturados provenientes del metabolismo del ácido araquidónico. [31]
- CD14: "Glicoproteína de membrana anclada por GPI, propia de los monocitos y de sus precursores. Asimismo, puede presentarse en forma soluble y actúa como receptor de lipopolisacárido. [20]
- CD40: Molécula de importancia crucial en la activación del linfocito B, ya que interacciona con la molécula CD40L (ligando del CD40) del linfocito T helper, suministrando una señal imprescindible para la activación, proliferación y diferenciación del linfocito B. Una mutación en el CD40L que impida su unión al CD40 origina una imposibilidad en el cambio de clase de las inmunoglobulinas, produciéndose únicamente el isotipo IgM (síndrome de hiper-IgM). [20]

## 1.4 Marco teorico.

A día de hoy, los biomarcadores son extensamente utilizados para la detección y prevención de una variedad de enfermedades como el cáncer, enfermedades cardiovasculares, etc., por su gran potencia y precisión. Con el incremento de eventos ENOS entre pacientes infectados con VIH, resulta interesante estudiar la relación que guardan algunos biomarcadores, con relación anteriormente probada con otras enfermedades, como infecciones diversas o canceres, con los eventos ENOS. Las diferentes estrategias de predicción usadas en la población no infectada con el virus no pueden ser las mismas que en los pacientes infectados, por las cualidades y circunstancias especiales que el virus causa, por lo tanto, es necesario buscar técnicas alternativas que permitan detectar o predecir varias enfermedades en dichos pacientes, para aplicar lo antes posible el tratamiento adecuado.

En la mayoría de la literatura de este campo, se remarca que no existe estudios suficientes y de suficiente calidad, en cuanto a la cantidad de pacientes en dichos estudios y el seguimiento de los mismos, para poder observar resultados con gran significación y conocer al detalle la relación entre los biomarcadores y las enfermedades. Por otro lado, es necesaria la incorporación de estos biomarcadores a las estrategias

de predicción de enfermedades para poder estudiar la calidad y eficiencia clínica real, ya que, hoy en día, dichos biomarcadores no son usados si no se requieren específicamente para un estudio. Finalmente es importante incorporar la componente tiempo en el estudio, lo que permite estudiar la evolución de dichos biomarcadores a lo largo del tiempo, ya que la mayoría de las investigaciones disponibles toman los valores de los biomarcadores en un momento concreto, generalmente cuando entran en el estudio, lo que limita en gran medida el estudio de la influencia de los biomarcadores en las diferentes enfermedades.

Algunas de las conclusiones encontradas en la literatura son las siguientes:

- Relación entre DD, IL-6, IL-7, sTNFR-I, sRNFR-II y el CD4s (CD4 souble), entre otros, con el desarrollo de eventos ENOS.
- Relación entre algunos biomarcadores como IL-6, ICAM-1, VCAM-1, CD14, CD163, CD4 y CD8 con un mayor riesgo de sufrir enfermedades cardiovasculares en pacientes con VIH.
- Es posible que la relación de DD y IL-6 con el desarrollo de eventos ENOS sea distina entre pacientes con un mayor valor de CD4 y pacientes con un valor de CD4 menor y los niveles inflamatorios de los apcientes son menores cuanto antes se comience a tratar la infección.

# 2 Objetivos.

Como se ha comentado anteriormente, existe un especial interés en detectar relaciones entre diversos biomarcadores y los diferentes eventos ENOS por el avance en detección y prevención que supondrían. En el caso de este estudio, se marca especial interés en la búsqueda de las relaciones entre los biomarcadores y si el paciente desarrolla o no un evento.

El presente estudio tiene dos objetivos principales. En primer lugar, la utilización de diferentes técnicas estadísticas, algunas más clásicas y otras más actuales, para modelizar los datos disponibles, con la intención de observar con cual de ellas se consiguen mejores resultados. Principalmente se modelizan para obtener las relaciones entre los diferentes biomarcadores, conteos de células y otra información interesante de la que se dispone, una variable binaria que indica si el paciente desarrolla un evento ENOS o no. También se tratará de aplicar los mismos procedimientos, pero esta vez para encontrar relaciones con otras variables, como pueden ser el tipo de evento que desarrolla el paciente o si al finalizar el estudio el paciente muere o no. Una vez realizados dichos análisis, se comparan entre ellos para poder extraer conclusiones acerca de cual de las técnicas usadas tiene una mayor eficiencia, mejores resultados y cual se adapta mejor a cada tipo de problema planteado a lo largo del estudio. Las técnicas que se usan son las siguientes:

- Regresión Logística (GLM): La cual permite obtener los riesgos de las diferentes variables usadas en el modelo, dejando entender como se relacionan las diferentes variables entre ellas, y devuelve la probabilidad de que un paciente con unas características determinadas termine desarrollando un evento ENOS o no.
- Análisis de Supervivencia: Con el análisis de supervivencia incluimos al tiempo, bien sea desde que entra hasta que muere, o desde que entra hasta que desarrolla un evento, etc., lo que permite observar como afectan diferentes variables y su influencia en ciertos eventos (muerte, desarrollo de un evento ENOS, etc.) a lo largo del tiempo.
- Arboles de decisión: Similar a la regresión logística en cuanto al output, probabilidad de que un paciente pertenezca a un grupo o a otro, pero diferente en funcionamiento, ya que crea un "mapa" o una serie de "reglas" optimas, según las cuales clasificar a los pacientes. Nos permite observar que variables y que valores de dichas variables, tienen un efecto significativo en que el paciente desarrolle o no un evento ENOS.
- Redes neuronales: Con el uso de redes neuronales se pierde, en cierta medida, el conocimiento que si que dan otras técnicas, sobre la influencia de las diferentes variables, pero se gana en precisión en cuanto a la predicción de eventos, ya que es una de las técnicas más potentes detectando patrones y relaciones en los datos, patrones que para otras tecnicas son muy dificiles o imposibles de detectar.

Finalmente, en segundo lugar, con los diferentes análisis realizados a lo largo del estudio, se obtienen ciertas conclusiones sobre los diferentes efectos y relaciones de los biomarcadores y otras variables, con las variables estudiadas (evento o no, tipo de evento, riesgos, etc.). Es interesante comparar las conclusiones obtenidas a lo largo del estudio, con la información disponible en la literatura, y buscar concordancias y discrepancias entre ellas.

# 3 Material

Para el presente estudio se dispone de la información medica de 557 pacientes infectados con VIH, recogida en el Hospital General Universitario de Elche por el propio personal del hospital mediante consultas, diferentes análisis y el seguimiento correspondiente de cada paciente. La obtención de los datos para el estudio se ha hecho de forma retrospectiva, se tienen ciertos casos con unas características determinadas, y para cada uno de estos casos se busca un control con características similares (mismo grupo de edad, sexo, patologías, etc.), pero con un diagnostico diferente. Todos los datos disponibles pertenecen a un periodo de 12 años, desde el año 2004 hasta el año 2015.

La base de datos usada para el estudio ha sido construida con los datos obtenidos de dos fuentes. La base de datos inicial, obtenida del Hospital General Universitario de Elche, contiene los pacientes que son parte del estudio, indicando cuales son casos y cuales controles, los diferentes biomarcadores que se usan en el estudio, y otras variables que pueden ser de interés, como el sexo del paciente o su edad. El resto de los datos son proporcionados por la Cohorte de la Red de Investigación en Sida (CoRIS), divididos en las siguientes bases de datos:

- ENOS: Contiene los eventos No-SIDA y otros eventos clínicos de todos los paciente e información relevante sobre estos, como la fecha del evento. Para mayor claridad ha sido necesario agrupar los eventos por categorías: eventos cardiovasculares, renales, de riñón, neoplásicos, metabólicos, de huesos, neuropsiquiatrico, otros eventos, o no-evento.
- BASAL: Información relacionada con la infección de VIH de cada paciente, como la fecha de entrada en la cohorte, fecha de detección de la infección, si recibe o no tratamiento, si el paciente tiene o no SIDA, etc.
- CD4 y CD8: Valor del conteo de células CD4 y CD8 en distintas fechas de cada paciente. Ha sido necesario modificar dicha base para que fuese posible su análisis, por lo que finalmente solo se usa el valor de CD4 y de CD8 inicial, el final, y en caso de que el paciente haya sufrido un evento ENOS, el valor más cercano a dicha fecha.
- CARGA VIRAL: De la misma forma que en la base anterior, contiene el valor de carga vírica de cada paciente en diferentes fechas, y finalmente solo se usa el valor inicial, el final, y el más cercano al evento ENOS, en caso de que exista.
- MUERTE: Información relacionada con la salida del paciente del estudio, bien por que se ha perdido durante el estudio, por muerte (indicando causa de muerte, fecha, etc.) o por finalizar el tratamiento.

El estudio requiere de la creación de varias variables que no se encuentran en las bases de datos mencionadas para realizar algunos de los análisis, como por ejemplo, CD4/CD8, que es un ratio que se suele usar en el estudio de los eventos ENOS, una variable binaria que indique si el paciente desarrolla o no un evento ENOS, o de la misma forma, otra serie de variables binarias que indiquen si el paciente desarrolla un tipo de evento en concreto o no. Todas estas variables son creadas en función de lo posible (por ejemplo, no existe un valor de CD8 siempre que hay valor de CD4, por problemas en la recogida de datos, por tanto, no siempre se puede crear la ratio CD4/CD8), y son añadidas al resto de variables disponibles.

En la siguiente tabla (Tab. 1) se muestran las principales variables usadas en el estudio, junto a su descripción y tipología (numérica o categórica)

| Nombre | Definición | Continua | Categórica | Niveles |
|---|---|---|---|---|
| cc.ex | Indica si el paciente es caso o control | | x | 0:Caso<br>1:Control |
| pcr | Biomarcador | x | | |
| tnf | Biomarcador | x | | |
| il6 | Biomarcador | x | | |
| il8 | Biomarcador | x | | |
| ifn | Biomarcador | x | | |
| nop | Biomarcador | x | | |
| vcam | Biomarcador | x | | |
| icam | Biomarcador | x | | |
| dd | Biomarcador | x | | |
| isopros | Biomarcador | x | | |
| mda | Biomarcador | x | | |
| cd14 | Biomarcador | x | | |
| cd163 | Biomarcador | x | | |
| edad | Edad del paciente | x | | |
| sexo | Sexo del paciente | | x | 0:Hombre<br>1:Mujer |
| cat | Forma en la que el paciente se contagió de VIH | | x | sex:Sexo<br>idu:Drogas |
| vch | Si el paciente está infectado o no por el virus de la Hepatitis C | | x | 0:No 1:Si |
| V39 | Biomarcador | x | | |
| eventsENOS* | Evento ENOS sufrido | | x | bone<br>cardiovascular<br>liver<br>metabolic<br>neoplastic<br>neuropsychiatric<br>noevento<br>others<br>renal |
| DROPY* | Si el paciente se pierde a lo largo del estudio | | x | 0:No 1:Si |
| DEATHY* | Si el paciente muere | | x | 0:No 1:Si |
| RECARTY* | Si el paciente recibe tratamiento por la infección de VIH | | x | 0:No 1:Si |
| AIDSY* | Si el paciente tiene SIDA | | x | 0:No 1:Si |
| RNAV** | Valor de la carga virica del paciente | x | | |
| CD4V** | Valor del conteo de células CD4 del paciente | x | | |
| CD8V** | Valor del conteo de células CD8 del paciente | x | | |
| CD4CD8** | Valor de la ratio CD4/CD8 | x | | |
| evenono | Si el paciente desarrolla un evento ENOS o no | | x | 0:No 1:Si |
| CD4200 | Si el conteo de células CD4 del paciente es inferior o superior a 200 | | x | 0:CD4<200<br>1:CD4>200 |

*Note:*
*Se dispone de la fecha **Se dispone de los valores iniciales, más cercanos al evento, y finales

Tabla 1: Variables del estudio.

Toda esta información es procesada para obtener una única base de datos con los datos más relevantes para el estudio. Para el presente estudio, únicamente se considera el primer evento de cada paciente, descartando la información relacionada con eventos posteriores. El proceso de modificar los datos para poder usarlos en el estudio, por la complejidad de las relaciones y estructuras de los datos, ha durado aproximadamente un mes y medio.

# 4 Modelos aplicados a ciencias de la salud

## 4.1 Regresión Logística

### 4.1.1 GLM

La regresión logística es un tipo de modelo para datos binarios perteneciente al conjunto de modelos lineales generalizados (GLM). Los GLM son un conjunto de modelos con los cuales es posible modelar variables tanto continuas como discretas, y que permiten que la variable dependiente Y tenga una distribución distinta de la normal, normalmente alguna distribución paramétrica. Algunos GLM son la regresión lineal clásica, el análisis de varianza (ANOVA), el análisis de la covarianza (ANCOVA), la regresión logística o los modelos log lineales para tablas de contingencia, entre otros.

### 4.1.2 Modelo

La regresión logística estudia la relación entre una variable categórica dependiente (Y), la cual únicamente tiene dos niveles, 0 y 1, y un conjunto de variables independientes explicativas (X), las cuales pueden ser cuantitativas o cualitativas. Se supone que la variable dependiente Y tiene una distribución de la familia exponencial.

Queremos modelizar la probabilidad condicionada $Pr(Y = 1|X = x)$, es decir, la probabilidad de que la variable Y tome valor 1, sabiendo que la variable explicativa toma x valor.

Llamamos $\pi_i$ a la probabilidad de que $y$ tome el valor 1 cuando $x = x_i$, por tanto:

$$\pi_i = P(y = 1|x_i) \qquad 1 - \pi_i = P(y = 0|x_i)$$

Cuando la variable Y es continua y con distribución normal $Y_i \sim N(\mu_i, \sigma^2)$, se puede ajustar un modelo lineal de la siguiente forma:

$$Y = \beta_0 + \sum_i \beta_i \times x_i$$

En este caso Y es una variable dicotomica que sigue una distribución condicional de $Y$ sobre $X = x$ de Bernoulli tal que $Y \sim bern(0, \pi_i)$ y:

$$Y_i = \begin{cases} 1 \text{ si } exito \\ 0 \text{ si } fracaso \end{cases}$$

Donde 1 siempre es algún tipo de suceso, 1=muerto/0=vivo, 1=enfermo/0=sano, etc. Por tanto, un ajuste lineal presenta una serie de problemas:

- El objetivo principal es modelizar las probabilidades de exito y fracaso, y con un modelo lineal, es muy probable obtener resultados por encima de 1 o por debajo de 0. Se busca un modelo que sea capaz de predecir la probabilidad de exito en un rango $[0, 1]$, para hacer que los resultados sean interpretables.
- Por otro lado, en un modelo lineal se presupone una relacción lineal entre la variable respuesta y las variables explicativas, lo cual supone un aumento constante de la probabilidad a medida que crece (o disminuye) el valor de $x$. Cuando hablamos de probabilidades en un rango $[0, 1]$, la relación entre dicha probabilidad y el valor de $x$ no es lineal, si la probabilidad ya es muy cercana a 0 o 1, un gran cambio en el valor de $x$, no supone un gran aumento en el valor de la probabilidad (Fig. 5)

Debido a esto, se han buscado diferentes alternativas de la forma:

$$\pi_i = g(\beta_0 + \sum_i \beta_i \times x_i) \quad o \quad g^{-1}(\pi_i) = \beta_0 + \sum_i \beta_i \times x_i$$

Donde $g$ es un nexo. Un nexo (o "link") es una función que se aplica a la variable respuesta Y para transformar el output, en concreto, para conseguir que las probabilidades de Y se ajusten a un rango 0-1 y no tengan una relación directamente lineal con las variables explicativas. Más adelante (Fig. 6) se muestran algunos de los distintos nexos que se pueden utilizar.

Aplicando una transformación logarítmica, se soluciona el problema de la relación lineal entre $Y$ y $X$, como se puede observar a continuación (Fig. 5), pero sigue sin solucionarse el problema de la acotación entre 0 y 1.

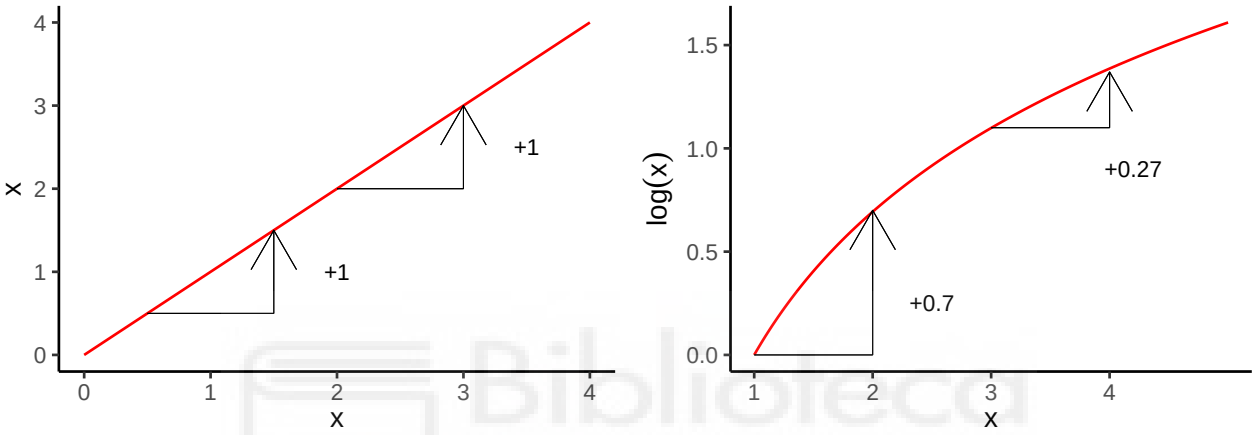


Figure 5: Transformación logarítmica

Un posible nexo ($g$) es el llamado logistico o modelo logit, el cual viene dado por:

$$logit(\pi_i) \Longrightarrow log\frac{\pi_i}{1+\pi_i} = \beta_0 + \sum_i \beta_i \times x_i$$

Escrito en terminos de $\pi_i$:

$$\pi_i = \frac{e^{\beta_0 + \sum_i \beta_i \times x_i}}{1 + e^{\beta_0 + \sum_i \beta_i \times x_i}} = \frac{1}{1 + e^{(-\beta_0 + \sum_i \beta_i \times x_i)}}$$

Como se observa (Fig. 6), los valores de la función logística están siempre comprendidos entre 0 y 1, lo que soluciona el problema de las probabilidades acotadas y hace que los parámetros del modelo sean interpretables.

Existen distintos tipos de nexos o transformaciones que se suelen aplicar, los cuales pueden generar soluciones destinas para los mismos datos, y cada uno se suele usar en situaciones concretas. Los nexos logit y probit son igual de sensibles por encima y por debajo de $x = 0.5$, mientras que el nexo log-log es poco sensible a variaciones en $x$ cuando $x < 0.5$.

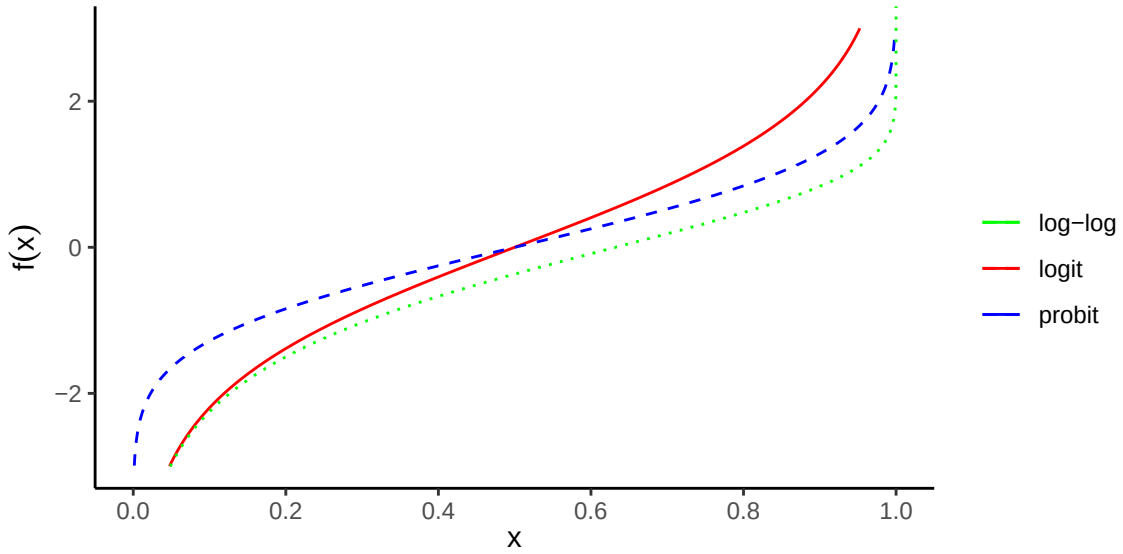


Figure 6: Función logística

#### 4.1.3 Razón de probabilidad (Odds Ratio -OR-)

Los OR indican en que medida se modifican las probabilidades de pertenecer a un grupo o a otro por una unidad de cambio en una variable continua, o por pasar de una categoría a otra, en las variables categóricas.

Tenido el siguiente modelo

$$logit\left(\pi_i\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

donde $x_1$ es una variable binaria y $x_2$ es una variable continua.

Para interpretar $\beta_1$, fijamos el valor de $x_2$.

Para $x_1 = 0$:

$$odds = e^{\beta_0 + \beta_1 (0) + \beta_2 x_2} = e^{\beta_0 + \beta_2 x_2}$$

Para $x_1 = 1$:

$$odds = e^{\beta_0 + \beta_1 (1) + \beta_2 x_2} = e^{\beta_0 + \beta_1 + \beta_2 x_2}$$

Por tanto, el OR de pasar de $x_1 = 0$ (nivel de referencia) a $x_1 = 1$ es:

$$OR = \frac{odd\ cuando\ x_1 = 1}{odd\ cuando\ x_1 = 0} = \frac{\beta_0 + \beta_1 + \beta_2 x_2}{\beta_0 + \beta_2 x_2} = e^{\beta_1}$$

Por otro lado, para interpretar $e^{\beta_2}$, fijamos el valor de $x_1$

Para $x_2 = k$, siendo $k$ cualquier valor:

$$odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 k}$$

Para $x_2 = k + 1$:

$$odds = e^{\beta_0+\beta_1 x_1+\beta_2(k+1)} = e^{\beta_0+\beta_1 x_1+\beta_2 k+\beta_2}$$

Por tanto, el OR de aumentar una unidad el valor de $x_2$ es:

$$OR = \frac{odd\ cuando\ x_1 = k+1}{odd\ cuando\ x_1 = k} = \frac{\beta_0+\beta_1 x_1+\beta_2 k+\beta_2}{\beta_0+\beta_1 x_1+\beta_2 k} = e^{\beta_2}$$

En resumen, $\beta_i$ da información sobre la importancia de cada una de las variables del modelo sobre la variable respuesta Y, pero si se quiere conocer como afecta el valor de cada variable a la probabilidad de que la variable Y tome un valor o otro, hay que recurrir a los OR ($e^{\beta_i}$). Si se tiene, por ejemplo, una variable binaria Y que indica si un paciente muere (1) o no (0), y una variable binaria X que indica tener una enfermedad (1) o no (0), si el $OR > 1$, el hecho de tener dicha enfermedad aumenta la probabilidad de que el individuo muera, mientras que si el $OR < 1$, tener la enfermedad, reduce la probabilidad de que el individuo muera. Que $OR > 1$ no garantiza la relación entre tener la enfermedad y morir, ya que esta podría estar influenciada por una tercera variable que no se recoge en el modelo.

### 4.1.4 Evaluación del modelo

Para resumir los datos y estudiar la bondad del ajuste, se suele recurrir a una tabla de clasificación $2 \times 2$. Buenos resultados en una tabla de contingencia no indica necesariamente un buen ajuste del modelo, pero es un buen indicio.

Para comenzar es necesario calcular los valores predichos, por tanto, se ajusta el modelo y se calculan los valores de $\pi_i$, las probabilidades de pertenecer a las categorías de la variable Y, con las que es posible calcular la categoría a la que pertenecería cada individuo de la siguiente forma:

$$\widehat{Y_i} = \begin{cases} 1 \text{ si } \pi_i(x) > c \\ 0 \text{ si } \pi_i(x) < c \end{cases}$$

Donde $c = 0.5$ normalmente, aunque puede ser otro valor.

La tabla tiene el siguiente formato:

| | Observado | | |
|---|---|---|---|
| Clasificado | Y=1 | Y=0 | TOTAL |
| $\widehat{Y_i} = 1$ | $\left\{\widehat{Y_i} = 1, Y = 1\right\}$ | $\left\{\widehat{Y_i} = 1, Y = 0\right\}$ | $\left\{\widehat{Y_i} = 1\right\}$ |
| $\widehat{Y_i} = 0$ | $\left\{\widehat{Y_i} = 0, Y = 1\right\}$ | $\left\{\widehat{Y_i} = 0, Y = 0\right\}$ | $\left\{\widehat{Y_i} = 0\right\}$ |
| TOTAL | $\{Y = 1\}$ | $\{Y = 0\}$ | n |

Tabla 2: Matriz de confusión.

En las casillas se indica el conteo de casos que cumplen las condiciones entre llaves, por ejemplo, donde pone $\left\{\widehat{Y_i} = 1, Y = 0\right\}$, se indica la cantidad de casos que realmente pertenecen a la categoría 0, pero el modelo predictivo las ha clasificado como categoría 1.

Los indicadores más comunes de las regresiones logísticas, especialmente en el ámbito sanitario, son la sensibilidad y la especificidad del modelo. La sensibilidad hace referencia a la probabilidad de detectar un verdadero positivo ($\widehat{Y_i} = 1, Y = 1$), y la escificidad es la probabilidad de detectar un verdadero negativo ($\widehat{Y_i} = 0, Y = 0$).

$$\text{Sensibilidad} = \widehat{P}(\widehat{Y_i} = 1/Y = 1) = \frac{\left\{\widehat{Y_i} = 1, Y = 1\right\}}{\{Y = 1\}}$$

$$\text{Especificidad} = \widehat{P}(\widehat{Y_i} = 0/Y = 0) = \frac{\left\{\widehat{Y_i} = 0, Y = 0\right\}}{\{Y = 0\}}$$

Los errores de clasificación se calculan de la siguiente forma:

$$\text{Falso Positivo (FP)} = \widehat{P}(\widehat{Y_i} = 1/Y = 0) = \frac{\left\{\widehat{Y_i} = 1, Y = 0\right\}}{\{Y = 0\}}$$

$$\text{Falso Negativo (FN)} = \widehat{P}(\widehat{Y_i} = 0/Y = 1) = \frac{\left\{\widehat{Y_i} = 0, Y = 1\right\}}{\{Y = 1\}}$$

$$\text{Error global} = \frac{\left\{\widehat{Y_i} = 0, Y = 1\right\} + \left\{\widehat{Y_i} = 1, Y = 0\right\}}{n}$$

Dependiendo de la situación en la que se utilice la regresión logística, se le puede dar más importancia a maximizar sensibilidad o especificidad, o a minimizar alguno de los errores de clasificación. Como se ha comentado anteriormente, generalmente, se suele usar $c = 0.5$ como punto de corte a la hora de clasificar, pero el punto de corte puede cambiar, cambiando así tanto la tabla de contingencia como el valor de la sensibilidad y la especificidad.

Otro método ampliamente usado para la evaluación de la regresión logística es la Receiver Operating Characteristic (curva ROC), con la cual se pretende obtener un valor continuo que describa la "calidad" del modelo. El principal propósito de una regresión logística es discriminar entre dos categorías, por ejemplo, en enfermos (1) y no enfermos (0), no tendría ninguna utilidad un modelo que categorice como enfermos a todos los pacientes, ya que tendría una sensibilidad del 100%, pero una especificidad del 0%. Con la curva ROC se trata de evaluar que existe una discriminación real por parte del modelo, y que la clasificación que realiza es mejor que lanzar una moneda al aire.

La curva ROC es un método gráfico basado en la tabla de contingencia, la sensibilidad, la especificidad, y el punto de corte (c). Se calcula, para todos los posibles puntos de corte de 0 a 1, la sensibilidad y el falso positivo (o 1-especificidad) y se representan, 1-especificidad en el eje x, y la sensibilidad en el eje y. Se traza una línea de 45º de la esquina izquierda inferior a la superior derecha, y cuanto más por encima se encuentre la curva ROC, mejor discrimina el modelo.

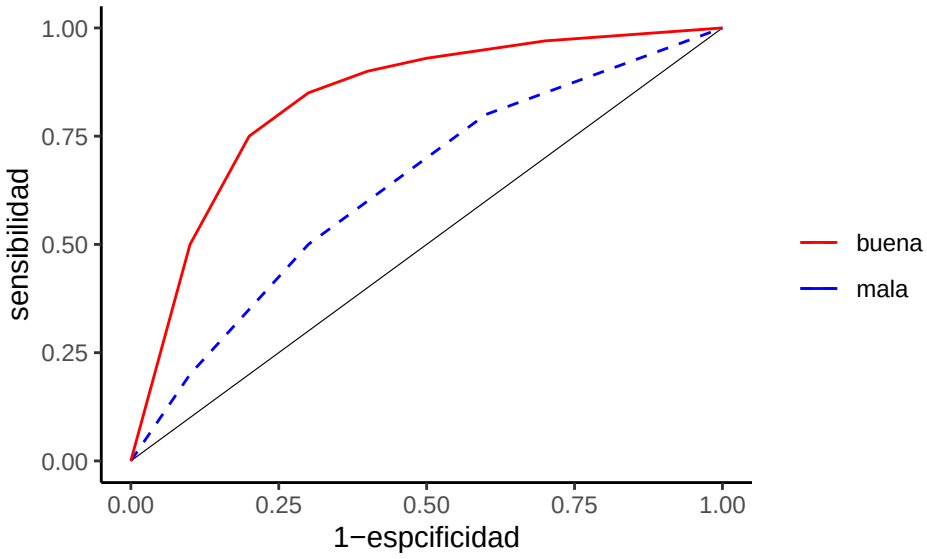


Figure 7: curva ROC

Se usa como medida de discriminación del modelo el área (A) por debajo de la curva, y normalmente se rige por las siguientes normas:

- $A > 0,9$ Discriminación muy buena.
- $0,9 > A > 0,8$ Discriminación buena.
- $0,8 > A > 0,7$ Discriminación aceptable.
- $0,7 > A$ Discriminación mala.
- $A = 0,5$ No hay discriminación.

## 4.2 Análisis de supervivencia

Un tipo de datos normalmente estudiados, sobre todo en el ámbito de la medicina, es el tiempo que transcurre desde un punto de inicio hasta el suceso de un evento, por ejemplo, desde que un paciente es ingresado con cierta enfermedad hasta que muere, tiempo que transcurre desde que se da cierto fármaco a un paciente hasta que se recupera, o, en otro ámbito como puede ser la ingeniería, el tiempo que transcurre desde que se pone en funcionamiento cierta maquina, hasta que se estropea. Este tipo de datos requieres técnicas especificas por la particularidad del propio dato tiempo hasta evento.

Normalmente se trabaja con un periodo de estudio, con un inicio y un final claramente fijados, dentro del cual se estudia el comportamiento de los diferentes individuos.

Como se puede observar en el siguiente diagrama (Fig. 8), no todos los individuos entran al estudio al mismo momento, ni todos abandonan el estudio a la vez ni por las mismas causas. Se suelen usar dos tipos de datos:

- Datos censurados: Estos son los individuos que han entrado al estudio y no han experimentado el suceso estudiado antes de que el estudio terminase, o bien individuos que no han llegado al final del estudio y la ultima información disponible de estos indica que no han experimentado el evento, pero por la razón que sea, ya no forma parte del estudio. (Denotado por "X")
- Eventos: Son los individuos que a lo largo del estudio han sufrido el evento estudiado. (Denotado por "D")

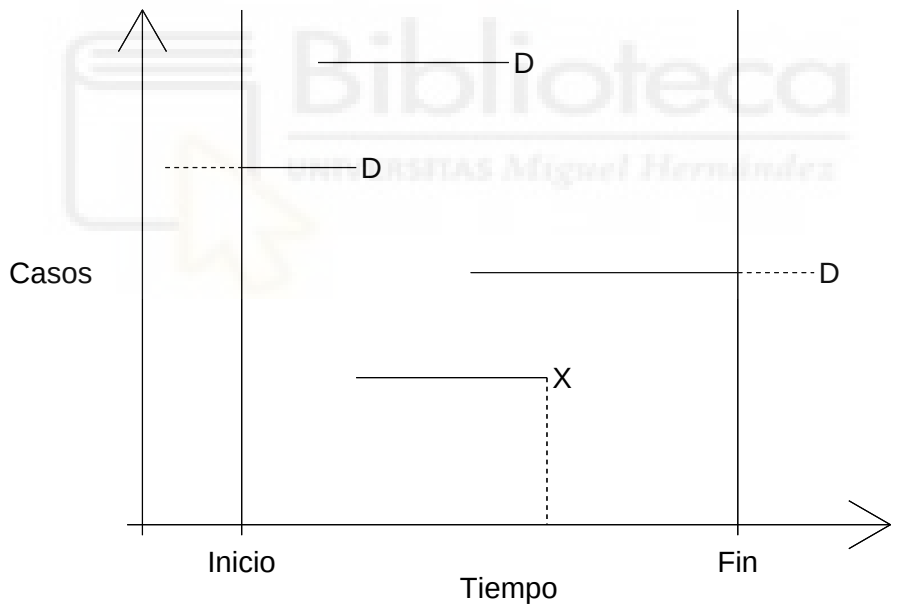


Figure 8: Entrada y salida de los pacientes en el estudio.

Por ejemplo, se estudia el tiempo transcurrido desde la detección de cierta enfermedad hasta la muerte. Los dos primero individuos entran al estudio en momentos diferentes, ya que al segundo se le detectó la enfermedad antes del inicio del estudio (censura por la izquierda), mientras que al primero se le detectó después del inicio, y ambos mueren en cierto momento. El tercer individuo entra al estudio en cierto momento, y no muere hasta después de que el estudio termine, por tanto, su muerte no se tiene en cuenta (censura por la derecha). Finalmente, el cuarto individuo, entra al estudio, y en cierto momento, se pierde el contacto con el y abandona el estudio, pero la ultima información que se tenia de dicho individuo, es que no había muerto (censura por la derecha).

A continuación, se introduce la notación básica del análisis de supervivencia. Por un lado, existen los siguientes parámetros:

$$T = \text{tiempo de supervivencia } (T \geq 0)$$

$$t = \text{valor concreto para T}$$

$$d = (0, 1) \begin{cases} 1 \text{ si evento} \\ 0 \text{ si censura derecha} \end{cases}$$

Donde T es una variable aleatoria igual o superior a 0 que denota el tiempo de supervivencia de un individuo, t es cualquier valor especifico de la variable T, y finalmente, d, la cual denota si un individuo experimenta el evento estudiado (muerte, enfermedad, etc.) o si, por otro lado, es censurado, por alguna de las razones mencionadas anteriormente (fin de estudio, excluido del estudio, etc.).

Por otro lado, tenemos dos funciones que se usan en todos los análisis de supervivencia:

La función de supervivencia $S(t)$ representa la probabilidad de que la supervivencia sea superior a un valor t.

$$S(t) = P(T > t)$$

Teóricamente, la función $S(t)$ se puede representar como una curva decreciente, continua, con rango de 0 a Infinito, y cuando $t = 0$, la probabilidad de supervivencia es de 1 (100%), mientras que cuando $t = \text{Inf}$, la probabilidad de supervivencia es 0 (0%).

Dichas propiedades corresponden a una curva de supervivencia teórica, en la practica, generalmente, se obtiene un gráfico escalonado a causa de que ningún estudio tiene una duración infinita, no todos los individuos del estudio experimentan el evento, y algunos de estos individuos se pierden a lo largo del estudio (censura).

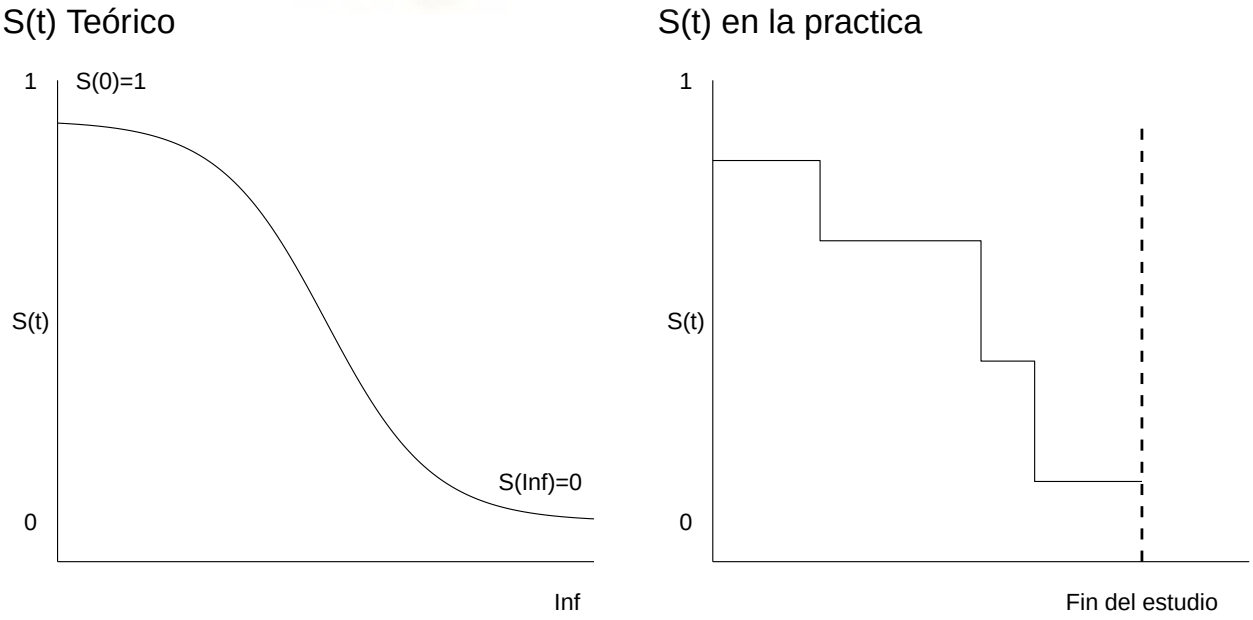


Figure 9: Función de supervivencia

La función de tasa de supervivencia (o riesgo) $h(t)$ viene dada por:

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

Indica el potencial instantáneo por unidad de tiempo de sufrir un evento, teniendo en cuenta que el individuo ha sobrevivido hasta el momento t. Cuando se habla de potencial instantáneo, se hace referencia al potencial que tiene el evento de suceder si se mantiene igual hasta el final, pero como en un análisis de supervivencia, con el avance del tiempo, disminuye la probabilidad de supervivencia (no sufrir un evento), dicho potencial no es constante.

El resultado de $h(t)$ depende tanto del valor de $t$ como de la unidad de tiempo en la que se mida ($\Delta t$), dando resultados distintos si se mide en días, meses, medios días, 3/4 de mes, etc.

Por ejemplo, cuando $t = 5$, $h(t) = 3.2$, pro tanto, teniendo en cuenta que el individuo ha sobrevivido hasta el momento 5, en ese momento tiene un potencial de sufrir el evento estudiado de 3.2, pero a medida que aumente el valor de $t$, $h(t)$ será mayor.

En resumen, $S(t)$ describe directamente la supervivencia, mientras que $h(t)$ es una medida del potencial instantáneo de sufrir un evento.

### 4.2.1 Kaplan-Meier

Kaplan-Meier se usa para estudiar la fracción de individuos que han sobrevivido durante una cierta cantidad de tiempo y sirve para obtener una visión general de la población de estudio. El principal objetivo de un análisis de supervivencia no es la estimación de $S(t)$, generalmente el interés reside en estudiar como cambian los tiempos de supervivencia entre diferentes grupos o en función de covariables.

El estimador de Kaplan-Meier viene dado por:

$$\hat{S}(t) = \prod_{t_j < t} \left(1 - \frac{r_j}{d_j}\right)$$

Con intervalos de confianza:

$$\hat{S}(t) \pm 1_{1-\frac{\alpha}{2}} \hat{\sigma}(t)$$

Donde:

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}$$

Siendo $t_j$ los diferentes momentos donde se observan eventos, $d_j$ la cantidad de individuos que experimentan el evento en el momento $t_j$ y $r_j$ las cantidades de individuos en riesgo, estos son los que no han sufrido el evento y no han sido censurados en el momento $t_j$.

### 4.2.2 Modelo de Riesgos Proporcionales de Cox

Generalmente, el modelo de regresión de Cox viene dado en términos de riesgos de la siguiente forma:

$$h(t, X_i) = h_0(t) \times e^{\sum_0^i \beta_i X_i}$$

Donde $X_i$ es un conjunto de variables explicativas.

El modelo está compuesto por dos partes, la primera, $h_0(t)$, definido como el riesgo basal en el momento $t$, y la segunda, el exponente de la suma de la expresión lineal $\beta_i X_i$ para todo $i$. El primer termino está en función del tiempo $t$, mientras que el segundo termino va en función de las diferentes variables explicativas.

En la practica, este modelo se suele usar con carácter descriptivo, siendo el interés principal la estimación e interpretación de los parámetros $\beta_i$.

### 4.2.3 Hazard Ratio (HR)

De la misma forma que sucede en la regresión logística, los parámetros ($\beta_i$) dan información en relación con la importancia de $X_i$ sobre el desarrollo del evento, pero si se quiere conocer como afecta un cambio en la variable (cambio de categoría en las variables categóricas, o un aumento en las variables continua), hay que recurrir a los hazard ratios, los cuales vienen dados de la siguiente forma:

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} = \frac{h_0(t) \times e^{\sum_0^i \beta_i X_i^*}}{h_0(t) \times e^{\sum_0^i \beta_i X_i}} = e^{\sum_0^i \beta_i (X_i^* - X_i)}$$

Donde $X^*$ es el cambio de categoría en una variable categórica ($X = 0$, $X^* = 1$) o el aumento de una unidad en una variable continua ($X = x$, $X^* = x + 1$), por tanto:

$$\widehat{HR} = e^{\beta_i}$$

Si el $HR < 1$, $X_i$ disminuye la probabilidad de experimentar el evento, mientras que si $HR > 1$, dicha variable aumenta la probabilidad de experimentarlo.

Es importante remarcar que en la regresión de Cox se asume que los $HR$ son constantes a lo largo del tiempo, ya que, como se comentaba anteriormente, el único elemento que depende del tiempo es el $h_0(t)$, el cual, como se muestra en la formula anterior, se cancela a la hora de calcular el HR.

### 4.2.4 Otros modelos.

Hasta el momento se ha hablado de un analisis de supervivencia en el cual podian suceder dos sucesos, el individuo es censurado del estudio o el paciente sufre el evento estudiado. El siguiente paso dentro del analisis de supervivencia son los modelos de riesgos competitivos, los cuales no son el objetivo del presente estudio, pero tienen como particularidad que estos pueden tener en cuenta diferentes tipos de evento, no solo uno. En los modelos estudiados hasta el momento, un ejemplo de evento a estudiar podria ser la muerte, mientras que en los modelos de riesgos competitivos, se podrian estudiar distintas causas de muerte con un solo modelo.

## 4.3 Árbol de clasificación

Los arboles de clasificación son un tipo de algoritmo recursivo ampliamente usado en el ámbito del Machine learning para clasificar a una población en 2 o más categorías. Siendo Y una variable respuesta aleatoria y X un conjunto de variables predictoras ($X_i$), se pretende estudiar la probabilidad condicionada de Y en función de las covariables $X_i$.

La función principal de un árbol de clasificación es crear una estructura lógica con las variables $X_i$ que sea capaz de clasificar una población dada en las diferentes categorías de $Y$ de la forma más eficiente posible.

### 4.3.1 Elementos de un árbol.

Un árbol de clasificación está formado por los siguientes elementos:

- **Nodo raíz**: Es el primer nodo, el cual no tiene nodos anteriores y realiza la primera partición en otros dos nodos
- **Nodo intermedio**: El nodo intermedio es el que tiene un nodo antecesor y realiza una división en otros dos nodos.
- **Nodos terminales**: Los nodos terminales son los que tienen un nodo antecesor, pero no realiza ninguna partición más, no tiene nodos predecesores, por tanto, sería el output del árbol.
- **Ramas**: Las ramas son las conexiones lógicas de cada nodo, un nodo suele tener dos ramas y, exceptuando el nodo raíz, todo nodo proviene de una rama.

Es común referirse a los nodos como nodo padre o nodo hijo, donde el nodo padre es aquel que se divide en dos sub-nodos, los cuales son los nodos hijos.

Cada nodo (raíz o intermedio) representa un "test" en relación con el atributo o variable explicativa ($X_i$) al que se haga referencia en dicho nodo, con dos posibles respuestas, si o no. Para una variable categórica con 5 categorías diferentes, un nodo puede reflejar la pertenencia o no a una o varias de las categorías, mientras que, en una variable continua, un nodo puede reflejar que el valor de dicha variable sea o no mayor a un cierto valor.

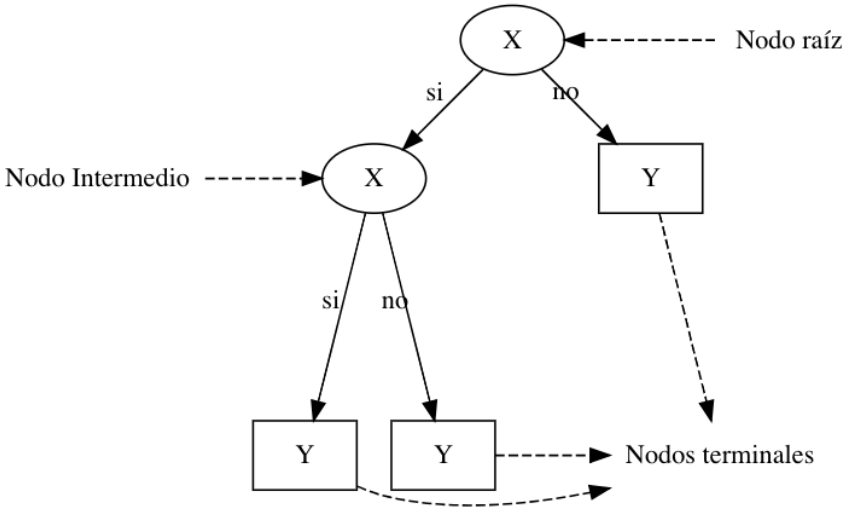


Figure 10: Estructura básica de un árbol

A continuación, se muestra un ejemplo de como funciona un caso simple de árbol de clasificación, con dos variables explicativas ($X_1$ y $X_2$) y una variable respuesta Y con dos categorías, 1 y 2. A la izquierda se puede observar como se realizan las particiones sobre el plano y a la izquierda la representación gráfica del árbol de clasificación.

Como se puede observar, en el plano se divide en dos secciones, una sección donde $X_2 > 7.5$ y otra sección donde $X_1 < 7.5$, y de la misma forma se subdividen estas dos secciones en función de si $X_1 < 6.5$ *o* $X_1 > 6.5$, en el caso de la primera sección, o en función de si $X_1 < 3$ *o* $X_1 > 3$, en el caso de la segunda sección.

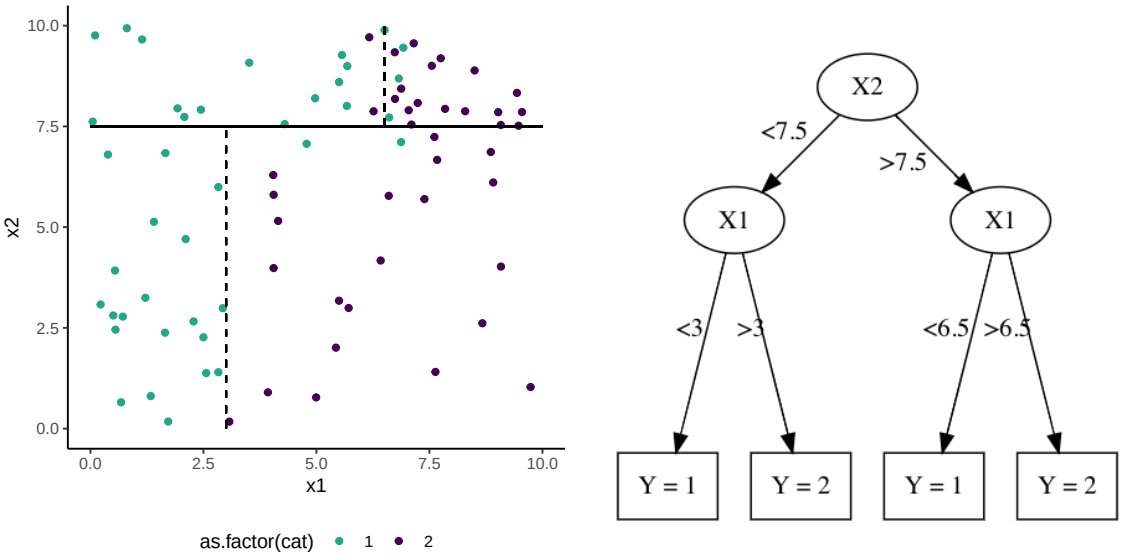


Figure 11: Funcionamiento sobre plano de un árbol de clasificación.

### 4.3.2 Algoritmo CART

La idea básica del algoritmo de Arboles de Clasificación y Regresión (CART) es seleccionar la mejor división de cada nodo para que los nodos hijos sean los mas "puros" posibles. La cantidad de posibles combinaciones de arboles crece exponencialmente a medida que crecen las variables usadas para formar el modelo, ya que para una variable categórica con $I$ categorías, existen $2^{I-1}-1$ posibles divisiones para dicha variable y para una variable continua con $K$ valores, existen $K-1$ posibles divisiones para esa variable, es por ello que el algoritmo tiene que ser lo más eficiente posible para que sea computacionalmente posible su calculo.

El proceso de formación del árbol es el siguiente:

1. **Se busca la mejor división para cada posible predictor**: Se prueban todas las posibles divisiones de cada predictor, tanto continuo como categórico, y se selecciona el que maximice el criterio de división. Este será el nodo raíz y se divide en dos nodos hijos.
2. **Buscar la mejor división del nodo anterior**: Una vez fijado el nodo raíz, hay que realizar la búsqueda del mejor nodo, por tanto, con los individuos de cada uno de los nodos, se vuelven a probar todas las posibles combinaciones y se selecciona el que mejor se ajuste al criterio de división.
3. **Parar o repetir**: Si se cumple algún criterio de parada, se finaliza el árbol, en caso contrario, se repite el paso 2 hasta que se cumple alguno de los criterios de parada.

#### 4.3.2.1 Criterio de división

Como se comentaba anteriormente, cuando se calcula la división optima, lo que se busca es que los nodos hijos sean lo mas puros posibles, lo que quiere decir que los individuos que contenga el nodo hijo deben ser mayoritariamente de una sola categoría, un nodo llega al máximo de pureza cuando todos los individuos del nodo pertenecen a una de las categorías de $Y$.

Si tenemos un nodo $t$, su mejor división es la que maximiza el criterio de división, el cual es el descenso en la impureza y viene dado por:

$$\Delta i(s,t) = i(t) - p_L \times (t_L) - p_R \times (t_R)$$

Donde $p_L$ y $p_R$ son las probabilidades de enviar un caso al nodo hijo $t_L$ (izquierdo) y de enviarlo al nodo hijo $t_R$ (derecho) respectivamente.

Cuando se habla de pureza en el algoritmo CART se hace referencia a la cantidad de individuos de cada tipo que hay en un nodo. Si en un nodo todos los individuos son del tipo $i$, ese nodo ha llegado al maximo de su pureza, mientras que si hay 50% de cada clase ($i$ y $j$), ese nodo es muy impuro. Lo ideal es que los nodos sean lo mas puros posibles para facilitar la tarea de clasificación a la hora de dividirse en dos nuevos nodos, cada vez más puros.

Existen diferentes criterios de división, pero en el presente estudio unicamente nos centraremos en el metodo de Gini.

Notación:

$$p(j,t) = \text{ Probabilidad de un caso de la clase j en el nodo t } = \pi(j) \times N_{w,j}(t)$$

$$p(t) = \text{ Probabilidad de un caso en el nodo t } = \sum_j p(j,t)$$

$$p(j \mid t) = \text{ Probabilidad de que un caso sea de la clase j sabiendo que está en el nodo t } = \frac{p(j,t)}{p(j)}$$

Donde:

$$N_{w,f} = \sum_{n \in h(t)} w_n \times f_n \times I(y_n = j)$$

Siendo $I(a = b)$ un indicador que toma valor 1 cuando $a = b$ y 0 en caso contrario, $w_n$ el peso asociado al caso n, $f_n$ el peso de la frecuencia asociada al caso n, $\pi(j)$ la probabilidad a priori de que Y=j y $h(t)$ los individuos usados para entrenar el árbol que caen en el nodo t.

#### 4.3.2.1.1 Variables categóricas:

En el caso de las variables categóricas, según el metodo de Gini, la impureza se mide como:

$$i(t) = \sum_{i,j} = C(i \mid j) \times p(i \mid t) \times p(j \mid t)$$

Donde $C(i \mid j)$ es el coste de clasificar erróneamente un caso de clase $j$ como un caso de clase $i$.

$p_L$ y $p_R$ se estiman de la siguiente forma:

$$p_L = \frac{p(t_L)}{p(t)} \quad \text{y} \quad p_R = \frac{p(t_R)}{p(t)}$$

#### 4.3.2.1.2 Variables Continuas

En el caso de las variables continuas, la impureza se calcula por el metodo de los Mínimos Cuadrados de la siguiente forma:

$$i(t) = \frac{\sum_{n \in h(t)} w_n \times f_n \times (y_n - \overline{y}(t))^2}{\sum_{n \in h(t)} w_n \times f_n}$$

y las probabilidades de enviar un caso a los nodos hijos vienen dadas por:

$$p_L = \frac{N_w \times (t_L)}{N_w \times p(t)} \quad \text{y} \quad p_L = \frac{N_w \times (t_R)}{N_w \times p(t)}$$

Donde:

$$N_w = \sum_{n \in h(t)} w_n \times f_n \quad \text{y} \quad \overline{y}(t) = \frac{\sum_{n \in h(t)} w_n \times f_n \times y_n}{N_w(t)}$$

### 4.3.3 Criterio de parada

En la construcción de arboles de clasificación es necesario fijar uno o varios criterios de parada para que no se sigan dividiendo los nodos de forma indefinida. Algunos de los criterios de parada que se suelen utilizar son los siguientes:

- Si la mejora que se consigue con la división del nodo t es menor a la mejora mínima establecida, no se divide el nodo
- Se puede fijar una cantidad de individuos por nodo mínima, por tanto, si la división de un nodo padre crea dos nodos hijos que tienen una cantidad de individuos menor a la fijada, el nodo no se divide.
- Si todos los individuos de un nodo tienen el mismo valor para la variable utilizada en dicho nodo para todas las categorías $j$ de la variable $Y$, el nodo no se divide.
- Si el nodo llega al máximo de pureza, todos los individuos de dicho nodo pertenecen a la misma categoría $j$ de la variable dependiente $Y$, el nodo no se divide.
- Otros criterios extras fijados por el usuario, como el tamaño máximo del árbol, etc.

### 4.3.4 Poda

Por la propia forma en la que se construyen los arboles de clasificación, es muy común que el árbol obtenido sufra de sobreajuste (del inglés "overfitting").

Cuando se habla de "overfitting" se hace referencia a un modelo ajustado demasiado bien a los datos con los que se ha entrenado, lo que genera muchos errores a la hora de generalizar y clasificar un nuevo conjunto de datos. Se entiende por entrenar un modelo la etapa en la que se pasa por el modelo un conjunto de datos que contiene tantos los valores de las variabes predictoras ($X_i$) como de la variable respuesta ($Y$) para que el modelo "detecte" patrones en los datos que hacen que el output sea uno u otro.

El árbol que se obtiene al final no es el árbol optimo que se busca, es que árbol que mejor se ajusta a los datos con los que se ha entrenado, por tanto, es necesario buscar un árbol optimo, el cual minimice tanto los errores propios del modelo como los errores de clasificación. En el árbol final se minimizan los errores del modelo, pero se disparan los errores de clasificación con un nuevo conjunto de datos.

En el siguiente gráfico (Fig. 12), se muestran los errores de clasificación de un conjunto de datos "train", usado para entrenar el modelo, y otro conjunto "test", usado para probar el modelo. Siendo cada punto un posible árbol, se observa como a medida que crece la complejidad de los arboles generados, los errores de clasificación de "train" decrecen hasta prácticamente 0, mientras que los errores de clasificación de "test" , en cierto momento, a causa de la alta complejidad del árbol, se disparan. El árbol optimo es el que minimiza ambos errores, siendo todo modelo posterior un modelo con overfitting, y todo modelo anterior, un modelo con underfit.

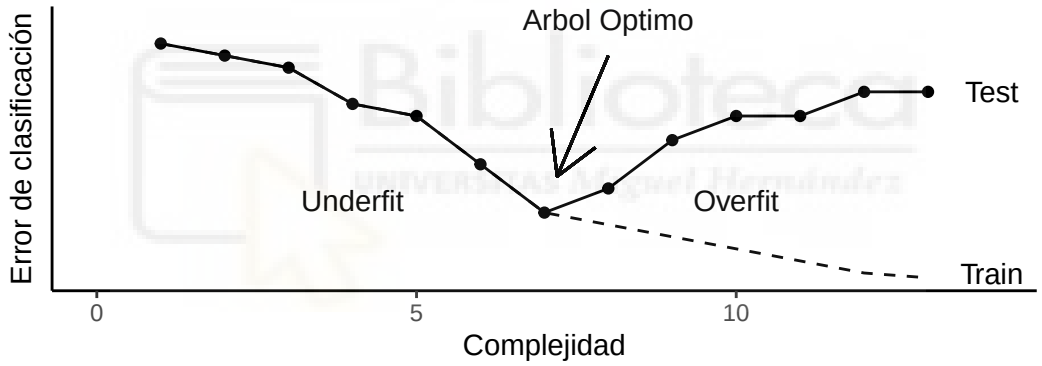


Figure 12: Error de clasificación.

Para solucionar dicho problema, los arboles de clasificación cuentan con una técnica llamada poda, en la que se "cortan" ramas del árbol hasta llegar a un árbol optimo.

Existen dos tipos de poda:

- **Poda a priori**: La poda a priori consiste en fijar una regla de parada mucho mas restrictiva desde un inicio, para evitar generar arboles demasiado complejos que sobreentrenen el modelo. La desventaja de la poda a priori es la dificultad que supone marcar las restricciones de crecimiento del árbol a priori, pudiendo resultar en arboles con "underfitting", no lo suficiente potentes, o con arboles con "overfitting", lo que causaría tener que recurrir de nuevo a la poda.
- **Poda a posteriori**: La poda a posteriori, es la técnica más conocida y usada y consiste en formar un árbol completo, el más complejo posible, para luego podar las diferentes ramas hasta conseguir el árbol optimo. Esta técnica es mucho más potente que la poda a priori y se consiguen mejores resultados, pero, por otro lado, se requiere mayor poder computacional para forma un árbol complejo que realmente no se va a utilizar. En el momento que podas por un nodo intermedio, este pasa a ser un nodo terminal, el cual da como output la categoría $j$ de la variable $Y$ de la que tiene mas casos.

## 4.4 Redes Neuronales

Las redes neuronales son una técnica de Deep learning inspirada en las redes neuronales bilógicas, copiando la estructura básica de una neurona, pero con una funcionalidad y forma de procesar información totalmente diferente. Una posible definición de red neuronal es la siguiente:

- Modelo computacional/matemático inspirado en el cerebro humano y basado en el procesamiento paralelo de la información por medio de nodos con grandes cantidades de conexiones entre estos para la detección de patrones en los datos.

Algunas de las ventajas de las redes neuronales frente a otras técnicas de Deep learning o machine learning son:

- **Sistema no lineal:** Al poder procesar la información de forma no lineal, permite el procesado de datos más "caóticos".
- **Tolerancia frente a fallos:** Ya que procesa la información de forma paralela entre los diferentes nodos, las redes neuronales son bastante tolerantes frente a los fallos de uno o varios nodos.
- **Adaptabilidad:** Una red neuronal tiene la capacidad de ajustar sus parámetros en tiempo real dependiendo de los inputs que recibe y de sus características, mientras que otros modelos no tienen esta adaptabilidad.

### 4.4.1 Neurona

El elemento básico de toda red neuronal son las neuronas, las cuales tienen 6 elementos básicos:

- Conjunto de inputs $x_j$, los cuales pueden ser tanto los datos de entrada como lo datos ya procesados por otro conjunto de neuronas.
- Conjunto de pesos $w_{ij}$, asociados a cada input $j$ y a cada neurona $i$.
- El umbral $\theta_i$, el cual sirve como indicador de si una neurona se activa o no.
- Sumatorio, el cual agrega todos los parámetros anteriores de la siguiente forma:

$$\sum_{i=1}^{n} w_{ij}x_j - \theta_i$$

- La función de activación, la cual genera el output de la neurona. Existen diferentes tipos de funciones de activación, por ejemplo, si se busca que el output sea binario, se suele usar una función de activación escalonada de la siguiente forma:

$$y_i = \begin{cases} 1 \text{ si } \sum_{i=1}^{n} w_{ij}x_j \geq \theta_i \\ 0 \text{ si } \sum_{i=1}^{n} w_{ij}x_j < \theta_i \end{cases}$$

- El output $y_i$, el cual puede ser la respuesta final de la red neuronal, o puede ser usada como input en otra neurona.

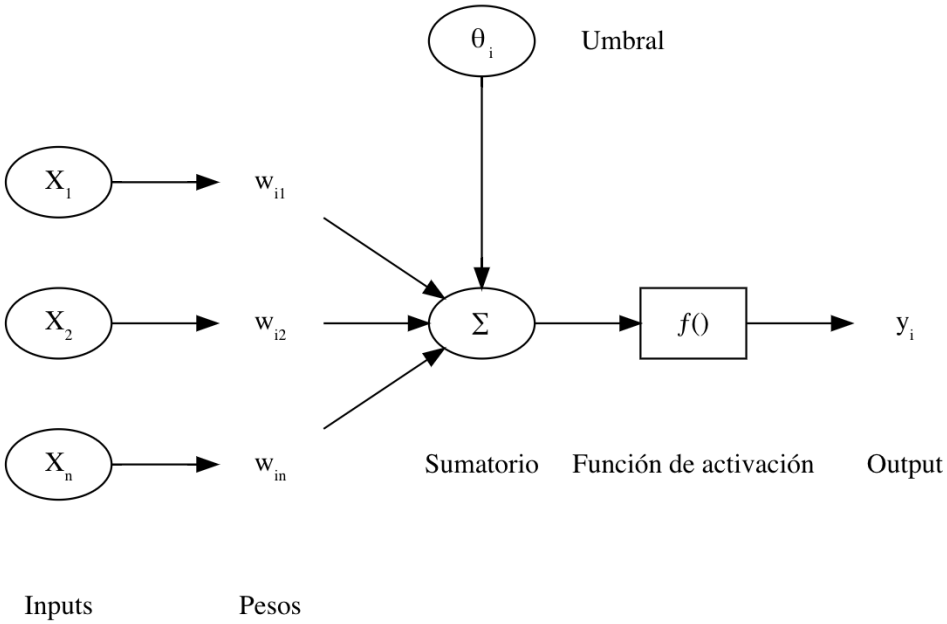


Figure 13: Estructura de una neurona.

Como comentábamos anteriormente, existen diferentes funciones de activación para diferentes casos, y dentro de una red neuronal se pueden utilizar distintos tipos de funciones de activación.

A continuación, observamos algunas de las funciones de activación más comúnmente utilizadas en redes neuronales.

En primer lugar, observamos la función sigmoide, generalmente usada cuando se busca predecir probabilidades, ya que se encuentra entre 0 y 1.

En según lugar observamos la función escalonada, la cual es usada cuando el output esperado es binario, si $x >$ a un limite establecido, entonces $y = 1$, y si $x <$ limite, $y = 0$.

Finalmente, la función ReLU, en la que si $x > 0$, $y = x$, mientras que si $x < 0$, $y = 0$. Resulta muy útil cuando se combina con más neuronas.

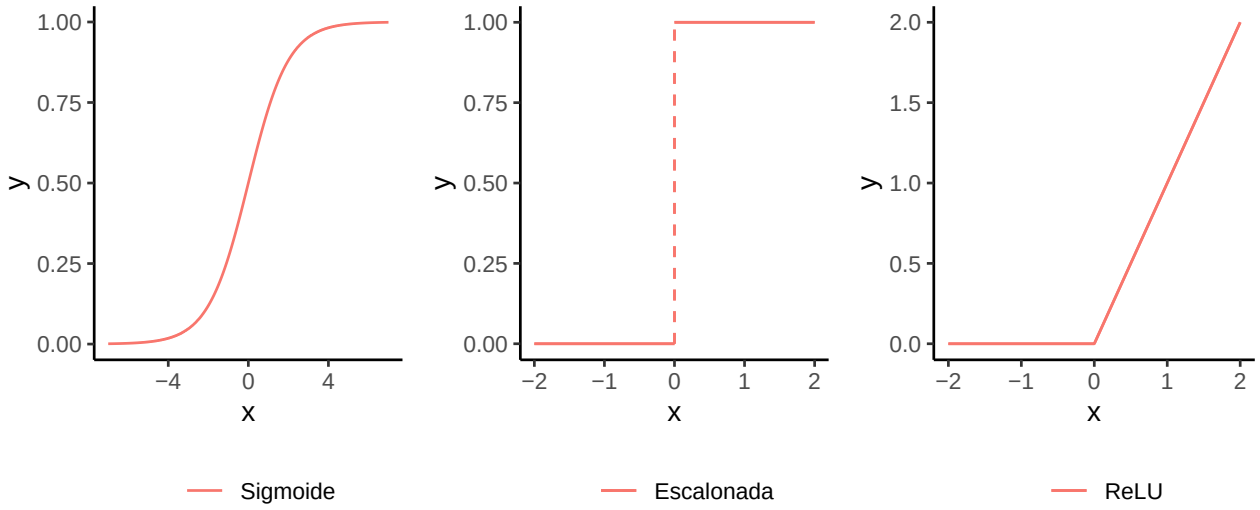


Figure 14: Funciones de activación.

### 4.4.2 Estructura de una red neuronal

Cuando se agrupan y conectan diferentes neuronas de la forma adecuada, se obtiene una red neuronal. La neuronas se suelen agrupar por capas, y la información es unidireccional, lo que quiere decir que la información uncialmente viaja en una dirección, desde la capa de entrada, hacia la capa de salida, aunque existen estructuras más complejas que permiten la circulación de la información en todas las direcciones. Existen tres tipos de capas, la capa de entrada, que es la que recibe la información, las capas ocultas, las cuales se encargan de procesar la información que entra en la red, y la capa de salida, la cual produce el output.

Existen diferentes tipos de estructuras para las redes neuronales, se pueden distinguir según la cantidad de capas, monocapa (1 única capa) o multicapa (2 o más capas), según el tipo de conexiones, no recurrentes (información unidireccional) o recurrentes (información multidireccional), o según la cantidad de conexiones, redes totalmente conectadas (todas las neuronas de una capa están conectadas con todas las neuronas de la capa siguiente) o redes parcialmente conectadas (no todas las neuronas conectadas).

A continuación, se observa un grafo de una red neuronal multicapa, no recurrente y totalmente conectada.

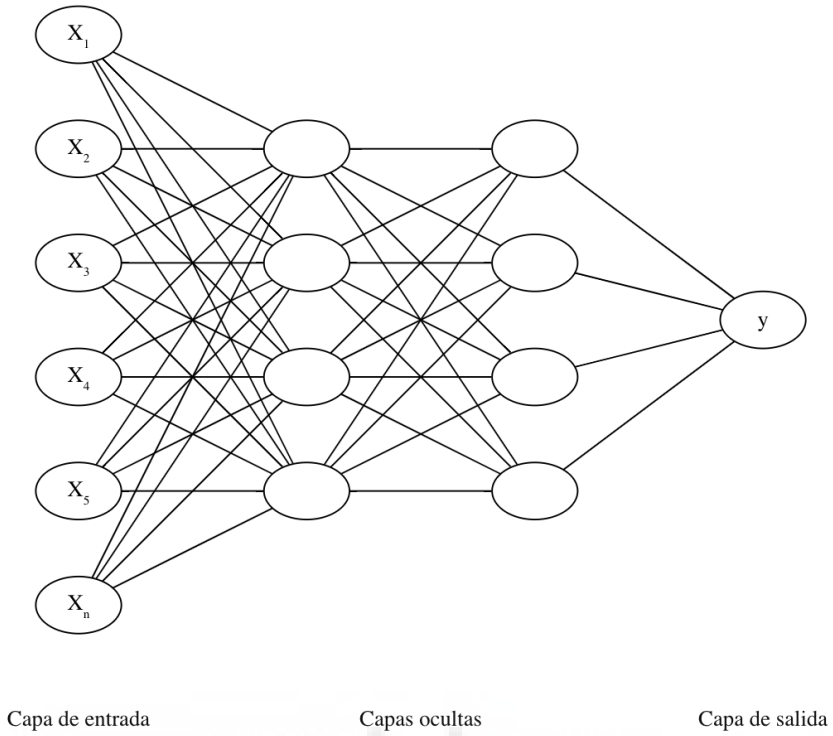


Figure 15: Ejemplo de una posible estructura de una red neuronal.

Toda red neuronal tiene las siguientes características:

- Cada conexión $(i, j)$ entre las neuronas $i$ y $j$ tiene asociado un peso $w_{i,j}$ independiente del resto.
- Cada neurona $i$ tiene asociado un umbral $\theta_i$
- En cada neurona $i$ se define una función que depende de los pesos asociados, del umbral, del estado de cada una de las neuronas $j$ conectadas con la neurona $i$. El valor de la función proporciona el estado de la neurona $i$.

Las estructuras unicapa (con ninguna capa oculta) resultan muy limitante a la hora de, por ejemplo, clasificar datos que no se puedan dividir con un hiperplano, es por eso que está mucho mas extendido el uso de redes neuronales multicapa, con varias capas ocultas, que permiten detectar patrones como más complejos y no lineales.

La selección de la estructura optima para una red neuronal es una de las principales dificultades de las redes neuronales ya que no existe ninguna norma general que indique cuantas capas y neuronas debería tener una red, generalmente se recurre a la prueba y error, aun que existen ciertas técnicas de poda y extensión para las redes neuronales que pueden facilitar en cierta medida dicha tarea.

Actualmente las redes neuronales son una de las técnicas de machine learning más utilizada y con mayores expectativas para el futuro gracias a su gran flexibilidad, precisión y eficiencia, se usan en todo tipo de campos, como puede ser la medicina, la biología, la economía, la optimización de recursos, etc. En los últimos años ha habido un crecimiento exponencial en la cantidad de tareas que las redes neuronales pueden resolver de forma eficiente (detección de complejos patrones, clasificación, análisis de textos, fotos, videos, etc.) y son usadas a diario por grandes empresas e investigadores para resolver algunos de los problemas más complejos.

# 5 Resultados

## 5.1 Descriptivos

La siguiente tabla (Tab. 3) contiene algunas de las variables cuantitativas más importantes para el estudio, sus medianas y la cantidad de información perdida (NAs) en función de si el paciente es caso (1) o control (0). Se comparan las medianas de cada variable según si son caso o control con el test de Kruskal-Wallis para concluir si existen o no diferencias significativas entre estas. Se encuentran diferencias significativas (p-valor $< 0{,}05$) en la mayoría de las variables de la tabla, exceptuando el valor de CD8 inicial, la carga viral inicial y los biomarcadores il8, c163 y cd40. En Carga viral inicial, las diferencias son prácticamente significativas, ya que el p-valor es casi 0.05. Las que más llaman la atención son el valor inicial de CD4, con un valor notablemente mayor en los pacientes control en comparación con los casos, y la ratio CD4/CD8, la cual, cuanto menor es su valor, más grave suele ser el estado del paciente.

| | 0 (N=122) | 1 (N=56) | Total (N=178) | p value |
|---|---|---|---|---|
| **Edad** | | | | 0.023 |
| Median | 44.93 | 48.53 | 46.01 | |
| **CD4 inicial** | | | | < 0.001 |
| Median | 374.00 | 119.00 | 298.50 | |
| **CD8 inicial** | | | | 0.422 |
| Median | 847.00 | 920.00 | 866.00 | |
| NAs | 34 | 25 | 59 | |
| **Carga viral inicial** | | | | 0.056 |
| Median | 60354.50 | 103600.00 | 77206.50 | |
| **CD4/CD8 inicial** | | | | < 0.001 |
| Median | 0.44 | 0.14 | 0.34 | |
| NAs | 34 | 25 | 59 | |
| **pcr** | | | | 0.031 |
| Median | 4.08 | 14.82 | 4.89 | |
| **tnf** | | | | 0.026 |
| Median | 1.01 | 6.58 | 2.13 | |
| **il6** | | | | < 0.001 |
| Median | 0.83 | 4.24 | 1.40 | |
| **il8** | | | | 0.609 |
| Median | 16.36 | 0.80 | 0.80 | |
| NAs | 120 | 43 | 163 | |
| **neop** | | | | < 0.001 |
| Median | 12.55 | 20.17 | 15.47 | |
| **dd** | | | | < 0.001 |
| Median | 3.35 | 79.69 | 13.06 | |
| **isopros** | | | | < 0.001 |
| Median | 20.73 | 46.52 | 25.77 | |
| **mda** | | | | 0.009 |
| Median | 9.29 | 12.39 | 9.65 | |
| **cd14** | | | | 0.020 |
| Median | 1824275.00 | 2320616.00 | 1956917.00 | |
| **c163** | | | | 0.807 |
| Median | 153203.50 | 173833.50 | 155002.00 | |
| **cd40** | | | | 0.605 |
| Median | 102.62 | 106.43 | 104.06 | |
| NAs | 2 | 0 | 2 | |

Tabla 3: Variables cuantitativas

En la siguiente tabla (Tab. 4) se puede observar las principales variables categóricas usadas en el estudio, usando la prueba de chi cuadrado ($X^2$) para comprobar si existen diferencias significativas entre las diferentes categorías. En las tres variables se encuentran diferencias significativas. La mayoría de los controles (80,5%) no desarrolla SIDA antes o durante el estudio, mientras que la mayoría de los casos (61,5%) si que desarrolla SIDA, sucede lo mismo con la variable muerte, la mayoría de los controles (97,6%) no muere durante el estudio, mientras que la mayoría de los casos (92,3%), si que mueren. Por otro lado, la variable ENOS, en principio, si que muestra diferencias significativas entre dos o más categorías.

| | 0 (N=41) | 1 (N=26) | Total (N=67) | p value |
|---|---|---|---|---|
| **ENOS** | | | | 0.014 |
| bone | 5 (12.2%) | 1 (3.8%) | 6 (9.0%) | |
| cardiovascular | 2 (4.9%) | 2 (7.7%) | 4 (6.0%) | |
| liver | 11 (26.8%) | 5 (19.2%) | 16 (23.9%) | |
| metabolic | 4 (9.8%) | 1 (3.8%) | 5 (7.5%) | |
| neoplastic | 3 (7.3%) | 11 (42.3%) | 14 (20.9%) | |
| neuropsychiatric | 11 (26.8%) | 1 (3.8%) | 12 (17.9%) | |
| others | 2 (4.9%) | 1 (3.8%) | 3 (4.5%) | |
| renal | 3 (7.3%) | 4 (15.4%) | 7 (10.4%) | |
| **SIDA** | | | | $< 0.001$ |
| Sí | 33 (80.5%) | 10 (38.5%) | 43 (64.2%) | |
| No | 8 (19.5%) | 16 (61.5%) | 24 (35.8%) | |
| **Muerte** | | | | $< 0.001$ |
| Sí | 40 (97.6%) | 2 (7.7%) | 42 (62.7%) | |
| No | 1 (2.4%) | 24 (92.3%) | 25 (37.3%) | |

Tabla 4: Variables cualitativas.

A continuación, se realiza un estudio gráfico de las diferentes variables para observar diferentes patrones y relaciones entre estas. A lo largo de la introducción se hace referencia a diferentes relaciones ya conocidas entre algunas de las variables disponibles en el presente estudio, por tanto, a continuación, se trata de verificar que dichas relaciones se encuentran en la base de datos de la que se dispone.

En primer lugar, en el siguiente gráfico (Fig. 16), se pueden observar los box plots de todos los biomarcadores de los que se dispone, estudiando su comportamiento cuando el valor de CD4 es mayor o menor a 200, el cual es un valor establecido por norma general, y por debajo del cual se dice que el sistema inmunológico es muy débil. En este gráfico no es relevante la comparación entre biomarcadores, ya que cada biomarcador se mueve en una escala diferente, por tanto, lo que se muestra es el logaritmo de los biomarcadores, para facilitar la interpretación.

Como se puede observar, algunos de los biomarcadores no se ven muy afectados por el hecho de que el valor de CD4 sea mayor o menor a 200, como por ejemplo cd14, tnf o il8, pero, por otro lado, observamos que algunos biomarcadores, como icam, vcam o isopros, entre otros, aumentan su valor cuando CD4<200, mientras que el biomarcador V39 baja su valor cuando CD4<200. Por tanto, si que se encuentra cierta relación entre ambas variables y pueden ser buenas variables explicativas para los posteriores modelos.

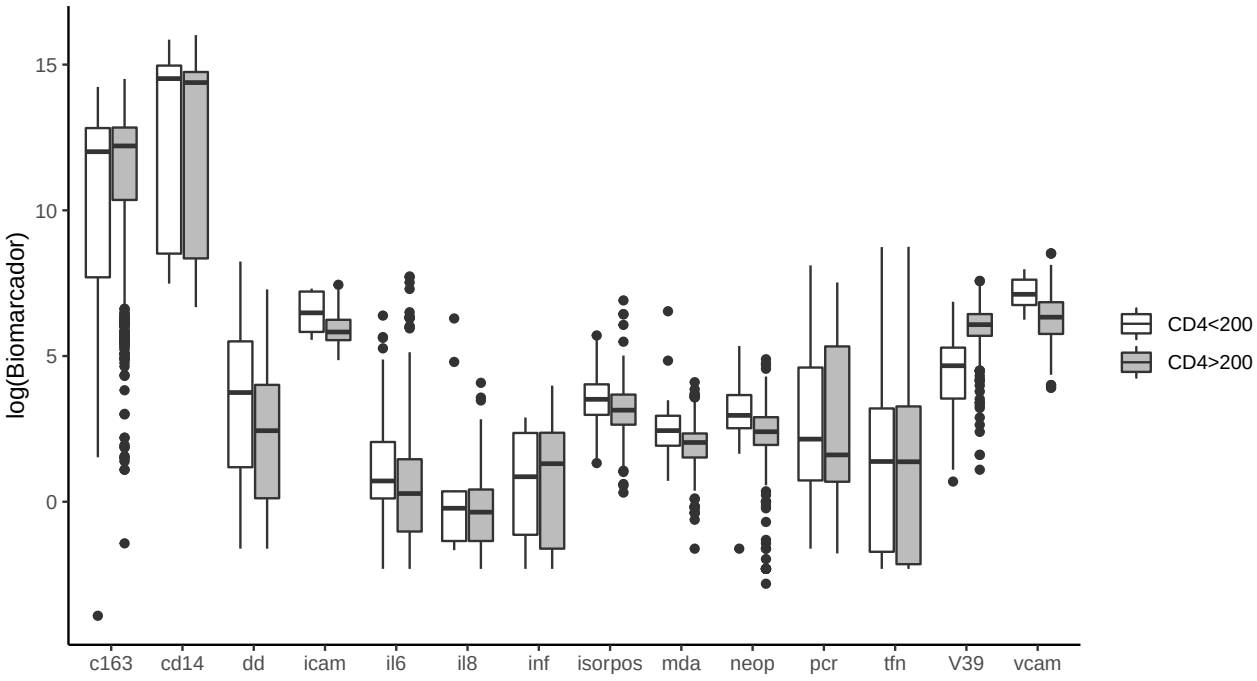


Figure 16: Distribución de los biomarcadores en función de si CD4 es mayor o menor a 200.

A continuación (Fig. 17), se muestra otra serie de boxplots, en este caso, como se distribuye el valor de CD4 inicial en función del evento ENOS. La línea roja representa el valor 200 en el CD4 inicial, que como se ha comentado anteriormente, es un valor por debajo del cual se considera que el sistema inmunológico está en situación critica.

En principio, si que se observan diferencias entre la distribución del valor de CD4 inicial en función del evento ENOS. Llama la atención que los eventos cardiovasculares y los eventos renales son más frecuentes cuando el valor de CD4 baja de 200, mientras que otros eventos, como los neuropsiquiatricos no tienen mayor relación con el nivel de CD4, ya que los individuos con eventos de dicho tipo tienen valores de CD4 en rangos "normales". En cuanto a los individuos que no sufren ningún tipo de evento ENOS, únicamente el 25-30% de los individuos tienen un valor de CD4 igual a 200 o inferior. En resumen, ciertos tipos de eventos parecen estar afectados por el valor de CD4 inicial, por tanto, estas variables pueden ser de utilidad a la hora de formular los modelos.

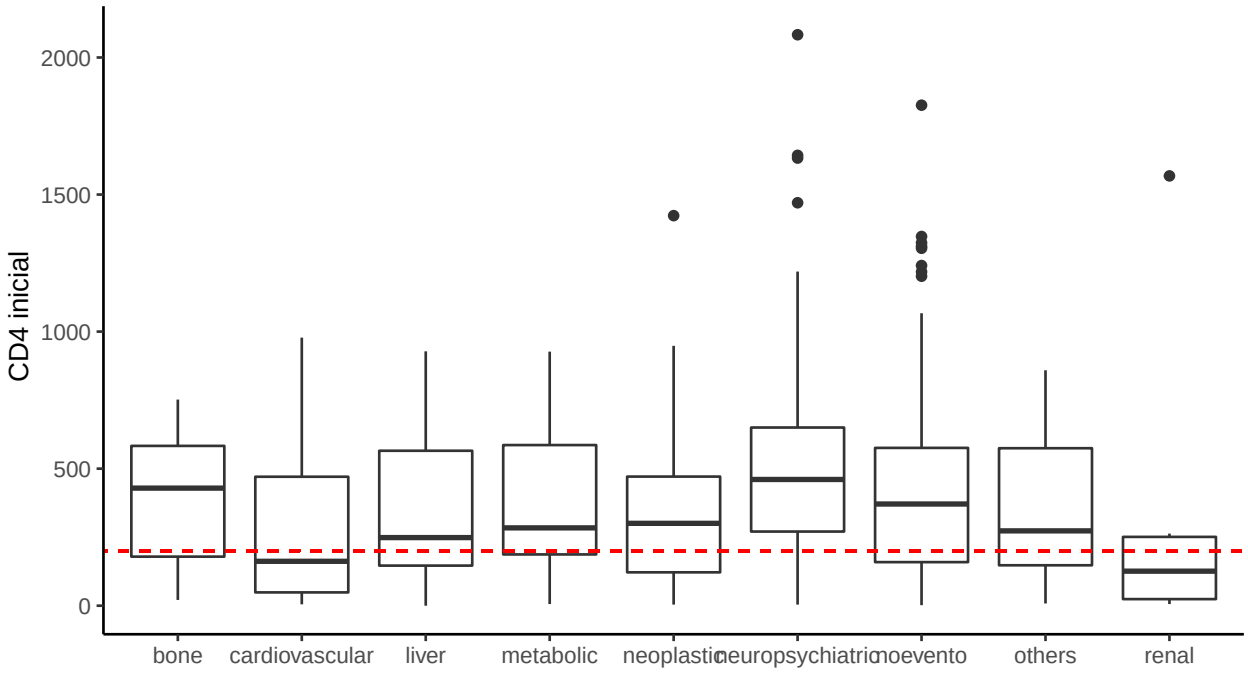


Figure 17: Distribución del CD4 inicial en función del evento sufrido.

Anteriormente se ha estudiado como se comporta el valor de CD4 con el que el individuo entra en el estudio, pero también resulta de interés evaluar como evoluciona el valor de CD4 a lo largo del estudio en función de ciertos parámetros.

Como se comentaba en la presentación de la base de datos, se cuenta con una medición del valor de CD4 de cada individuo en tres momentos distintos, al inicio del estudio, en algún momento cercano al evento y al final del estudio (ultima muestra disponible del individuo).

A continuación (Fig. 18), observamos 3 gráficos donde se representa la evolución del valor de CD4 a lo largo del tiempo en función de ciertos parámetros. Cada punto representa la mediana del valor de CD4 en un momento concreto según la categoría a la que pertenezca, mientras que las barras representan los intervalos de confianza de dicha mediana.

En el primer gráfico, se estudia la evolución en función de si el paciente ha recibido tratamiento contra el VIH o no. Como se puede observar, los pacientes que reciben tratamiento consiguen aumentar en gran medida su conteo de CD4 a lo largo del tiempo, mientras que en los que no reciben tratamiento, el conteo cada vez es menor, lo que supone un mayor riesgo para el paciente y la destrucción del sistema inmunológico. Los intervalos de confianza son mayores en los pacientes que no reciben tratamiento debido a la menor cantidad de pacientes que no reciben, en comparación con los que, si que reciben, los cuales son la mayoría.

En el segundo gráfico se estudia la diferencia en la evolución de los pacientes que mueren y de los que no mueren. Se puede observar que los pacientes que viven tienen una evolución positiva, aumentando el conteo de CD4 a lo largo del tiempo, mientras que los individuos que mueren, por lo general comienzan con un valor de CD4 bastante bajo (<200) y el aumento en el conteo de CD4 es bastante bajo, los niveles se mantienen bajos a lo largo del tiempo, lo que causa un mayor riesgo de muerte a causa del débil sistema inmunitario.

Finalmente, el tercero se compara la evolución en función de si el paciente comienza con niveles de CD4 inferiores o superiores a 200. Se puede observar que los que comienzan con valores muy bajos, por lo general, no suelen mejorar demasiado y se mantienen constantes por debajo de 200, lo que indica, como ya se ha cometido, un sistema inmunológico muy débil y puede llevar a enfermedades muy graves o a la muerte.

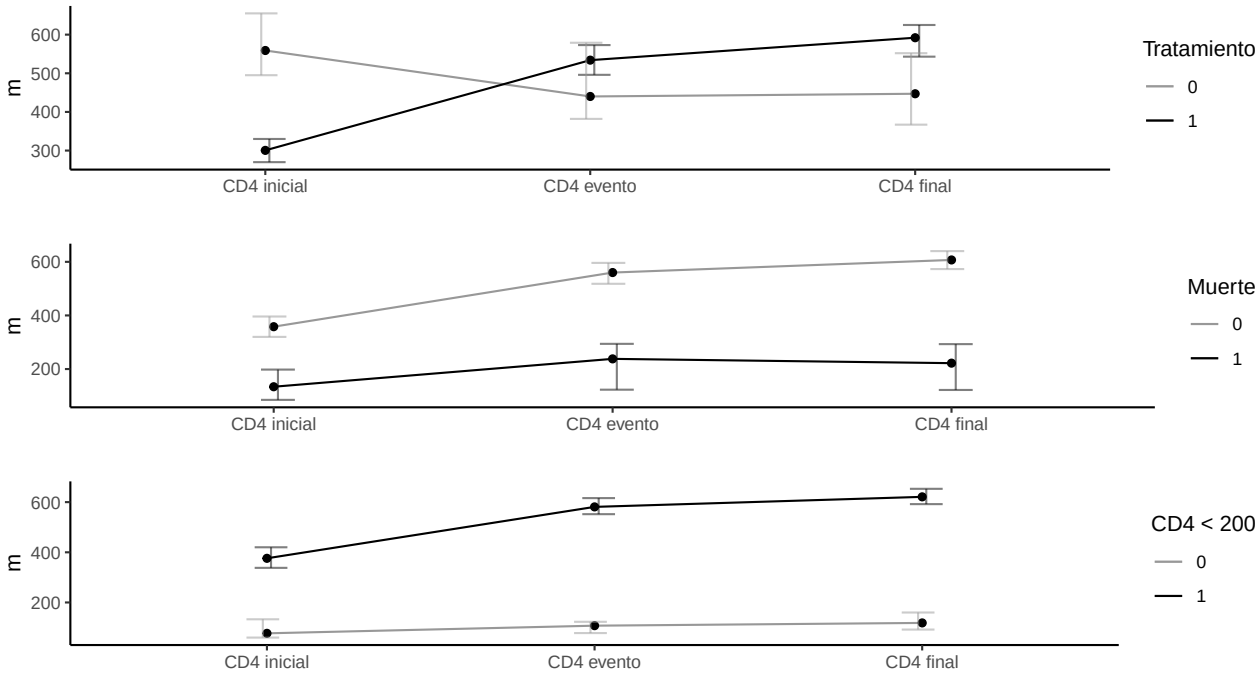


Figure 18: Evolución del CD4 en fucnión de diversos parametros.

Como complemento a la información anterior sobre la evolución del valor de CD4 a lo largo del estudio, observamos el siguiente gráfico (Fig. 19), en el cual se representa la evolución del CD4 en función del tipo de evento ENOS.

En primer lugar, llama la atención al gran tamaño de los intervalos de confianza, debido a que el tamaño muestra de cada tipo de evento es pequeño. En segundo lugar, parece que, por lo general, en todos los tipos de eventos, después del evento, el valor de CD4 se estabiliza, pero no se dispone de la información de lo que sucede entre el evento y la ultima medición, por tanto, estamos observando una estimación de lo que realmente sucede. El único tipo de evento que tiene un comportamiento distinto es el renal, en el cual baja el conteo de CD4 después del evento.

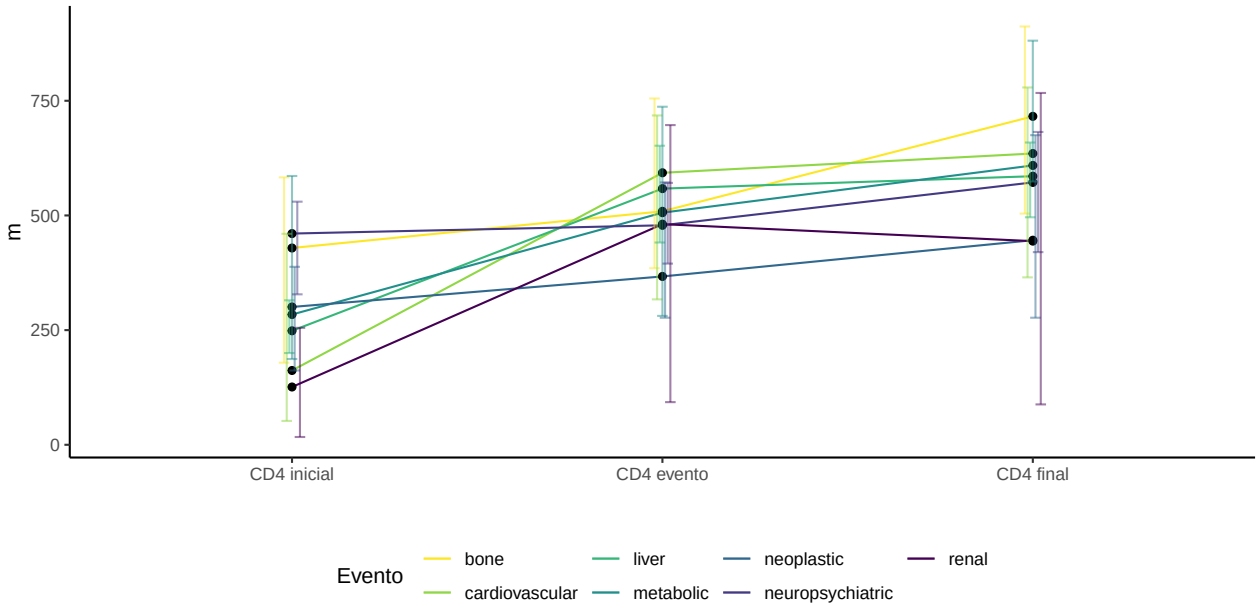


Figure 19: Evolución del CD4 en función del evento sufrido.

Continuamos estudiando el comportamiento de los diferentes tipos de eventos, esta vez en función de la edad. Para ello se han dividido a los individuos en 4 grupos de edad, menores de 25 años, entre 26 y 45 años, entre 46 y 65 años y, finalmente, entre 66 y 75 años. Debido a que en el estudio no se cuenta con una base de datos homogénea en la que haya la misma cantidad de pacientes sin eventos como de pacientes con algún evento, el siguiente gráfico se ha realizado según la frecuencia relativa de cada categoría para mejorar la comprensión. En primer lugar (Fig. 20), se observa como existe una clara relación entre edad y los pacientes sin ningún tipo de evento, a medida que aumenta la edad, es más raro observar un paciente sin evento, en el intervalo [46-65], el porcentaje de pacientes sin evento ya es menor al 50%, mientras que el resto de los eventos cada vez son más frecuentes. En segundo lugar, los eventos neuropsiquiatricos parecen no estar muy relacionados con la edad, siendo más o menos similares en todos los intervalos, aunque un tanto menor en el intervalo [46-65], a causa de que en dicho intervalo los eventos neoplásicos cobran importancia. Finalmente, llama la atención como tanto los eventos de hueso como los cardiovasculares cobran importancia según avanza la edad del paciente, siendo dos de los más frecuentes en el intervalo [66-75].

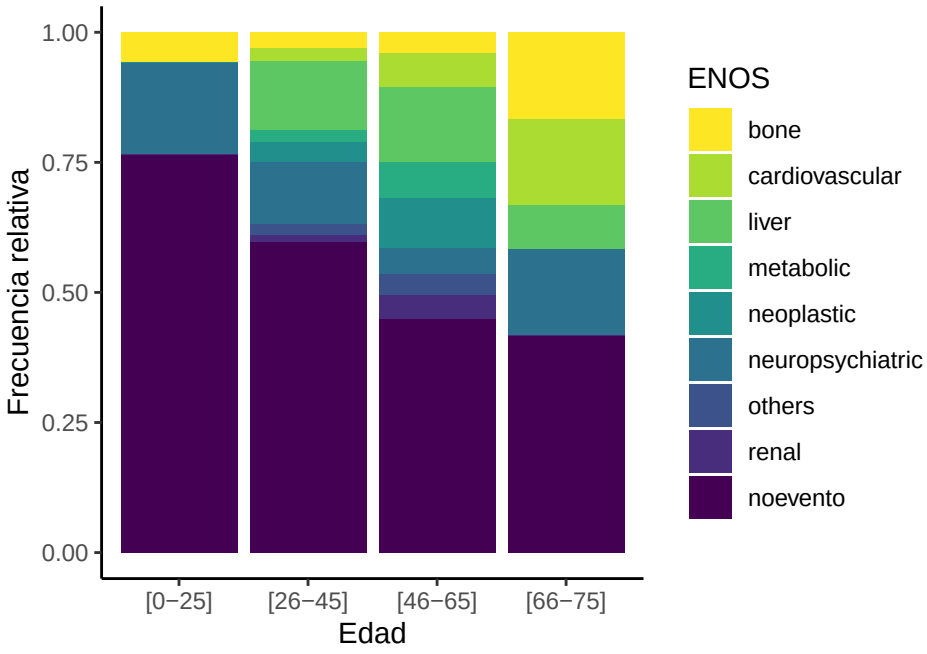


Figure 20: Frecuencia relativa de eventos en función del rango de edad.

Como se comentaba en la introducción, la infección de VHC y VHB son dos de las grandes causas de eventos ENOS, principalmente, el VHC está directamente relacionado con los eventos de hígado. El 27% de los infectados con VIH, también están infectado por el VHC. Se recomiendo tratar los riesgos del VHC para reducir la probabilidad de eventos ENOS. En la siguiente figura (Fig. 21) se observan las diferencias entre eventos según si los pacientes están o no infectados por el VHC. Los pacientes negativos en VHC sufren muchos menos eventos ENOS (~ 40%) que los pacientes positivos en VHC (~ 75%). Llama la atención especialmente los eventos ENOS relacionados con el hígado, los cuales son la mayoría de los eventos ENOS que sufren los pacientes positivos de VHC, mientras que, en los pacientes negativos, son prácticamente insignificantes. Esto se debe que el una de la principales consecuencias de infectare con VHC es desarrollar una inflamación de hígado. De la misma forma que en la figura anterior, se hace uso de la frecuencia relativa para facilitar la interpretación.

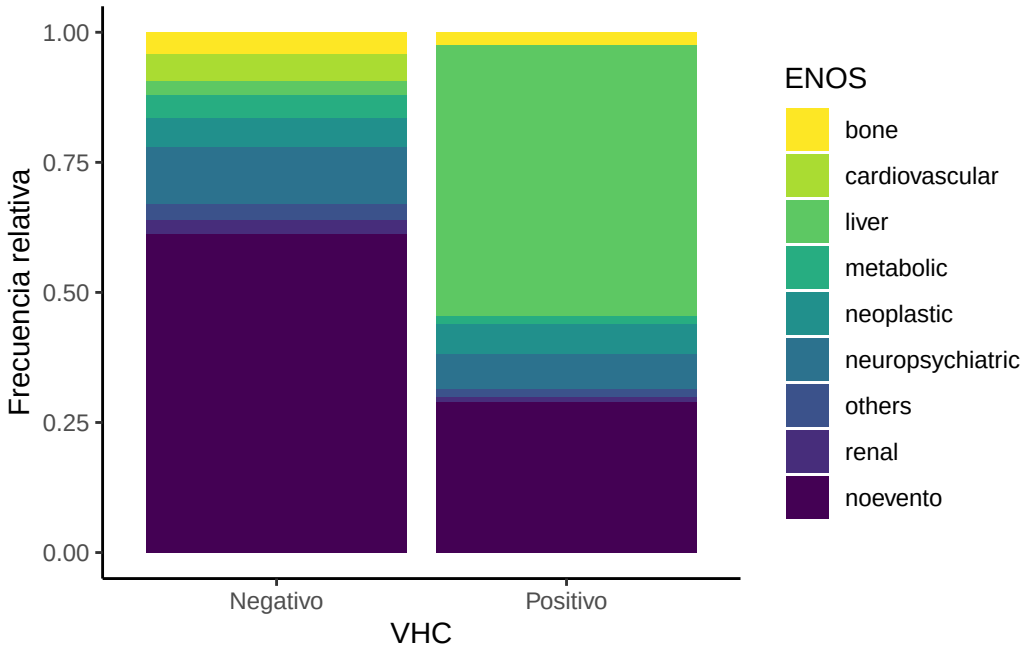


Figure 21: Frecuencia relativa de eventos en función de VHC.

## 5.2 Modelos

En siguiente apartado se muestra la aplicación de los diferentes modelos explicados en el apartado 4 a la base de datos disponible para el presente estudio.

Para el ajuste de los modelos se utiliza el 90% de los datos disponibles seleccionados de forma aleatoria para que no existan sesgos en los resultados y se trabaje con una submuestra de datos representativa de la base original. El 10% de los datos restantes se utilizará para evaluar la potencia de los diferentes modelos, comparando los outputs que den los modelos con los valores reales, principalmente para obtener modelos sin "overfitting".

### 5.2.1 Regresión logística

Para el ajuste de modelos de regresiones logísticas se tiene en cuenta principalmente dos factores:

- **La significación de las variables**: La significación de una variable dentro de un modelo indica si dicha variable tiene una influencia real sobre los resultados del modelo, por tanto, las variables que no sean significativas, generalmente, son excluidas.
- **La potencia predictora**: Una vez formulado el modelo, se evaluará, con el 10% de los datos reservados, la potencia o fiabilidad del propio modelo a la hora de predecir los valores de Y con un nuevo data set. Si estos no son buenos, se debe reformular el modelo para obtener los mejores resultados posibles. Principalmente nos centramos en la precisión o "accuracy" y en el valor de Kappa, el cual es una medida de concordancia y compara la precisión esperada con la precisión observada, se considera concordancia moderada un valor de Kappa entre 50% y 70%, si es mayor del 70%, la concordancia es elevada.

Para la comparación entre dos modelos que tengan la misma variable dependiente (Y), se utiliza el Criterio de Información de Akaike (AIC), el cual es una medida de calidad relativa y a la hora de comparar 2 o más modelos, a menor AIC, mejor es el modelo.

Para facilitar la interpretación de los parámetros de los modelos ($\beta_i$), estos serán transformados para obtener los odds ratios y poder observar como afecta cada variable a las probabilidades del modelo.

#### 5.2.1.1 Evento sí o no.

En primer lugar, se trata de encontrar relación entre ciertas variables y que el paciente sufra o no un evento ENOS del tipo que sea. Las variables que resulten significativas en los próximos modelos explicarán cuales son las posibles causas de que un paciente sufra un evento, y en que medida afecta cada variable a dicho resultado. La cantidad de pacientes en el estudio que han sufrido un evento ENOS (304) es similar a la cantidad de pacientes que no lo han sufrido (250).

En primer lugar, se han realizado destinas pruebas con las variables CD4, CD8, CD4/CD8 y Carga Viral, pero por falta de significación dentro del modelo, han sido eliminadas. Tras extensas pruebas, el mejor modelo que se ha obtenido es el Modelo 1, el cual contiene la edad, si es positivo o negativo en VHC, si recibe o no tratamiento, el biomarcador neop, la variable CD4200, la cual indica si el valor de CD4 inicial del paciente es mayor o menor a 200, y la interacción estas dos variables.

Para poder comparar distintas técnicas, comparamos con el modelo 2, el contiene las mismas variables que un buen modelo obtenido con Arboles de Clasificación, el cual veremos posteriormente.

| | Modelo 1 | Modelo 2 |
|---|---|---|
| Intercept | 5.34* | 1.97 |
| | [4.21; 6.47] | [−2.60; 6.54] |
| Edad | 0.98* | 0.94* |
| | [0.96; 0.99] | [0.89; 1.00] |
| VHC positivo | 0.33* | 0.55 |
| | [−0.14; 0.79] | [−0.66; 1.76] |
| Tratamiento | 0.44* | |
| | [−0.12; 0.99] | |
| Neop | 1.01 | |
| | [1.00; 1.02] | |
| CD4 > 200 | 2.01* | 7.24* |
| | [1.28; 2.74] | [4.57; 9.92] |
| Neop x CD4 > 200 | 0.98* | |
| | [0.96; 1.00] | |
| V39 | | 1.00 |
| | | [1.00; 1.00] |
| RNA inicial | | 1.00 |
| | | [1.00; 1.00] |
| CD4 inicial | | 1.00 |
| | | [1.00; 1.00] |
| Bnp | | 1.00 |
| | | [1.00; 1.00] |
| Isopros | | 1.00 |
| | | [0.97; 1.03] |
| CD14 | | 1.00 |
| | | [1.00; 1.00] |
| AIC | 632.10 | 118.24 |
| BIC | 661.42 | 142.67 |
| Log Likelihood | -309.05 | -49.12 |
| Deviance | 618.10 | 98.24 |
| Num. obs. | 487 | 85 |

* 1 outside the confidence interval

Tabla 5: Modelos.

Como se puede observar, el modelo 2, con el cual se obtienen buenos resultados en otra técnica, con la regresión logística no se obtienen apenas variables significativas

Si evaluamos su capacidad clasificadora, obtenemos una precisión y un valor de Kappa muy bajos, lo cual es esperado observando que la mayoría de las variables no son significativas.

| Predicciones | Referencia | |
|---|---|---|
| | No | Si |
| No | 4 | 4 |
| Si | 1 | 2 |

| | Valor |
|---|---|
| Accuracy | 0.5454545 |
| Kappa | 0.1269841 |

Tabla 6: Matriz de confusión y metricas.

En cuanto al modelo 1, todas sus variables son significativas y se obtienen resultados ligeramente mejores cuando se evalúa su capacidad clasificadora. Una precisión del 57,1% y un valor de Kappa del 20%. Se observa que clasifica bastante bien a los individuos que no sufren evento, pero, posiblemente a causa de que dentro de los pacientes que sufren un evento, existen diferentes tipos de eventos, con diferentes características, el modelo no es muy bueno clasificando a los individuos que si que sufren un evento.

| Predicciones | Referencia | |
|---|---|---|
| | No | Si |
| No | 21 | 21 |
| Si | 3 | 11 |

| | Valor |
|---|---|
| Accuracy | 0.5714286 |
| Kappa | 0.2000000 |

Tabla 7: Matriz de confusión y metricas.

A pesar de ello, se pueden extraer ciertas conclusiones de los Odds ratio del modelo 1. Por cada año extra que tenga el paciente, tiene un 2% menos de probabilidad de sufrir un evento, por otro lado, si es positivo en VHC, tiene un 67% menos de probabilidad de sufrir un evento, de la misma forma, si el paciente recibe tratamiento contra el VIH, tiene un 56% menos de probabilidad de sufrir un evento. Finalmente, las variables neop y CD4200 no resultan significativas, pero su interacción si, por tanto, si el conteo de CD4 del paciente es igual o menor a 200, por cada unidad extra de neop, su probabilidad de sufrir un evento disminuye en un 2%.

Como se puede observar, en las conclusiones de el modelo se hacen ciertas afirmaciones que contradicen los que se ha observado en el apartado de descriptivos o que directamente son anti intuitivas, esto posiblemente se debe a la falta de una o más variables confesoras que hace que los parámetros del modelo se ajusten sin tener en cuenta la información faltante.

#### 5.2.1.2 CD4 $<$ 200

En los siguientes modelos se estudia la relación entre una serie de variables explicativas y la variable CD4200, la cual indica si el valor de CD4 del paciente es menor o mayor a 200.

Es posible que existan problemas a la hora de ajustar el modelo ya que existen muchos más pacientes con CD4 $> 200$ (481) que pacientes con CD4 $< 200$ (76)

Tras realizar diferentes pruebas dos posibles modelos. El primero cuenta con 3 biomarcadores (neop, mda y cd14) y con la variables cualitativa cc.ev que indica si el paciente es caso o control. El segundo modelo es igual que el anterior, exceptuando que se susittuye la el biomarcador cd14 por la edad del paciente. En ambos modelos todas las variables usadas son significativas.

| | Modelo 1 | Modelo 2 |
|---|---|---|
| Intercept | 13.55* | 4.31* |
| | [13.10; 14.00] | [4.09; 4.53] |
| Neop | 0.52* | 0.71* |
| | [0.21; 0.82] | [0.54; 0.88] |
| Mda | 0.25 | 0.45 |
| | [−0.78; 1.27] | [−0.11; 1.01] |
| CD14 | 0.76 | |
| | [0.50; 1.01] | |
| cc.ve | 0.22* | 0.46* |
| | [−0.36; 0.80] | [0.15; 0.76] |
| Edad | | 0.84* |
| | | [0.68; 0.99] |
| AIC | 336.12 | 338.88 |
| BIC | 357.17 | 359.94 |
| Log Likelihood | -163.06 | -164.44 |
| Deviance | 326.12 | 328.88 |
| Num. obs. | 498 | 499 |

* 1 outside the confidence interval



Tabla 8: Modelos.

Como se puede observar en la tabla 8, el AIC del primer modelo es menor, por tanto, en principio es mejor que el segundo modelo, pero es necesario evaluar la potencia de cada modelo para obtener resultados concluyentes.

Ambos modelos ofrecen resultados más o menso similares, una precisión bastante alta (83,9%). Si nos fijamos en como clasifican a los individuos observamos que ambos modelos clasifican casi a la perfección a los individuos que tienen un valor de CD4 $> 200$, pero bastante mal a los individuos que tienen un valor de CD4 $< 200$, debido a que, como comentábamos anteriormente, la base con la que se ajustan los modelos no cuenta con la misma cantidad de individuos de cada tipo, lo que crea un sesgo en los modelo, y posiblemente por esto, los valores de Kappa son del 23%, muy bajos.

Por la forma en la que clasifica a los individuos, parece que el modelo 1 es ligeramente mejor, ya que pierde un poco en cuanto a cesibilidad, pero mejora ligeramente en cuanto a especificidad.

| Predicciones | Referencia | |
|---|---|---|
| | <200 | >200 |
| <200 | 2 | 2 |
| >200 | 7 | 45 |

| | Valor |
|---|---|
| Accuracy | 0.8392857 |
| Kappa | 0.2317073 |

Tabla 9: Matriz de confusión y metricas.

| Predicciones | Referencia | |
|---|---|---|
| | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 8 | 47 |

| | Valor |
|---|---|
| Accuracy | 0.8571429 |
| Kappa | 0.1734317 |

Tabla 10: Matriz de confusión y metricas.

Por tanto, centrándonos únicamente en le modelo 1, se puede concluir que, a mayores valores de los biomarcadores incluidos en el modelo, la probabilidad de que el valor de CD4 del paciente sea igual o menor a 200, disminuye.

#### 5.2.1.3 Muerte

El objetivo de los siguientes modelo es averiguar cuales son las variables más relevantes para que un paciente muera o no, y en que medida afectan. Para ellos se formulan los 3 siguientes modelos. El primer modelo contiene la edad del paciente, su valor inicial de CD4, una variable binaria que indica si el valor de CD4 del paciente es menor o mayor a 200, una variable que indica si el paciente tiene SIDA y la interacción entre el valor inicial de CD4 y la variable que indica si es mayor o menor a 200. En el segundo modelo se añaden las interacciones entre la variable que indica si el paciente tiene SIDA con el resto de variable continuas. Finalmente, en el tercer modelo se añade una variable que indica si el paciente recibe tratamiento contra el VIH y las interacciones de esta con el resto de las variables continuas.

| | Modelo 1 | Modelo 2 | Modelo 3 |
|---|---|---|---|
| Intercept | 0.251 | 0.236 | 0.087 |
| | [−0.617; 1.119] | [−0.788; 1.260] | [−1.704; 1.878] |
| Edad | 1.029* | 1.036* | 1.064* |
| | [1.013; 1.044] | [1.017; 1.055] | [1.029; 1.100] |
| CD4 inicial | 0.997* | 0.996* | 0.996* |
| | [0.994; 0.999] | [0.993; 0.999] | [0.993; 0.999] |
| CD4 > 200 | 0.221* | 0.192* | 0.201* |
| | [−0.304; 0.747] | [−0.355; 0.738] | [−0.350; 0.753] |
| SIDA | 1.730* | 2.777* | 2.209 |
| | [1.378; 2.083] | [1.129; 4.425] | [0.493; 3.925] |
| CD4 inicial x CD4 > 200 | 1.003* | 1.004* | 1.004* |
| | [1.001; 1.006] | [1.001; 1.007] | [1.002; 1.007] |
| Edad x SIDA | | 0.983 | 0.987 |
| | | [0.950; 1.016] | [0.953; 1.022] |
| CD4 inicial x SIDA | | 1.001 | 1.001 |
| | | [1.000; 1.002] | [1.000; 1.002] |
| Tratamiento | | | 4.064* |
| | | | [1.994; 6.134] |
| Edad x Tratamiento | | | 0.964 |
| | | | [0.924; 1.005] |
| CD4 inicial x Tratamiento | | | 1.000 |
| | | | [0.999; 1.001] |
| AIC | 324.850 | 323.581 | 324.704 |
| BIC | 350.149 | 357.314 | 371.086 |
| Log Likelihood | -156.425 | -153.791 | -151.352 |
| Deviance | 312.850 | 307.581 | 302.704 |
| Num. obs. | 501 | 501 | 501 |

* 1 outside the confidence interval

Tabla 11: Modelos.

En estos tres casos se ilustra claramente como la significación y los valores de las variables del modelo cambian en función del resto de variables incluidas en el modelo, por la influencia que estas tienen. Por ejemplo, en el modelo 1, la variable SIDA es significativa, mientras que en el modelo 2, como se añaden las interacciones, la variable deja de ser significativa.

También se muestra como el valor de AIC no es una regla fija inamovible, ya que el que menor AIC tiene es el modelo 2, pero tras evaluar la capacidad clasificadora de los modelos, el que mejor resultados ofrece es el modelo 1.

En las siguientes tablas (Tab. 12) se muestra como el modelo 1 clasifica 2 de 4 pacientes que mueren y 51 de 52 pacientes que no mueren de forma correcta. Esto se traduce en una precisión muy elevada (94,6%) y un valor de Kappa moderado (54,3%).

| Predicciones | Referencia | |
|---|---|---|
| | No | Si |
| No | 51 | 2 |
| Si | 1 | 2 |

| | Valor |
|---|---|
| Accuracy | 0.9464286 |
| Kappa | 0.5434783 |

Tabla 12: Matriz de confusión y metricas.

Por tanto, centrándonos únicamente en el modelo 1, observando los odds ratio podemos concluir que por cada año de más que tenga el paciente, la probabilidad de muerte aumenta en un 3%, que la probabilidad de muerte aumenta en un 73% si el paciente tiene SIDA. Finalmente, observando la variable CD4_V inicial y CD4200, se puede concluir que en los paciente que tienen un conteo de CD4 superior a 200, una unidad extra de CD4 prácticamente no modifica la probabilidad de muerte, pero en los paciente con un valor de CD4 inferior a 200, por cada unidad extra de CD4 inicial, la probabilidad de muerte disminuye en un 0,3%.

#### 5.2.1.4 Evento vs resto de eventos.

Con los siguientes modelos se trata de detectar que variables influyen sobre la aparición de un tipo en concreto de evento ENOS frente a la aparición de otro o ningún evento.

Nos centramos únicamente en los dos eventos más comunes dentro de la base de datos disponible, los cuales son liver (hígado) y neuropsychiatric (neurosiquiátrico), para ellos es necesario crear dos variables dicotómicas que indiquen si se da el evento estudiado (1) o no (0)

Únicamente se dispone de 73 pacientes que sufren un evento de hígado, frente a 483 paciente que sufren otro o ningún evento, y de 54 paciente que sufren un evento neuro frente a 503 pacientes que sufren otro ningún evento. Esta disparidad entre los datos supone un gran problema a la hora de formular el modelo.

En primer lugar, en cuanto al modelo para eventos de hígado, se seleccionan para el modelo las variables con las que se han obtenido mejores resultados en las diferentes pruebas realizadas, y se obtiene el siguiente modelo.

El modelo incluye 3 variables explicativas, las cuales son el biomarcador mda, el valor inicial de CD4 y la variables binaria que indica si el valor de CD4 inicial es mayor o menor a 200. únicamente la variable binaria y el biomarcador son significativas.

| | Modelo 1 |
|---|---|
| Intercept | 0.03 |
| | [−1.32; 1.38] |
| CD4 > 200 | 5.69* |
| | [4.38; 7.00] |
| Mda | 1.03* |
| | [1.00; 1.05] |
| CD4 inicial | 1.00 |
| | [1.00; 1.00] |
| AIC | 373.06 |
| BIC | 389.92 |
| Log Likelihood | -182.53 |
| Deviance | 365.06 |
| Num. obs. | 501 |

* 1 outside the confidence interval

Tabla 13: Modelos.

Si comprobamos la capacidad clasificadora del modelo, se observa que clasifica todos los evento como si no fuesen eventos de hígado, incluso los que, si que lo son, debido a la gran cantidad de individuos que no sufrían ese tipo de evento en los datos utilizados para calcular el modelo. El modelo no "sabe" cuales son las

características que diferencian a los individuos que sufren eventos de hígado de los que no. Esto se traduce en una precisión del 83,9% y un valor de kappa del 0%, extremadamente bajo.

| Predicciones | Referencia | |
|---|---|---|
| | Otro | Hígado |
| Otro | 47 | 9 |
| Hígado | 0 | 0 |

| | Valor |
|---|---|
| Accuracy | 0.8392857 |
| Kappa | 0.0000000 |

Tabla 14: Matriz de confusión y metricas.

Se pueden extraer varias conclusiones de los odds ratios, como que, si el valor de CD4 inicial es menor a 200, la probabilidad de que el paciente sufra un evento de hígado aumenta en un 469% y por cada unidad extra en el biomarcador mda, aumenta la probabilidad de sufrir el evento en un 3%. Estas conclusiones no se pueden tomar como 100% reales debidas a que el modelo no ofrece buenos resultados.

En segundo lugar, de igual forma que en el modelo anterior, para modelizar las probabilidades de que se produzca un evento neuro usamos las variables que mejores resultados han ofrecido tras realizar varias prueba y análisis. Obtenemos el siguiente modelo, el cual incluye la edad, los valores iniciales y del evento de CD4 y el biomarcador mda. Únicamente los valores de CD4 son significativos.

| | Modelo 1 |
|---|---|
| Intercept | 0.83 |
| | [−0.73; 2.39] |
| Edad | 0.97 |
| | [0.94; 1.00] |
| CD4 evento | 1.00* |
| | [1.00; 1.00] |
| CD4 inicial | 1.00* |
| | [1.00; 1.00] |
| Mda | 0.96 |
| | [0.91; 1.01] |
| AIC | 312.34 |
| BIC | 333.42 |
| Log Likelihood | -151.17 |
| Deviance | 302.34 |
| Num. obs. | 501 |

* 1 outside the confidence interval

Tabla 15: Modelos.

Si comprobamos la capacidad clasificadora del modelo observamos que sucede lo mismo que con el modelo anterior, el modelo no ha "aprendido" a diferenciar entre individuos que sufren un evento neuro, de los pacientes que sufren otro o ningún evento.

Nuevamente, obtenemos una precisión muy alta, del 87,5%, pero un valor de kappa del 0%, por el alto acierto en individuos que no sufren el evento, pero haber clasificado como que no lo sufren a todos los individuos que si que lo sufren.

| Predicciones | Referencia | |
|---|---|---|
| | Otro | Neuro |
| Otro | 49 | 7 |
| Neuro | 0 | 0 |

| | Valor |
|---|---|
| Accuracy | 0.875 |
| Kappa | 0.000 |

Tabla 16: Matriz de confusión y metricas.

Como sucedía en el modelo anterior, se pueden extraer ciertas conclusiones del modelo, pero no son necesariamente del 100% fiables, debido a los malos resultados que ofrece el modelo. Un aumento de una unidad en los valores de CD4 iniciales y en el momento del evento, supone un aumento en la probabilidad de sufrir un evento neuro.

### 5.2.2 Análisis de supervivencia

En el presente apartado se aplican dos de las técnicas más extendidas de análisis de supervivencia. En primer lugar, se estudian lo datos según las curvas de Kaplan-Meier, para comprobar si existen diferencias significativas en la supervivencia de los pacientes según ciertas categorías. En segundo lugar, se ajustan diferentes modelo de riesgos proporcionales de Cox, para estudiar las variables que afectan en la supervivencia de los pacientes.

#### 5.2.2.1 Desde entrada hasta muerte

En primer lugar, se centra el estudio en el tiempo transcurrido desde que el paciente entra en el estudio hasta que muere.

Comenzamos calculando las curvas de Kaplan-Meier según dos variables, si el paciente experimenta algún evento ENOS, y si su valor de CD4 es mayor o menor a 200. Como se puede observar a continuación, no parecen existir demasiadas diferencias en la supervivencia de los pacientes que sufren un evento ENOS y CD4 $> 200$, y los pacientes que no sufren un evento y CD4 $> 200$, por tanto, que el paciente sufra un evento no parece ser una característica que influya en gran medida en la supervivencia del paciente. Por otro lado, se observan grandes diferencias significativas cuando el valor de CD4 del paciente es menor a 200, sufra o no un evento, aunque cuando sufre un evento y CD4 $< 200$, no transcurre mucho tiempo hasta que la probabilidad de supervivencia baja del 50%. En el resto de las categorías, no se baja en ningún momento del 50% de probabilidad de supervivencia.

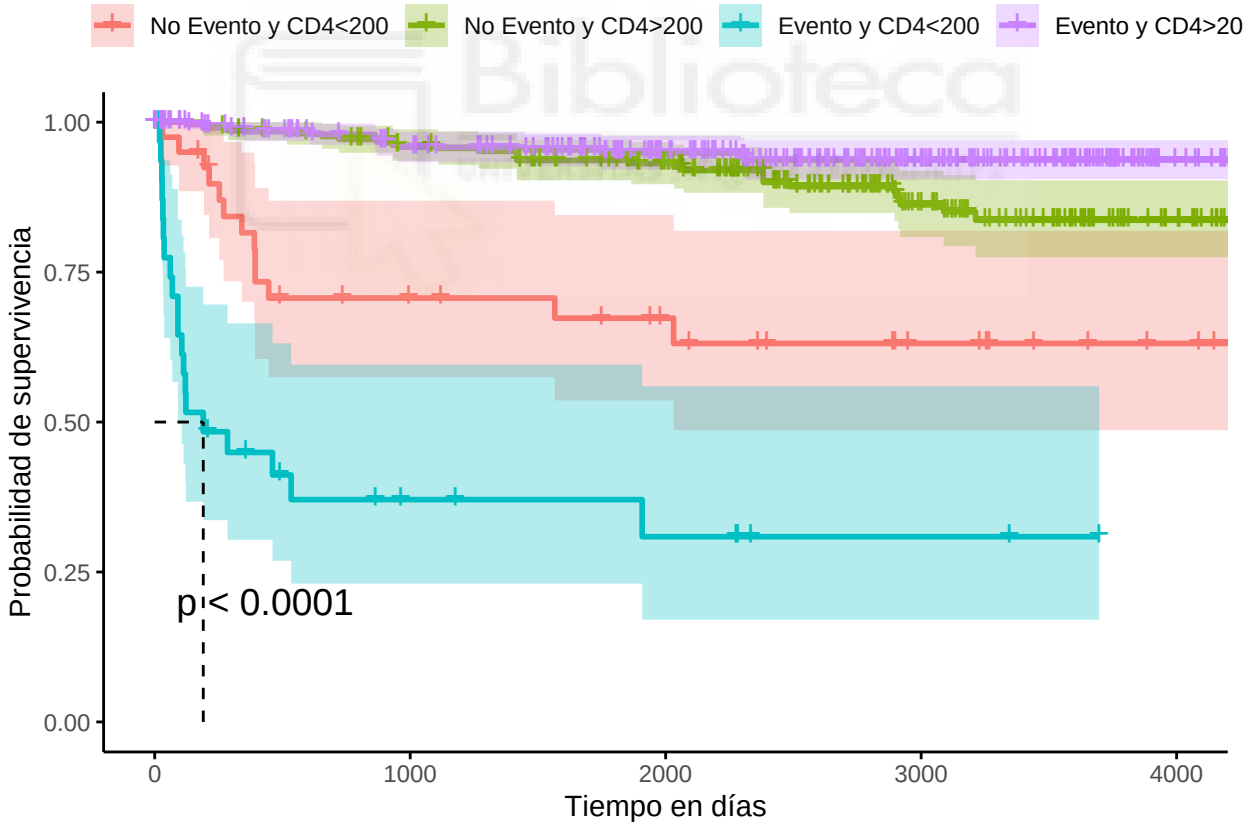


Figure 22: Función de Kaplan-Meier

Para conocer que variables que influyen en la supervivencia de los pacientes y en que medida, se calcula un

modelo de Cox. Tras realizar diferentes pruebas, se obtiene el siguiente modelo.

El modelo contiene 3 biomarcadores (neop, mda y il8), si el paciente recibe o no tratamiento contra el VIH y la edad del paciente. Todas las variables son significativas.

| | Modelo 1 |
|---|---|
| IL8 | 1.08* |
| | [1.04; 1.13] |
| Neop | 1.10* |
| | [1.04; 1.15] |
| Mda | 1.14* |
| | [1.04; 1.23] |
| Tratamiento | 0.03 |
| | [−2.23; 2.29] |
| Edad | 1.29* |
| | [1.16; 1.43] |
| AIC | 65.20 |
| $R^2$ | 0.59 |
| Max. $R^2$ | 0.83 |
| Num. events | 15 |
| Num. obs. | 64 |
| Missings | 493 |
| PH test | 0.01 |

* 1 outside the confidence interval



Tabla 17: Modelos.

En el siguiente gráfico de bosque, se representa gráficamente la influencia de cada una de las variables del modelo sobre la probabilidad de muerte del paciente. Es importante remarcar que mientras que en las curvas de Kaplan-Meier se estudiaba la probabilidad de supervivencia, en el modelo de Cox lo que se estudia es la probabilidad de muerte.

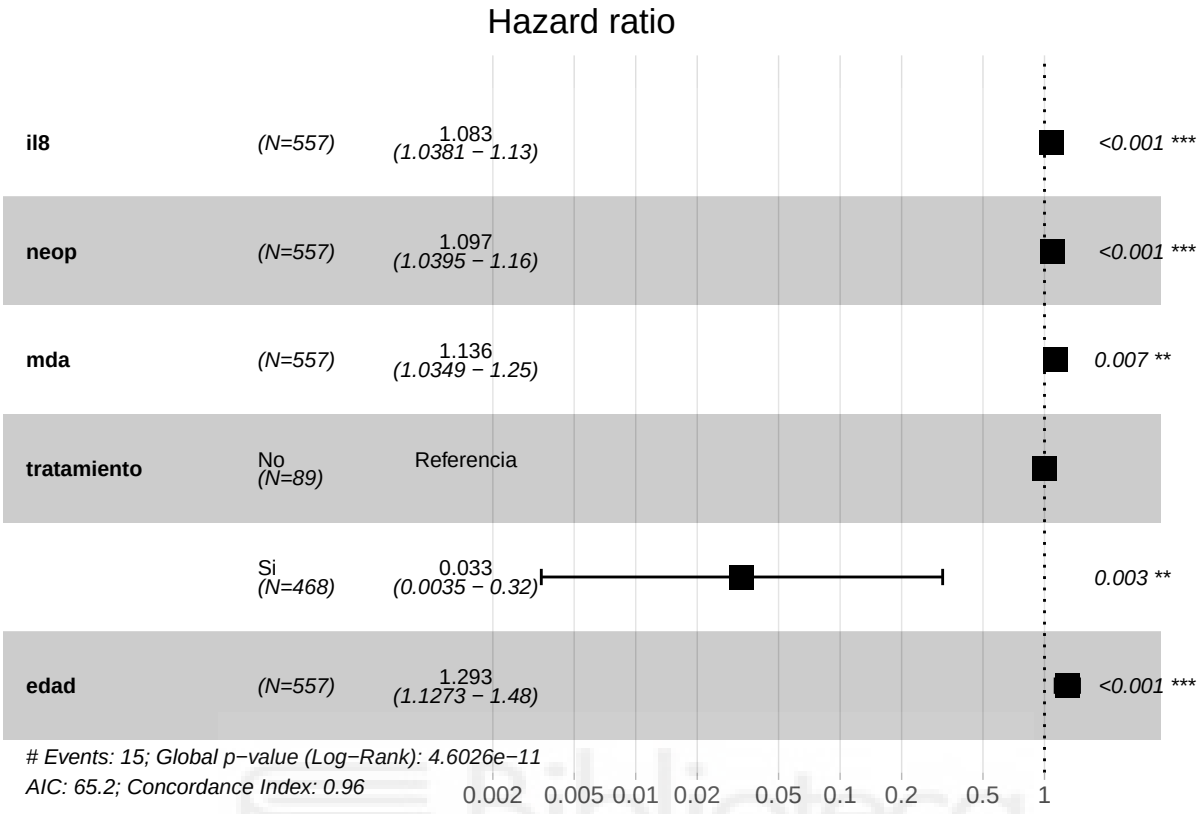


Figure 23: Hazard ratio

Observando los Hazard ratio del modelo, podemos concluir que lo que más reduce la probabilidad de muerte en un paciente infectado con VIH es el recibir tratamiento, reduciéndola en un 96,7%. Por otro lado, la edad influye de forma negativa, a mayor edad del paciente, más probable es la muerte, en concreto, por cada año de más, la probabilidad crece en un 29,3%. Lo mismo sucede con los 3 biomarcadores incluidos en el modelo, por ejemplo, por una unidad extra de il8, la probabilidad de muerte aumenta en un 8,3%, o por una unidad extra de mda, aumenta en un 13,6%.

#### 5.2.2.2 Desde entrada hasta evento

En segundo lugar, nos centramos en el estudio del tiempo transcurrido desde que el paciente entra en el estudio hasta que desarrolla un evento ENOS.

Como podemos observar en las siguientes curvas de Kaplan-Meier, existe una diferencia significativa en los tiempos de aparición de un evento en función de si el paciente recibe o no tratamiento contra el VIH. Los que no reciben tratamiento tienen una probabilidad menor del 50% de no sufrir un evento antes de 2500 días (~7 años), mientras que los que reciben tratamiento, no bajan del 50% hasta pasados los 3125 días (~8,5 años).

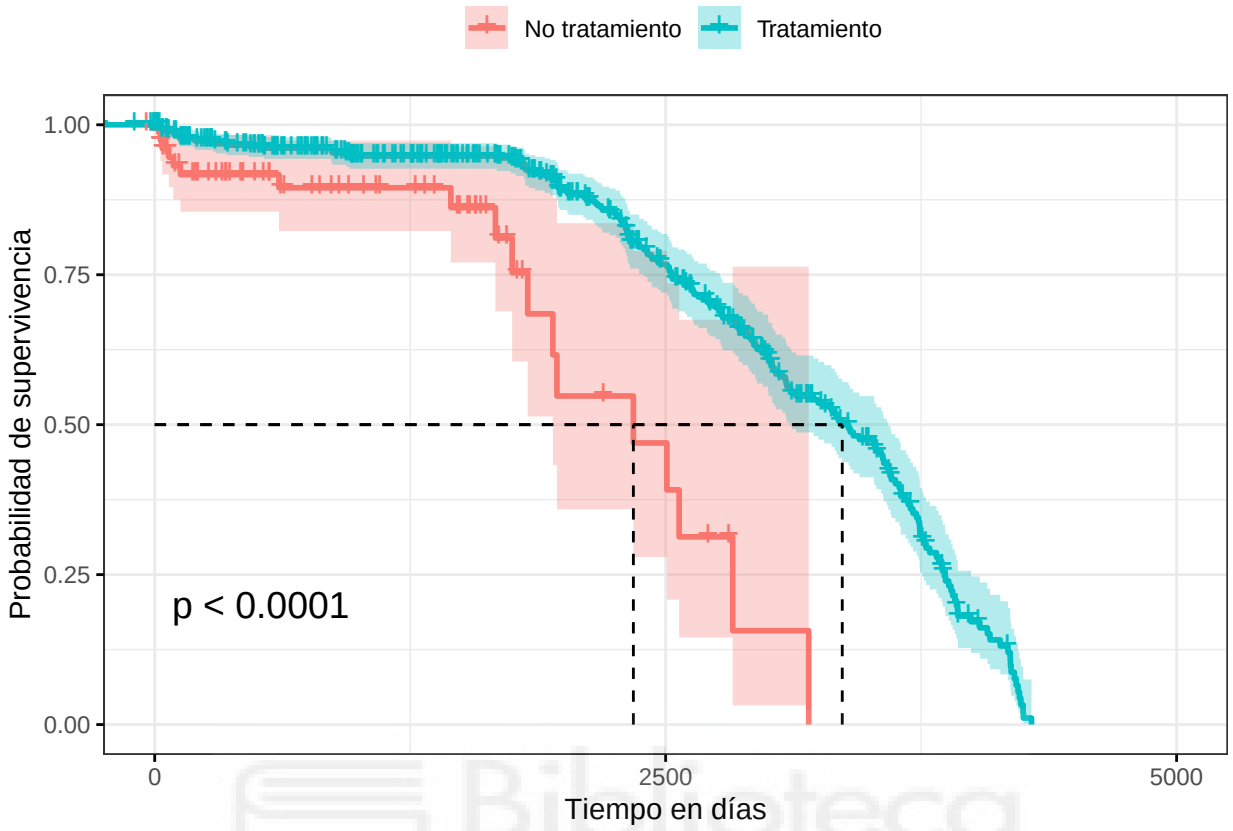


Figure 24: Función de Kaplan-Meier

En cuanto al modelo de Cox, tras realizar distintas aproximaciones, se ha obtenido el siguiente modelo, el cual incluye un biomarcador (neop), y tres variables binarias, las cuales indican si el paciente recibe o no tratamiento (RECART_Y), si su valor inicial de CD4 es mayor o menor a 200 (CD4200) y si está o no infectado por el VHC (vhc). Todas las variables son significativas.

| | Modelo 1 |
|---|---|
| Neop | 1.02* |
| | [1.01; 1.02] |
| Tratamiento | 0.33* |
| | [−0.27; 0.93] |
| CD4 > 200 | 0.23* |
| | [−0.28; 0.75] |
| VHC | 0.45* |
| | [−0.00; 0.91] |
| AIC | 1805.81 |
| $R^2$ | 0.10 |
| Max. $R^2$ | 0.97 |
| Num. events | 196 |
| Num. obs. | 543 |
| Missings | 14 |
| PH test | 0.00 |

* 1 outside the confidence interval

Tabla 18: Modelos.

En el siguiente gráfico de bosque se representan de forma gráfica la influencia de cada variable del modelo sobre la probabilidad de desarrollar un evento ENOS.

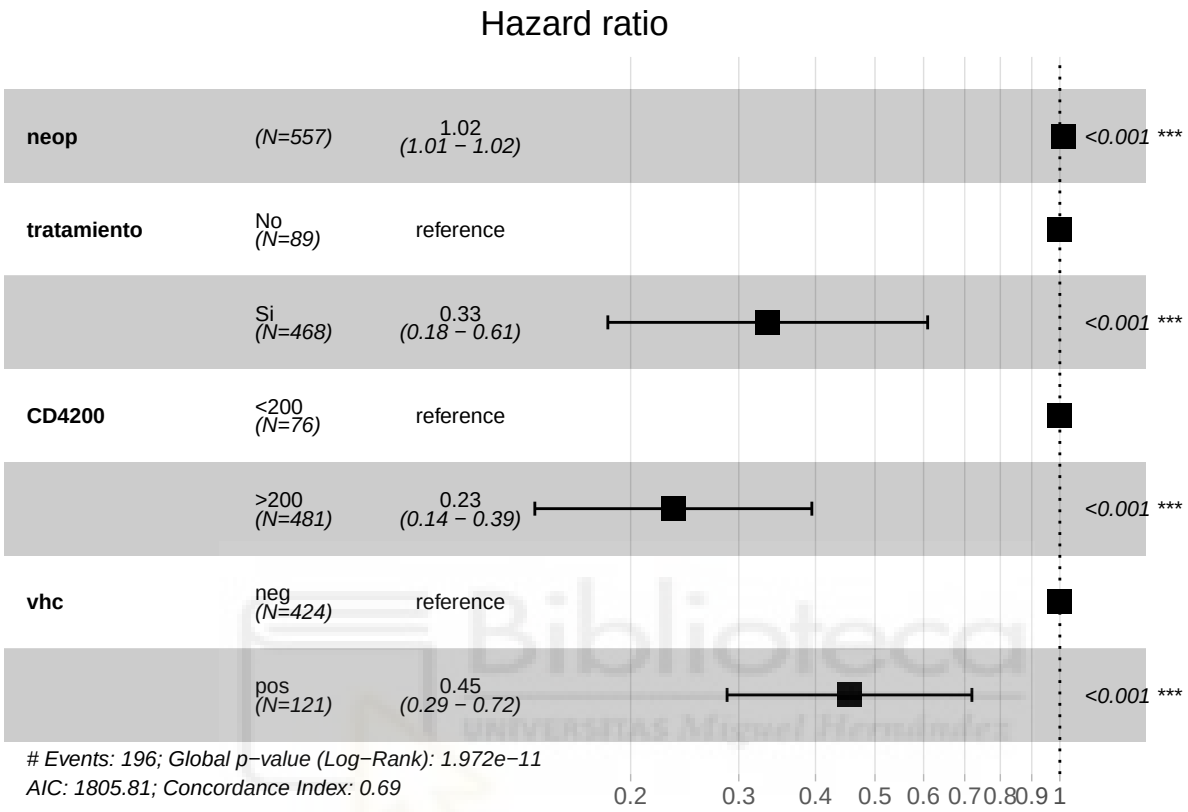


Figure 25: Hazard Ratio.

La variables que mayor efecto tienen sobre la probabilidad de desarrollar un evento son "tratamiento", la cual indica que si el paciente recibe tratamiento se reduce la probabilidad de sufrir un evento en un 67%, CD4200, la cual indica que si el valor de CD4 es mayor a 200, se reduce la probabilidad en un 77%, y finalmente vhc, la cual indica que si el paciente está infectado, se reduce la probabilidad de sufrir un evento ENOS en un 55%. Por otro lado, el modelo indica que por cada unidad extra del biomarcador neop, la probabilidad de sufrir un evento aumenta en un 2%.

#### 5.2.2.3 Desde evento hasta muerte

Finalmente, estudiamos el tiempo transcurrido desde que el paciente desarrolla un evento hasta que muere. Para ello, en primer lugar, en cuanto a las curvas de Kaplan-Meier, se representa una curva para cada tipo de evento ENOS. Utilizamos únicamente los más comunes, el resto los incluimos junto a los pacientes que no desarrollan evento en la categoría "otros".

Como se puede observar, si el paciente ha tenido un evento ENOS, independientemente del tipo, la supervivencia similar, exceptuando los eventos "neoplastic", los cuales tienen una mortalidad mucho mayor, la probabilidad de supervivencia baja del 50% antes de los 3 años.

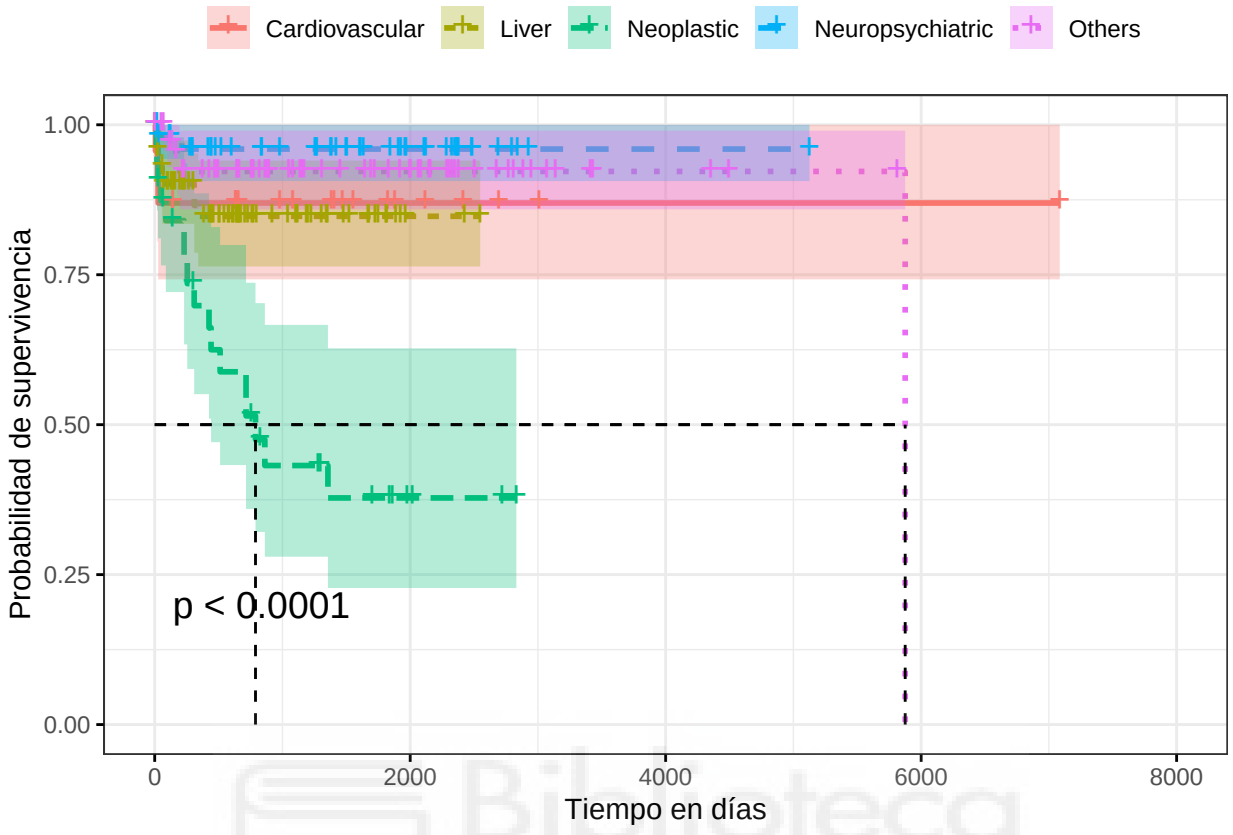


Figure 26: Función de Kaplan-Meier

En cuanto al modelo de Cox, tras realizar diversas aproximaciones se ha obtenido el siguiente modelo (modelo 1), el cual incluye 5 biomarcadores, como il8, neop o dd, y la edad del paciente. Por otro lado, se ha modelizado la supervivencia con una única variable explicativa, la cual es la misma que se ha utilizado para representar las curvas anteriores, e indica los diferentes tipos de eventos (modelo 2)

| | Modelo 1 |
|---|---|
| IL8 | 1.93* |
| | [1.47; 2.38] |
| Neop | 1.79* |
| | [1.39; 2.19] |
| DD | 0.97* |
| | [0.95; 0.99] |
| Mda | 1.72* |
| | [1.34; 2.09] |
| CD40 | 0.93* |
| | [0.87; 0.98] |
| Edad | 1.71* |
| | [1.34; 2.07] |
| AIC | 45.28 |
| $R^2$ | 0.75 |
| Max. $R^2$ | 0.92 |
| Num. events | 13 |
| Num. obs. | 30 |
| Missings | 527 |
| PH test | 0.00 |

* 1 outside the confidence interval

Tabla 19: Modelos.

| | Modelo 2 |
|---|---|
| Evento Hígado | 1.52 |
| | [0.16; 2.87] |
| Evento Neoplastico | 5.96* |
| | [4.67; 7.25] |
| Evento Neuropsiquiatrico | 0.35 |
| | [−1.49; 2.18] |
| Otros eventos | 0.74 |
| | [−0.67; 2.15] |
| AIC | 368.96 |
| $R^2$ | 0.12 |
| Max. $R^2$ | 0.79 |
| Num. events | 38 |
| Num. obs. | 253 |
| Missings | 304 |
| PH test | 0.01 |

* 1 outside the confidence interval

Tabla 20: Modelos.

Los resultados en el modelo 2 son los esperados si se observan las curvas anteriores, en la única categoría en la que se encuentran diferencias significativas es en los eventos “neoplastic”.

En el siguiente gráfico de bosque se representan los Hazard ratio de las variables del modelo 1.

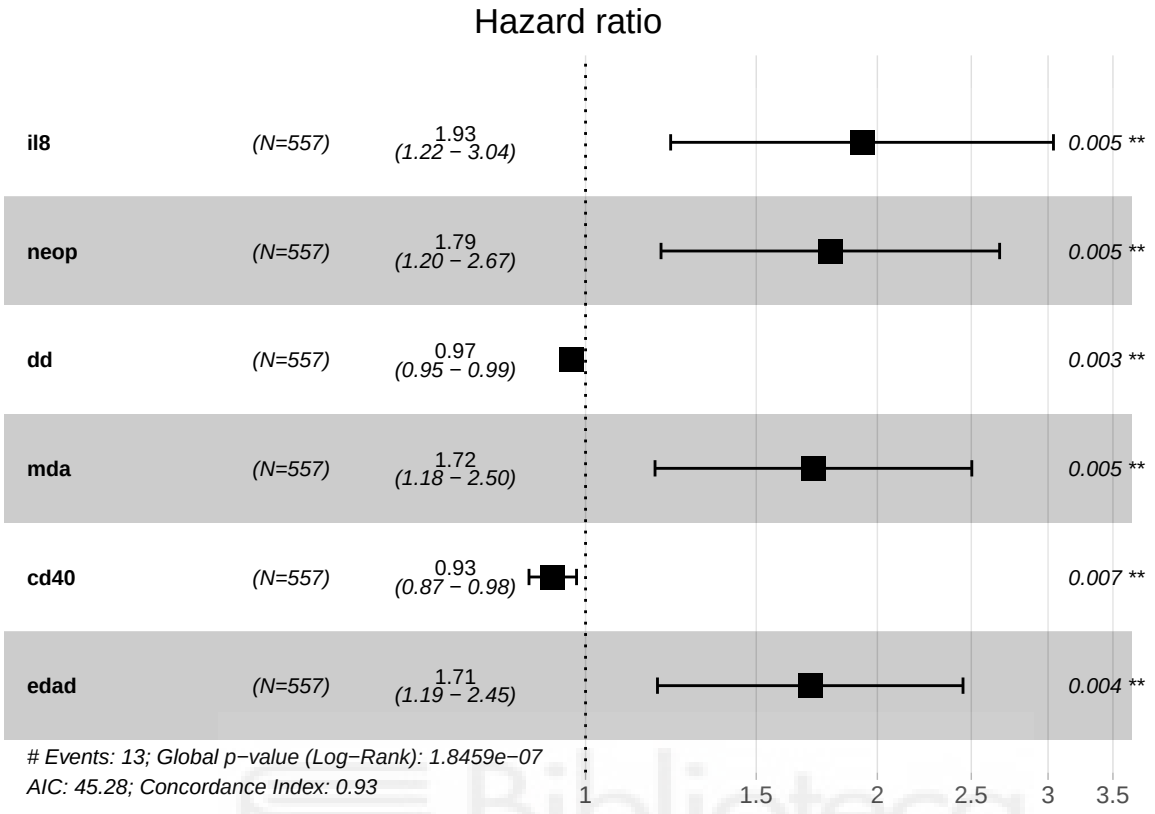


Figure 27: Hazard Ratio.

Observando los hazard ratio se pueden extraer varias conclusiones. Por cada unidad extra en el biomarcador dd, la probabilidad de muerte se reduce en un 3%, de la misma forma, una unidad extra de cd40, disminuye la probabilidad de muerte en un 7%. Por otro lado, el resto de las variables aumentan la probabilidad de muerte, por ejemplo, por cada año extra que tenga el paciente, la probabilidad de muerte aumenta en un 71%, o por cada unidad extra de il8, la probabilidad de muerte aumenta en un 93%.

### 5.2.3 Arbol de clasificación

Por el propio funcionamiento del algoritmo que genera los arboles de clasificación, se ha optado por introducir en el modelo todas las variables, dejando que el propio modelo seleccione las mejores variables explicativas para cada variable respuesta que se estudie con cada modelo, exceptuando las variables que puede causar problemas como el código de identificación de cada paciente, su fecha de nacimiento, etc.

De igual forma que en las técnicas anteriores, se ha dividido la base de datos, de forma aleatoria, en dos bases de datos, "train" (90%), con la cual se calculan los modelos, y "test" (10%), con la cual se evaluará su capacidad para clasificar.

#### 5.2.3.1 Evento sí o no

Con el siguiente modelo se buscan las variables que influyen en que el paciente sufra un evento, del tipo que sea, o no. Tras realizar varias pruebas, con diferentes grupos de variables y con diferentes podas, se obtiene el siguiente modelo:

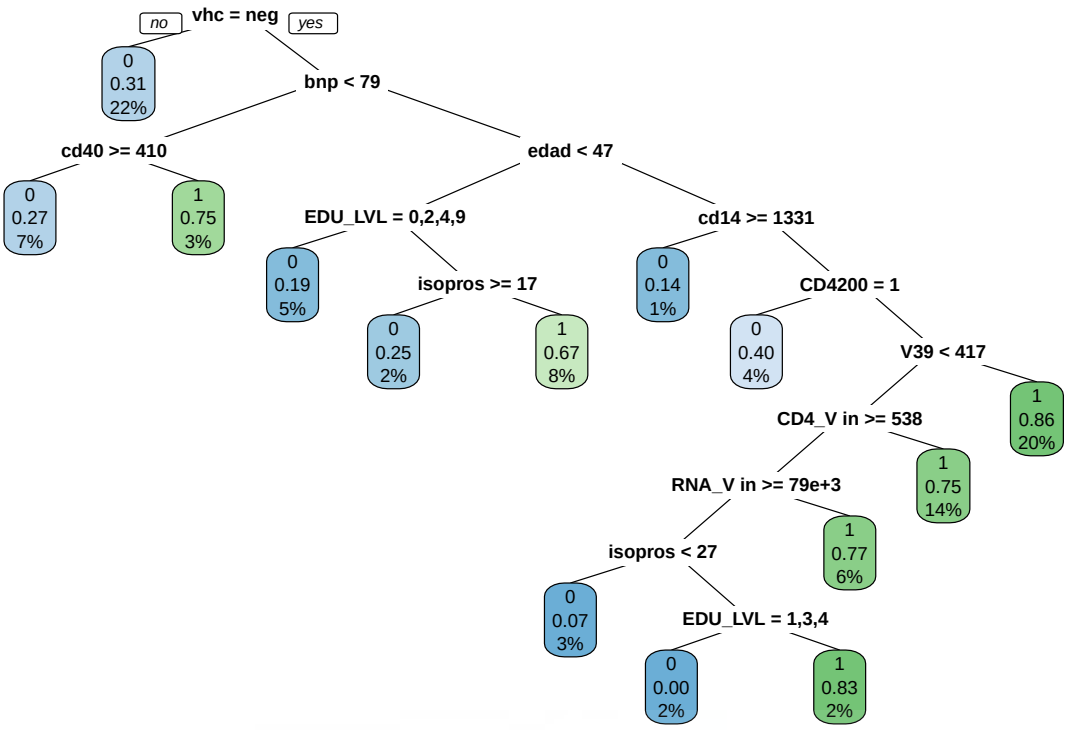


Figure 28: Arbol de clasificación evento sí o no

En la siguiente tabla (Tab. 21) se muestra la importancia de las variables, marcadas en verde las variables que se usan en el arbol de clasificación. Observamos que no necesariamente se usan las variables más importantes, ya que lo que se busca es un arbol de clasificación lo más preciso posible, para lo cual no es necesario que las variables sean las de más importancia.

Como podemos observar, algunas de las variables más importantes, usadas en el modelo, son el nivel educativo del paciente, si es positivo o negativo en VHC, la edad, y algunos biomarcadores, como V39, bpn o isopros, entre otros.

| | Valor |
|---|---|
| EDU_LVL | 14.823005 |
| vhc | 14.485839 |
| edad | 14.415903 |
| V39 | 10.768724 |
| RNA_V inicial | 9.541914 |
| CD4_V inicial | 9.026917 |
| bnp | 8.820156 |
| isopros | 7.463347 |
| cd40 | 6.989775 |
| cd14 | 5.937757 |
| CD4200 | 4.183935 |

Tabla 21: Importancia de las variables.

A continuación, se evalúa la capacidad predictora del modelo, y como se puede observar en las siguientes tablas (Tab. 22), se obtiene una clasificación bastante buena, con una precisión del 76,7% y un valor de

Kappa moderado, del 53,3%, por tanto, este modelo es capaz de predecir si un paciente desarrollará un evento ENOS con bastante fiabilidad.

| Predicciones | Referencia | |
|---|---|---|
| | No | Sí |
| No | 19 | 5 |
| Sí | 8 | 24 |

| | Valor |
|---|---|
| Accuracy | 0.7678571 |
| Kappa | 0.5333333 |

Tabla 22: Matriz de confusión y metricas.

Cuando se trabaja con arboles de clasificación no se dispone de valores numéricos, como los Odds Ratio en la regresión logística, con los que poder extraer conclusiones de en que medida afecta cada variable explicativa a la variable respuesta, pero, por otro lado, se obtiene una estructura lógica (Fig. 28) con la que se pueden extraer conclusiones. Por ejemplo, un paciente negativo en VHC, con un valor de bnp menor a 79, menor de 47 años, un valor de cd14 mayor o igual a 1331, con un valor de CD4 menor a 200 y un valor de V39 menor a 417, es muy probable que desarrolle un evento ENOS, mientras que si se dan las condiciones anteriores, exceptuando que su valor de CD4 sea mayor a 200, es más probable que el paciente no desarrolle un evento ENOS. Por tanto, para cada camino del arbol, existe una interpretación diferente, y no se puede interpretar un nodo sin tener en cuenta el anterior o los siguientes.

#### 5.2.3.2 Eventos 1

Para este arbol, prescindimos de los pacientes sin evento y de los pacientes de los eventos minoritarios para estudiar únicamente que variables afectan a los eventos más comunes: liver, cardiovascular, neoplastic o neuropsychriatic.

Por desgracia, la base de datos no cuenta con tantos pacientes que hayan sufrido los evento estudiados, como pacientes que no han sufrido evento, por tanto, para el siguiente modelo se pierde una gran parte de la población. Tras realizar varias pruebas, sin necesidad de podar, se obtiene el siguiente modelo:

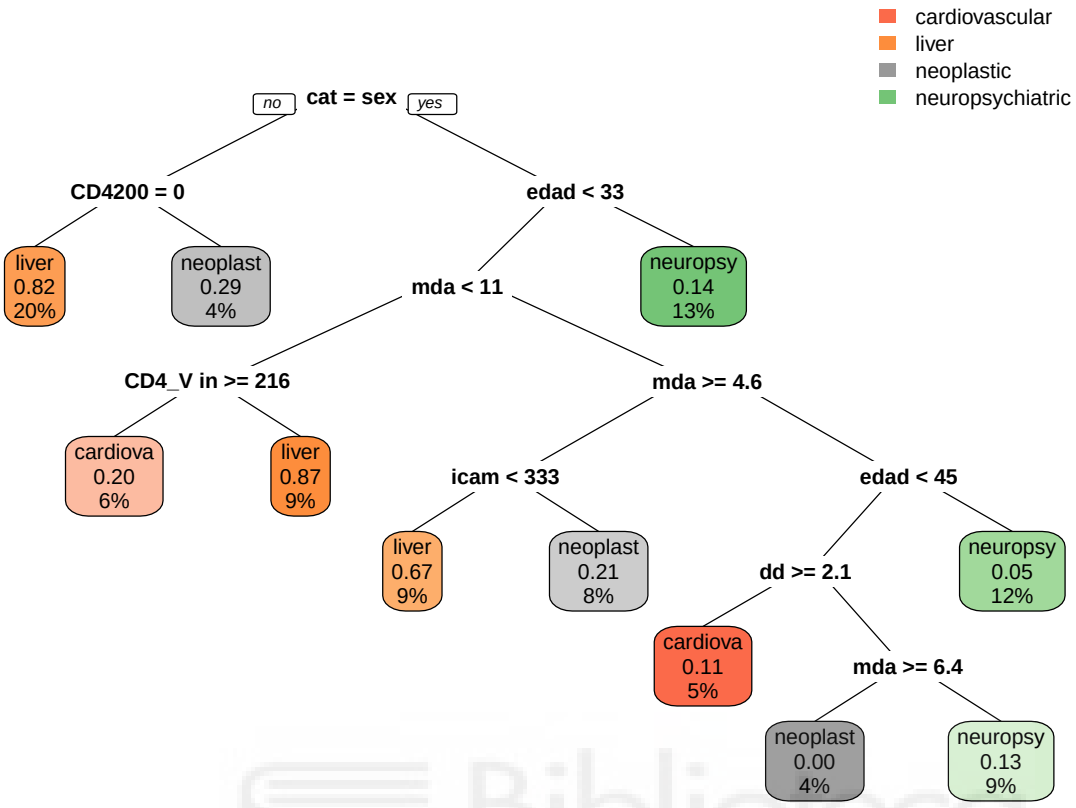


Figure 29: Arbol de clasificación entre eventos.

Como se puede observar, en este arbol de clasificación se usan muchas menos variables para discriminar entre tipos que en el modelo anterior, posiblemente debido a que anteriormente, todos los individuos que habían sufrido un evento ENOS se encontraban agrupados en un único grupo, mientras que, al clasificar a los individuos según el evento sufrido, sus características serán más similares.

En la siguiente tabla (Tab. 23) observamos la importancia de las variables usadas en el modelo, y como se puede observar, en el modelo se incluyen algunos biomarcadores como mda, icam y dd, y otras variables como el CD4 inicial, la variable CD200, que indica si el valor inicial de CD4 es mayor o menor a 200, la edad, y la forma en la que se infectaron de VIH los pacientes (cat).

| | Valor |
|---|---|
| mda | 15.481220 |
| edad | 14.205358 |
| cat | 8.765030 |
| CD4_V inicial | 6.213160 |
| icam | 4.311508 |
| dd | 4.086631 |
| CD4200 | 3.146970 |

Tabla 23: Importancia de las variables

Si evaluamos la capacidad para clasificar, observamos que se obtienen resultados muy buenos, con una precisión del 89,7% y un valor de Kappa muy alto, del 83,1%. Por tanto, este modelo es muy bueno prediciendo si un paciente sufrirá alguno de estos eventos. En la tabla de doble entrada (Tab. 24) se muestra

como clasifica a todos los individuos de forma correcta, exceptuando un individuo que sufre un evento de hígado, y es clasificado como neuropsiquiatrico, y un paciente que sufre un evento neuropsiquiatrico y es clasificado como evento de hígado. Es importante tener en cuenta que ya que no se han tenido en cuenta los pacientes que no sufren evento para este arbol de clasificación, las conclusiones se deben abordar con cierta cautela, ya que al modelo le falta esa pieza de información, y es posible que un paciente que no va a sufrir ningún tipo de evento no se ajuste adecuadamente a este modelo.

| | cardiovascular | liver | neoplastic | neuropsychiatric |
|---|---|---|---|---|
| cardiovascular | 2 | 0 | 0 | 0 |
| liver | 0 | 8 | 0 | 1 |
| neoplastic | 0 | 0 | 1 | 0 |
| neuropsychiatric | 0 | 1 | 0 | 6 |

| | Valor |
|---|---|
| Accuracy | 0.8947368 |
| Kappa | 0.8318584 |

Tabla 24: Matriz de confusión y metricas.

Como comentábamos en el arbol anterior, cada camino del arbol (Fig. 29) representa una conclusión distinta, por ejemplo, si un paciente se infecta de VIH por compartir agujas, y su valor de CD4 es menor a 200, posiblemente sufrirá un evento de hígado, posiblemente debido al uso de drogas, mientras que si un paciente es infectado por medio de relaciones sexuales, es mayor a 33 años, su valor de mda es mayor a 11, y su CD4 inicial es menor a 216, es probable que sufra un evento cardiovascular.

#### 5.2.3.3 Eventos 2

El siguiente arbol de clasificación es muy similar al anterior, con la diferencia de que en este modelo incluimos todos los tipos de eventos que no habíamos tenido en cuenta en el modelo anterior y los clasificamos como "otros", junto a los pacientes que no sufren ningún evento ENOS. Por tanto, si un paciente clasificado como "otros", es posible que no sufra ningún evento, o que sufra uno del cual no tenemos demasiados registros. Tras realizar varias pruebas, sin necesidad de podar, se obtiene el siguiente modelo:

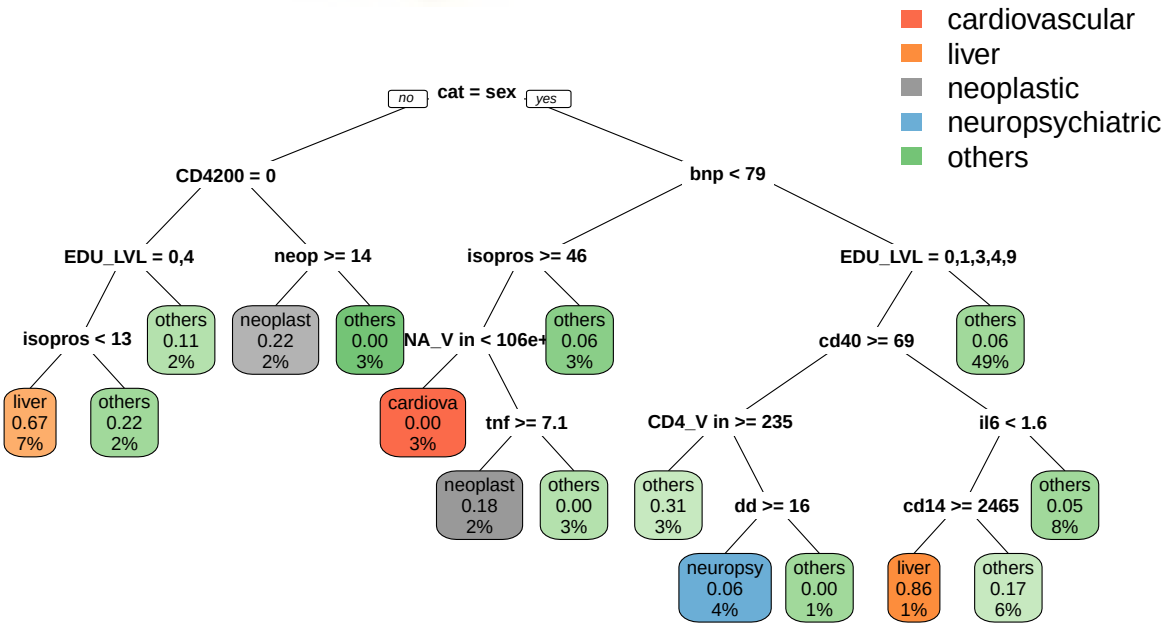


Figure 30: Arbol de clasificación entre eventos.

En la siguiente tabla (Tab. 25) se muestra la importancia de cada una de las variables usadas en el modelo, algunas de las cuales son CD4200, que indica si el valor de CD4 inicial es mayor o menor a 200, algunos biomarcadores como neop, isopros o dd, entre otros, y otras variables como valor de Carga Vírica inicial.

| | Valor |
|---|---|
| cd40 | 10.650555 |
| neop | 10.051207 |
| isopros | 9.994727 |
| cat | 8.724338 |
| EDU_LVL | 8.264236 |
| tnf | 7.898868 |
| CD4_V inicial | 6.924822 |
| il6 | 6.333320 |
| dd | 6.252944 |
| CD4200 | 6.027066 |
| bnp | 5.585516 |
| cd14 | 5.332519 |
| RNA_V inicial | 5.145433 |

Tabla 25: Importancia de las variables

Si evaluamos la capacidad clasificadora del arbol, obtenemos resultados ligeramente peores al arbol anterior, pero igualmente bastante buenos, con una precisión de 78,57% y un valor de Kappa del 53,8%, y si observamos la taba de doble entrada, la clasificación es relativamente buena, aunque se podría mejorar disponiendo de una mayor cantidad de individuos en cada tipo de evento para el calculo del arbol.

| | cardiovascular | liver | neoplastic | neuropsychiatric | others |
|---|---|---|---|---|---|
| cardiovascular | 0 | 0 | 0 | 0 | 1 |
| liver | 0 | 6 | 0 | 0 | 1 |
| neoplastic | 0 | 0 | 1 | 0 | 1 |
| neuropsychiatric | 0 | 0 | 0 | 3 | 0 |
| others | 2 | 3 | 0 | 4 | 34 |

| | Valor |
|---|---|
| Accuracy | 0.7857143 |
| Kappa | 0.5387783 |

Tabla 26: Matriz de confusión y metricas.

Como comentábamos anteriormente, en cuanto a conclusiones, cada camino del árbol es una conclusión diferente, por ejemplo, si el paciente se infecta por compartir agujas con una persona infectada, tiene un valor de CD4 inicial inferior a 200 y su nivel educativo es universitario o nulo, probablemente pertenezca al grupo "otros", peri si el paciente se infecta por transmisión sexual de IVH, tiene un valor de bnp superior a 79, un valor de isopros inferior a 46 y una carga vírica inicial superior a 106000, es probable que sufra un evento cardiovascular.

### 5.2.4 Redes neuronales

Para la elaboración de las siguientes redes neuronales redes neuronales se han seleccionado algunas de las variables que se ha observado en los modelos anteriores que mejores resultados ofrecen, como son algunos biomarcadores (tnf, il6, dd, etc.), algunas variables binarias (CD4200, RECART_Y o AIDS_Y) y otras variables como la edad o el valor inicial de CD4. De igual forma que se ha hecho a lo largo del estudio con el resto de los modelos, se ha utilizado el 90% de la base de datos para formular las redes neuronales, y el 10% restante para evaluar la potencia y capacidad predictiva de las mismas.

#### 5.2.4.1 Evento sí o no

En primer lugar, construimos una red neuronal que sea capaz de predecir si el paciente desarrollará un evento ENOS de cualquier tipo, o no.

Tras varias pruebas estas son las dos mejores redes neuronales que se han obtenido.

En primer lugar, tenemos una red neuronal con dos capas ocultas, con 5 neuronas en cada capa y una única neurona de output que devuelve la probabilidad de que el paciente sufra un evento ENOS. Para todas las neuronas se ha utilizado la función sigmoide de activación. La red neuronal tiene la siguiente estructura:

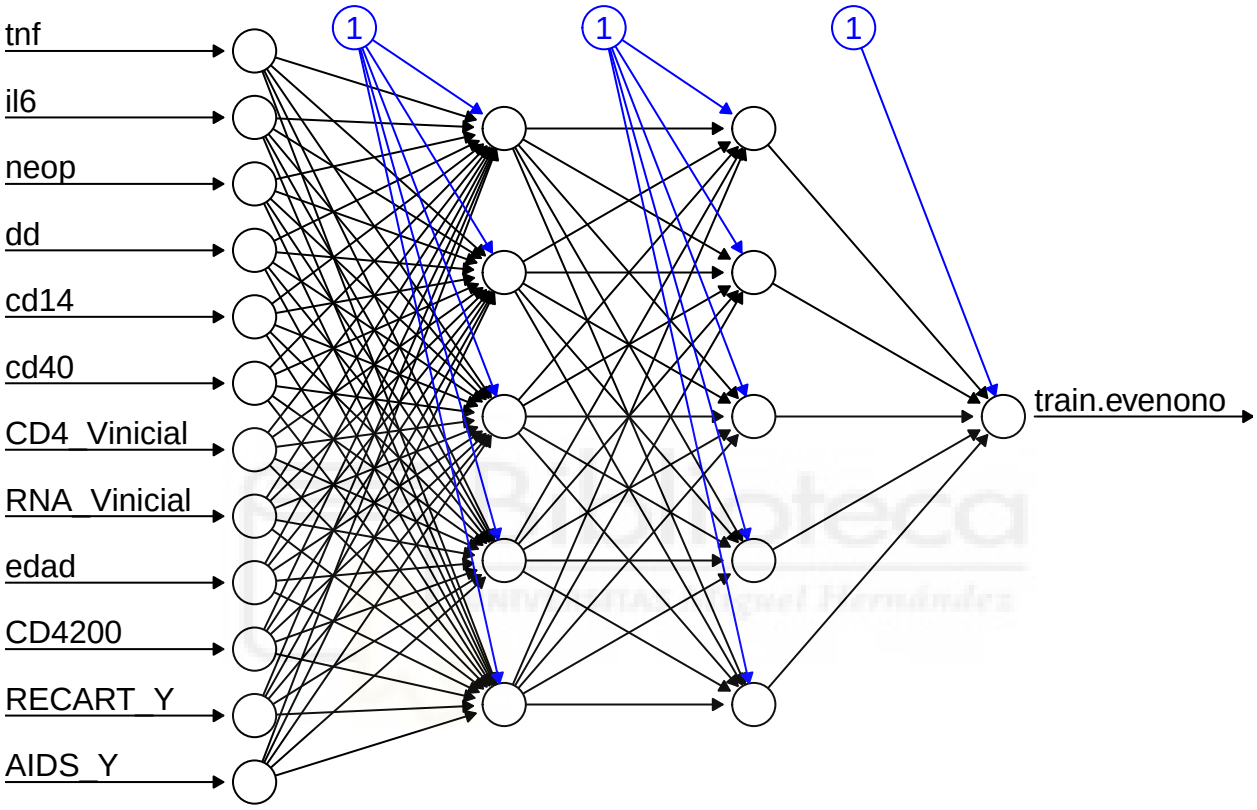


Figure 31: Red neuronal evento sí o no.

Si evaluamos si capacidad predictora, observamos que es relativamente buena, con una precisión del 65,5%, aunque el el valor de Kappa sea bastante bajo (30,1%). Se clasificando la mayoría de los individuos correctamente, por tanto, esta es una posible red neuronal para predecir si el paciente sufrirá o no un evento.

| Predicciones | Referencia | |
|---|---|---|
| | No | Sí |
| No | 15 | 10 |
| Sí | 9 | 21 |

| | Valor |
|---|---|
| Accuracy | 0.6545455 |
| Kappa | 0.3010033 |

Tabla 27: Matriz de confusión y metricas.

Por otro lado, otra posible red neuronal es la siguiente, en este caso, se ha añadido el biomarcador V39, lo que ha hecho que sea necesario cambiar por completo la estructura de la red. La red cuenta con 3 capas

ocultas, la primera con 5 neuronas, y las dos siguientes con 2 neuronas cada capa. Nuevamente todas las neuronas cuentan con la función sigmoide como función de activación y hay una única neurona de output la cual devuelve la probabilidad de que le paciente sufra un evento ENOS.

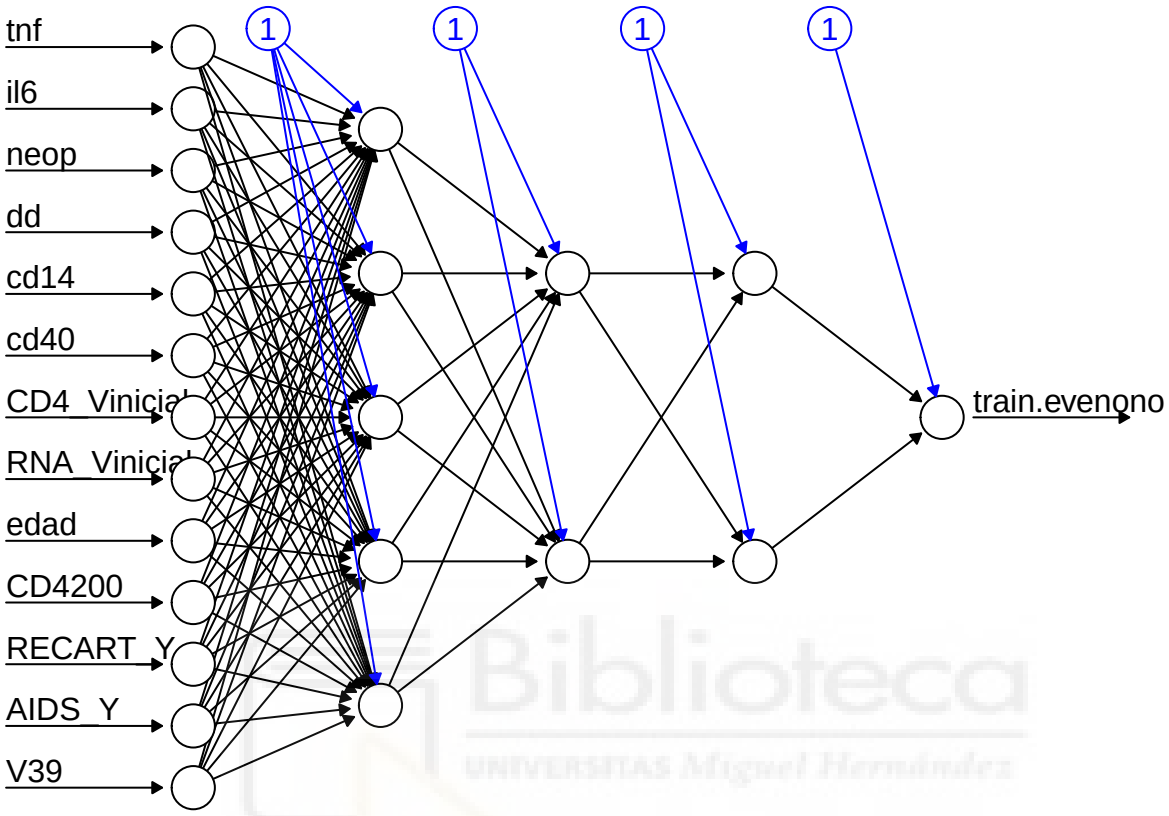


Figure 32: Red neuronal evento sí o no.

Evaluando nuevamente su capacidad predictiva, se observa que con esta red neuronal se obtienen resultados notablemente mejores,aunque siguen siendo regulares, con una precisión del 70,9% y un valor de Kappa del 37,3%.

| Predicciones | Referencia | |
|---|---|---|
| | No | Sí |
| No | 10 | 2 |
| Sí | 14 | 29 |

| | Valor |
|---|---|
| Accuracy | 0.7090909 |
| Kappa | 0.3732194 |

Tabla 28: Matriz de confusión y metricas.

La redes neuronales son una clase de modelos conocidos como “modelos de caja negra”, lo que quiere decir que no son modelos de fácil interpretación, como puede ser una regresión logística, en la cual con los parámetros del modelo se puede obtener la medida en la que influye cada variable sobre el resultado. Lo que se pierde en interpretabilidad del modelo, se gana en potencia del propio modelo.

#### 5.2.4.2 Muerte

A continuación, manteniendo el mismo set de variables que en los modelos anteriores, construimos una red neuronal que sea capaz de predecir si un paciente va a morir o no.

Tras diversas pruebas, cambiando diferentes parámetros de la red neuronal y la estructura de la misma, obtenemos la siguiente red, la cual cuenta con una única capa oculta con dos neuronas y una única capa de output, la cual devuelve la probabilidad de que un paciente muera. Nuevamente, todas las neuronas cuentan con la función sigmoide como función de activación, debido a su gran potencia y adaptabilidad. La estructura de la red es la siguiente.

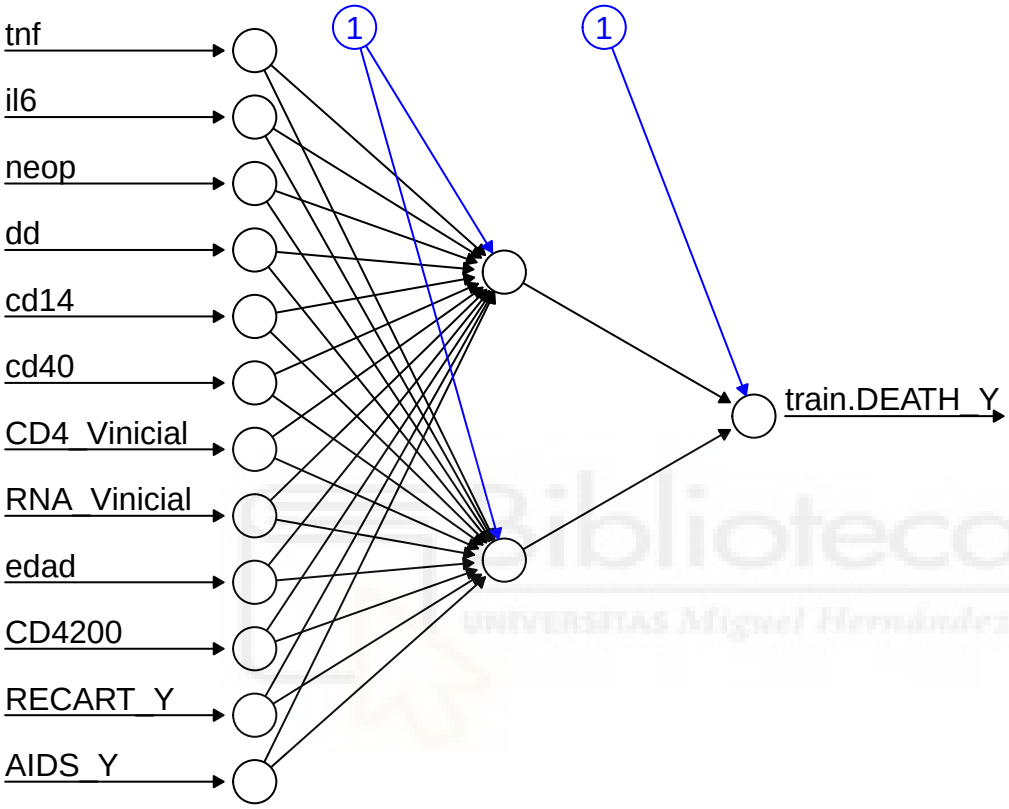


Figure 33: Red neuronal muerte sí o no.

Si se evalúa la capacidad predictora de la red se observan resultados muy buenos, una precisión del 90,9% y un valor moderado de Kappa, 57,2%. Si se compara con las dos redes anteriores se observa que no necesariamente por que la estructura de la red sea muy compleja, se obtienen resultados mejores, ya que, en este caso, con una única capa oculta y dos neuronas, se obtienen resultados muy buenos.

| Predicciones | Referencia | |
|---|---|---|
| | No | Sí |
| No | 46 | 0 |
| Sí | 5 | 4 |

| | Valor |
|---|---|
| Accuracy | 0.9090909 |
| Kappa | 0.5723173 |

Tabla 29: Matriz de confusión y metricas.

#### 5.2.4.3 Tipo evento

Finalmente, se construye una red neuronal para resolver un problema bastante mas complejo que los anteriores, en este caso la función de la red neuronal es predecir que tipo de evento va a sufrir el paciente. Esta situación es bastante mas compleja, no solo por la variedad de outputs que puede devolver la red, si no por que la base de datos utilizada en el estudio no cuenta con la misma cantidad de eventos de todos los tipos, lo que dificulta la tarea de la red de "aprender" cuales son las características de cada tipo de evento.

En primer lugar, únicamente se predicen 4 tipos de eventos, de los que más casos tenemos, y otros, categoría que contiene a los pacientes que no sufren un evento, los cuales son la mayoría, y otros tipos de eventos de los cuales se tienen pocos casos.

Tras diversas pruebas se obtiene la siguiente red neuronal, la cual cuenta con una estructura bastante más compleja y requiere de mayor poder computacional para calcularla. La red cuenta con dos capas ocultas, la primera con 20 neuronas, y la segunda con 9 neuronas, por otro lado, cuenta con 5 neuronas de output, una por cada categoría, y cada una de ellas devuelve la probabilidad de que el paciente sufra el correspondiente evento. La categoría con mayor probabilidad es la que se le asigna a cada paciente. Nuevamente todas las neuronas cuentan con la función sigmoide como función de activación. La estructura de la red es la siguiente.

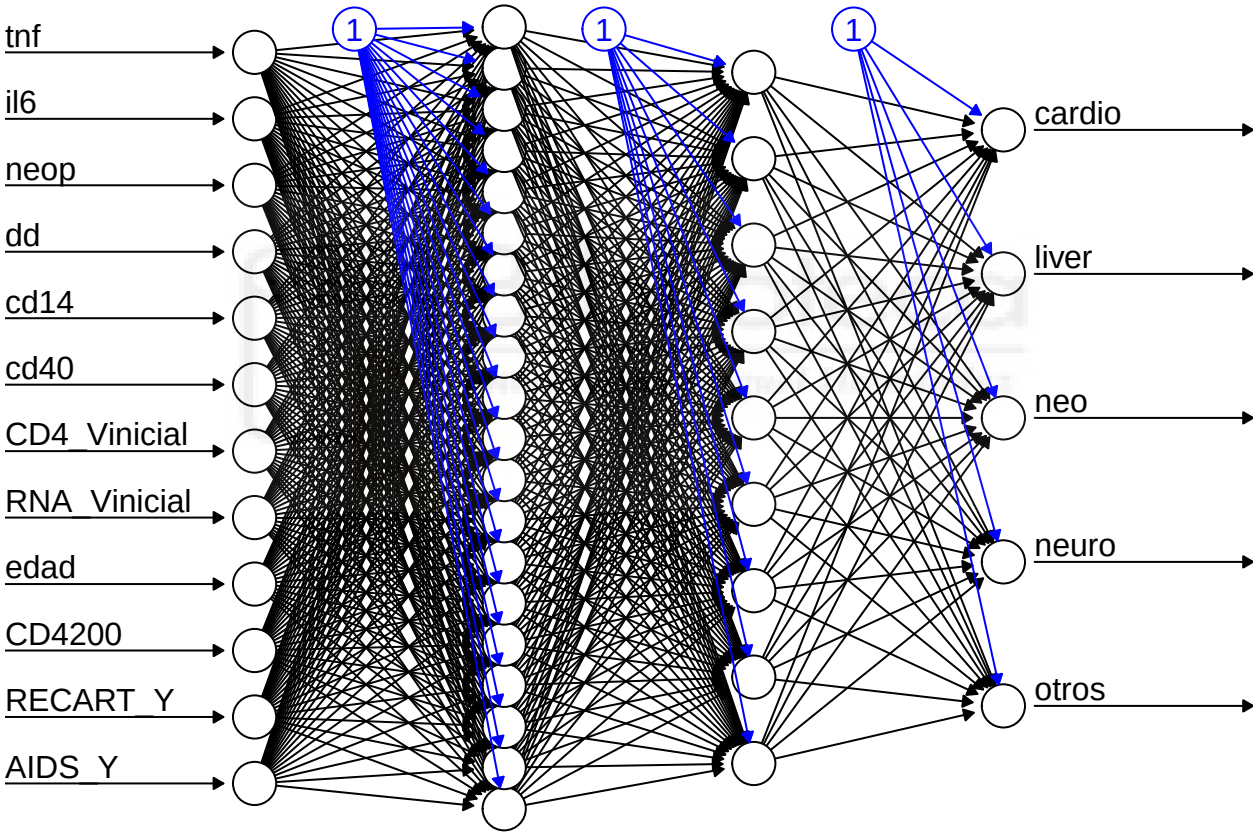


Figure 34: Red neuronal eventos.

Si evaluamos la capacidad predictora de la red, observamos resultados bastante buenos para la complejidad del problema, una precisión del 76,3% y un valor de Kappa del 53,9%. Observamos que las categorías que mejor predice la red son la de liver, neuro y otros, resultados similares a los obtenidos en un modelo anterior similar utilizando arboles de clasificación.

| | cardio | liver | neo | neuro | otros |
|---|---|---|---|---|---|
| cardio | 0 | 0 | 0 | 1 | 0 |
| liver | 2 | 8 | 0 | 1 | 2 |
| neo | 0 | 0 | 0 | 0 | 1 |
| neuro | 0 | 0 | 0 | 2 | 1 |
| otros | 0 | 1 | 1 | 3 | 32 |

| | Valor |
|---|---|
| Accuracy | 0.7636364 |
| Kappa | 0.5393041 |

Tabla 30: Matriz de confusión y metricas.

Como se comentaba anteriormente, es muy complicado extraer conclusiones sobre como afectan las variables a los resultados, pero si que se puede observar que si en la base de datos se contase con más casos de cada tipo de evento ENOS, se podrían obtener resultados bastante mejores, la red neuronal construida es un buen primer acercamiento a la resolución de un problema tan complejo como es la predicción de si un paciente va a sufrir un evento ENOS y de que tipo, o si no lo va a sufrir.

# 6 Conclusiones

A lo largo del presente estudio se han puesto en practica 4 de las técnicas estadísticas más relevantes en la actualidad, cada una de ellas con sus flaquezas y sus fortalezas, lo que causa que no todas las técnicas sean igual de efectivas para todos los casos. En este apartado se van a discutir los resultaos obtenidos, comparando los diferentes modelos y técnicas empleadas, para tratar de concluir cuales son las mejores técnicas para cada caso.

Se parte de la base de que el estudio ha tenido varias limitaciones, como usar únicamente el primer evento de cada paciente y el tiempo disponible para la realización del propio estudio, entre otras, lo que ha generado que no se pueda extraer el mayor potencial de cada una de las técnicas, por ejemplo, los arboles de clasificación se podrían mejorar poniendo en practica un random forest, el cual es muy similar a un árbol de clasificación, pero elimina muchos de los problemas encontrados a la hora de aplicar el árbol, es por esto que en el estudio se han dejado varias vías de investigación futura abiertas, algunas de las cuales se irán comentando a lo largo de este apartado.

Se pueden dividir las técnicas empleadas en dos categorías, por un lado, las que se han usado para comprender la influencia de ciertas variables sobre el resultado y para clasificar, como son la regresión logística, el árbol de clasificación y las redes neuronales, y, por otro lado, las que únicamente se han empleado para entender la influencia sobre una variable respuesta de el resto de las variables explicativas, como es el análisis de supervivencia. Para facilitar la comparación de las distintas técnicas y los casos en los que se han aplicado, se comparan por un lado las técnicas pertenecientes a la primera categoría, y el análisis de supervivencia por separado.

En la siguiente tabla (Tab. 31) se muestra un resumen de los resultados de los distintos modelos en cada una de las técnicas, siendo "P" la precisión, "K" el valor de Kappa, y denotando con una "x" si dicha técnica no existe dicho modelo.

| | Regresión Logística | | Árbol de Clasificación | | Redes Neuronales | |
|---|---|---|---|---|---|---|
| | P | K | P | K | P | K |
| Evento sí o no | 57% | 20% | 76% | 53% | 70% | 37% |
| CD4 < 200 | 83% | 23% | x | x | x | x |
| Muerte | 94% | 54% | x | x | 90% | 57% |
| Tipo de evento | 0% | 0% | 89% | 83% | 76% | 53% |

Tabla 31: Resumen de los resultados.

A rasgos generales se observa que la regresión logística (RL) es la que peores resultados ofrece, en comparación con los arboles de clasificación y las redes neuronales, ya que, aunque las precisiones sean altas, los valores de Kappa son bastante bajos. En el único caso en el que se obtienen resultados aceptables con RL es el modelo "Muerte", con un Kappa moderado y una precisión bastante elevada, pero incluso en ese caso, la red neuronal ofrece resultados ligeramente mejores. En el caso del modelo "Muerte", aunque se obtengan resultados ligeramente superiores con una red neuronal, la complejidad de estructurar una red neuronal no compensa la mejora que se obtiene. Sería interesante profundizar, de cara al futuro, en alguno método de suavizado para mejorar los resultados de la regresión logística. Los principales beneficios de la RL es lo simple que es de aplicar y lo comprensibles que son los resultados de cara a un publico no experto en la materia, pero por otro lado, como aspectos negativos, la RL no trabaja bien con datos perdidos ni outliers, y generalmente la RL no trabaja bien con pequeñas poblaciones y datos desequilibrados, lo que supone un problema, especialmente para este estudio, ya que una de las carencias que presenta el estudio es la falta de población y lo desequilibrado que está la cantidad de paciente en cuanto a distintas categorías estudiadas. Este tipo de carencias se denotan de forma muy clara en el modelo "Tipo de evento", el cual es una RL que clasifica a los individuos en dos grupos, los que sufren un evento concreto frente al resto, pero, como se ha observado a lo largo del estudio, de cada tipo de evento existen muy pocos pacientes, por lo que se obtienen

resultados muy malos. Otra posible vía de investigación futura en relación con el modelo "Tipo de evento" es el uso de una regresión logística multinomial, la cual funciona de igual forma que la RL estudiada, pero permite que el output sea más de dos categorías.

Por otro lado, en cuanto a arboles de clasificación, es con la técnica con la que, por lo general, mejores resultados se han obtenido. Únicamente hay dos arboles de clasificación, "Evento sí o no", que clasifica a los individuos en función de si desarrollarán un evento o no, y "Tipo de evento", el cual, a diferencia de la regresión logística, no clasifica en un evento concreto o el resto, lo que hace es clasificar a los individuos en función del tipo de evento que desarrollará. Ambos modelos ofrecen los mejores resultados en función del resto de técnicas para el mismo problema. Los puntos fuertes de un árbol de clasificación son lo bien que trabaja con poca cantidad de datos y con datos desequilibrados, a diferencia de la regresión logística, y lo comprensible que son los resultados, generando un diagrama lógico que hace muy fácil su comprensión para un lector no experto en la materia. Por otro lado, uno de los principales puntos débiles de un árbol de clasificación es la falta de datos numéricos sobre la influencia de cada una de las variables utilizadas en el mismo, en comparación con, por ejemplo, con la RL, la cual ofrece valores numéricos no relativos sobre la importancia y como afecta cada una de las variables sobre el resultado. Una de las posibles vías futuras de investigación relacionada con los arboles de clasificación son los random forest, los cuales generan una cantidad dada de arboles de clasificación y calculan una especie de árbol "medio", con lo cual se consiguen mejores resultados a la hora de datos incompletos o muy complejos.

Finalmente, respecto a los resultados obtenidos con las redes neuronales, dos de los modelos presentan resultados aceptables, con precisiones altas y valores de Kappa moderados. Los puntos positivos de las redes neuronales es que son muy flexibles y pueden trabajar con datos muy complejos y encontrar patrones en estos que otras técnicas no pueden, pero por otro lado, como puntos negativos, son bastante mas complejas de entender y aplicar que el resto de las técnicas, con prácticamente infinitas estructuras posibles para resolver un mismo problema, computacionalmente son mucho mas pesadas y lentas que el resto de técnicas y finalmente, mientras que en las otras técnicas tienes más o menos información sobre los patrones encontrados en los datos y como afectas las variables al resultado, con las redes neuronales se pierde esta información y pasas a tener únicamente un input y un output. Queda como vía de investigación futura el profundizar más ampliamente en los diferentes parámetros que se pueden modificar y probar diferentes estructuras, ya que la principal limitación que se ha encontrado en la aplicación de las redes neuronales ha sido el tiempo.

En la siguiente tabla (Tab. 32) se muestra un resumen de los resultados de los distintos modelos formulados usando el análisis de supervivencia, en concreto el $R^2$ de los modelos de Cox.

| | R^2 |
|---|---|
| Desde la entrada hasta la muerte | 59% |
| Desde la entrada hasta el evento | 10% |
| Desde el evento hasta la muerte | 75% |

Tabla 32: Resumen de los resultados.

En primer lugar, se han definido tres periodos de estudio distintos. Desde que el paciente entra en el estudio, hasta el momento de la muerte, hasta el momento en el que desarrolla un evento y desde que desarrolla un evento hasta el momento en el que el paciente muere.

Como se puede observar, con los diferentes modelos de Cox, por lo general, se explica más variabilidad cuando el evento que se estudia es la muerte, hasta un 75% cuando se estudia desde el evento hasta el momento de la muerte. El análisis de supervivencia son un conjunto de diferentes técnicas estadísticas, muy comúnmente utilizadas en la bioestadística, que tienen en cuenta el tiempo que tarda un evento en suceder, por tanto, es muy interesante su uso dentro de estudios que se extienden en el tiempo, como es el caso del presente estudio. Como vía futura de investigación existen los modelos de riesgos competitivos, los cuales pueden tener en cuenta diferentes tipos de eventos, como por ejemplo distintas causas de muerte o distintos eventos ENOS.

Como conclusión, en primer lugar, la regresión logística es una técnica muy potente, pero que necesita que los datos sean demasiado "perfectos" para poder obtener buenos resultados, lo cual no siempre es el caso, en segundo lugar, el árbol de clasificación no necesita que los datos sean tan "perfectos como en la regresión logística para dar buenos resultados, pero no se tiene tanta información sobre la influencia de las variables sobre el resultado final, a diferencia de otras técnicas. En tercer lugar, Las redes neuronales son una de las técnicas más novedosas, potentes y sobre la cual se investiga para mejorarla a diario, es capaz de detectar patrones en los datos que otras técnicas no pueden, pero esto viene con dos principales problemas, es una técnica un tanto mas compleja que las anteriores, tanto de entender como de aplicar, y es un modelo de"caja negra", no se sabe muy bien las relaciones que se hacen entre las variables ni como estas afectan al resultado final. Finalmente, en cuarto lugar, El análisis de supervivencia, en concreto los modelos de Cox son una buena herramienta cuando el tiempo transcurrido entre dos momentos de un estudio es relevante para el resultado, es capaz de ofrecer información clara sobre como y en que medida afecta cada variable al resultado final, por lo que, normalmente, es utilizado para entender como se relacionan las variables, más que para clasificar a una serie de individuos, pero como punto negativo, no es una técnica tan potente como otras existentes y requiere de una serie de asunciones que no siempre se cumplen.

# 7 Conocimientos adquiridos

A lo largo de la carrera, muchas veces no es posible profundizar en el funcionamiento a nivel interno de las distintas técnicas, por tanto, todo el proceso de buscar información, comprenderla y tener que redactar una explicación sobre cada una de las técnicas utilizadas me ha servido para desarrollar la capacidad de comprender con relativa facilidad una gran cantidad de técnicas por mi cuenta. También este proceso de búsqueda y recopilación de información sobre las técnicas me ha ayudado a saber donde buscar información fiable, conocer algunos libros de referencia en el campo de la estadística, etc.

Por otro lado, la redacción y estructuración del presente estudio me ha ayudado a entender como se debe redactar un informe, la información que hay que incluir y como redactar conceptos estadísticos de forma comprensible, entre otras cosas, lo que será de gran ayuda una vez entre en el mercado laboral.

Finalmente, de las técnicas utilizadas en el presente estudio, a lo largo de la carrera, únicamente se estudia la regresión logística, por tanto, el tener que estudiar las diferentes técnicas que no había visto a lo largo de la carrera me ha servido para rellenar ciertas carencias curriculares que podría tener de cara a un futuro.

# 8 Anexos

## 8.1 Código general

```
#Librerias utilizadas.
library(readr)
library(ROCR)
library(tidyverse)
library(caret)
library(MASS)
library(dplyr)
library(e1071)
library(stargazer)
library(rpart)
library(rpart.plot)
library(knitr)
library(randomForest)
library(arsenal)
library(knitr)
library(texreg)
library(kableExtra)

#Lectura base de datos.
base <- read_csv("base.csv", col_types = cols(X1 = col_skip()))

#creación de variables necesarias.
base$eventosrecod1<-recode(base$eventsENOS, bone = "others",
                           metabolic = "others", noevento = "others", renal = "others")

#Creación de bases de datos "train" para ajuste de modelos y
#"test" para evaluación de los mismos.
set.seed(1234)
train <- sample_frac(base, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base[-sample_id,]
```

## 8.2 Código Modelos.

### 8.2.1 Regresión logísitca.

#### 8.2.1.1 Evento o no.

```
#Calculo modelos.
mod1 <- glm(evenono ~ CD4CD8inicial + edad + vhc + RECART_Y ,
            data = train, family = "binomial")

mod2 <- glm(evenono ~ EDU_LVL + vhc + edad + V39 + `RNA_V inicial` + `CD4_V inicial` +
              bnp + isopros + cd14 + CD4200, data = train, family = "binomial")

#Calculo de Odds ratio.
coef.vector <- list(exp(mod$coefficients), exp(mod3$coefficients))
```

```r
#Evaluación modelo 1.
test$precticed <- predict(mod1, newdata = test)
test$precticed <- ifelse(test$precticed > 0.5, 1, 0)

m1<-confusionMatrix(as.factor(test$precticed), as.factor(test$evenono))

#Evaluación modelo 2.
test$precticed <- predict(mod2, newdata = test)
test$precticed <- ifelse(test$precticed > 0.5, 1, 0)

m2<-confusionMatrix(as.factor(test$precticed), as.factor(test$evenono))
```

#### 8.2.1.2 CD4 < 200

```r
#Calculo de modelos.
mod1 <- glm(as.factor(CD4200) ~ (scale(neop) + scale(mda) + scale(cd14) +
              as.factor(cc.ev)), data = train, family ="binomial")

mod2 <- glm(as.factor(CD4200) ~ (scale(neop) + scale(mda)  + as.numeric(scale(edad))  +
           as.factor(cc.ev)), data = train, family ="binomial"( link = "probit"))

#Calculo Odds ratio.
coef.vector <- list(exp(mod1$coefficients), exp(mod2$coefficients))

#Evaluación modelo 1.
test$precticed <- predict(mod1, newdata = test)
test$precticed<- ifelse(test$precticed > 0.5, 1, 0)

m<-confusionMatrix(as.factor(test$precticed), as.factor(test$CD4200))

#Evaluación modelo 2.
test$precticed <- predict(mod2, newdata = test)
test$precticed <- (test$precticed - min(test$precticed))/(max(test$precticed) -
                                                          min(test$precticed))
test$precticed<- ifelse(test$precticed > 0.5, 1, 0)

mat<- confusionMatrix(as.factor(test$precticed), as.factor(test$CD4200))
```

#### 8.2.1.3 Muerte

```r
#Calculo de modelos.
mod1 <- glm(as.factor(DEATH_Y) ~ (edad  + `CD4_V inicial`*as.factor(CD4200) +
            as.factor(AIDS_Y)  ), data = train, family ="binomial"( link = "probit"))


mod2 <- glm(as.factor(DEATH_Y) ~ (edad  + `CD4_V inicial`*as.factor(CD4200) +
             as.factor(AIDS_Y) + as.factor(AIDS_Y)*edad +
             `CD4_V inicial`*as.factor(AIDS_Y)), data = train,
            family ="binomial"( link = "probit"))

mod3 <- glm(as.factor(DEATH_Y) ~ (edad  + `CD4_V inicial`*as.factor(CD4200) +
```

```
              as.factor(AIDS_Y) + as.factor(AIDS_Y)*edad +
              `CD4_V inicial`*as.factor(AIDS_Y) + as.factor(RECART_Y)*edad +
              as.factor(RECART_Y)*`CD4_V inicial`), data = train,
              family ="binomial"( link = "probit"))

#Calculo Odds ratio.
coef.vector <- list(exp(mod1$coefficients),exp(mod2$coefficients),exp(mod3$coefficients))

#Evaluación modelo 1.
test$precticed <- predict(mod1, newdata = test, type = "response")
test$precticed<- ifelse(test$precticed > 0.5, 1, 0)
mat<- confusionMatrix(as.factor(test$precticed),as.factor(test$DEATH_Y))
```

#### 8.2.1.4 Evento vs resto de eventos.

```
#Creación de variables necesarias para los dos siguientes modelos.
liverotro<- ifelse(base$eventsENOS=="liver", 1, 0)
neurootro<- ifelse(base$eventsENOS=="neuropsychiatric", 1, 0)
base$liverotro <- liverotro
base$neurootro <- neurootro

#Actualización base de datos "train" y "test"
#con la nuevas variables.
set.seed(1234)
train <- sample_frac(base, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base[-sample_id,]

#Calculo de modelos.
modliver <- glm(as.factor(liverotro) ~ (as.factor(CD4200) + mda + `CD4_V inicial`),
                data = train, family ="binomial")

modneuro <- glm(as.factor(neurootro) ~ (edad + `CD4_V evento` + `CD4_V inicial` + mda),
                data = train, family ="binomial")

#Evaluación modelo liver.
test$precticed <- predict(modliver, newdata = test)
test$precticed<- ifelse(test$precticed > 0.5, 1, 0)

mat<- confusionMatrix(as.factor(test$precticed), as.factor(test$liverotro))
coef.vector <- exp(modliver$coefficients)

#Evaluación modelo neuro.
test$precticed <- predict(modneuro, newdata = test)
test$precticed<- ifelse(test$precticed > 0.5, 1, 0)

mat<- confusionMatrix(as.factor(test$precticed), as.factor(test$neurootro))
coef.vector <- list(exp(modneuro$coefficients))
```

### 8.2.2 Analisis de supervivencia

#### 8.2.2.1 Desde entrada hasta muerte

```
#Calculo variables de tiempo necesarias
DDD<- ifelse(base$DROP_Y == "1", 1, ifelse(base$DEATH_Y== "1", 2, 1))
FFF<- if_else(base$DROP_Y == "1", base$DROP_D, if_else(base$DEATH_Y=="1",
                                                       base$DEATH_D, base$L_ALIVE))

base$status <- DDD
base$FFF <- FFF

tiempo<- base$FFF - base$ENROL_D
base$tiempo <- tiempo

#Calculo modelo de Kaplan-Meier
ggsurvplot(surv_fit(Surv(tiempo, status) ~  evenono + CD4200, data=base), data=base,
           pval = TRUE, conf.int = TRUE, risk.table = TRUE,surv.median.line = "hv",
           ggtheme = theme_bw(),ncensor.plot = TRUE, fontsize = 2)
```

```
#Calculo modelo de cox
res.cox <- coxph(Surv(tiempo, status) ~ (il8+neop+mda+RECART_Y+edad) , data = base)
coef.vector <- exp(res.cox$coefficients)
```

#### 8.2.2.2 Desde entrada hasta evento

```
#Creación de variables de tiempo necesarias
DDD<- ifelse(base$DROP_Y == "1", 1, ifelse(base$evenono== "1", 2, 1))
FFF<- if_else(base$DROP_Y == "1", base$DROP_D, if_else(base$evenono=="0",
                                                       base$CEP_D, base$L_ALIVE))

base$status <- DDD
base$FFF <- FFF

tiempo<- base$FFF - base$ENROL_D
base$tiempo <- tiempo

#Calculo modelo de Kaplan-Meier
ggsurvplot(surv_fit(Surv(tiempo, status) ~ RECART_Y , data=base), data=base, pval = TRUE,
           conf.int = TRUE, risk.table = TRUE,surv.median.line = "hv",
           ggtheme = theme_bw(),ncensor.plot = TRUE, fontsize = 2)
```

```
#Calculo modelo de Cox
res.cox <- coxph(Surv(tiempo, status) ~ (neop+RECART_Y+CD4200+vhc), data = base)
coef.vector <- exp(res.cox$coefficients)
```

#### 8.2.2.3 Desde evento hasta muerte

```
#Creación de variables de tiempo necesarias.
DDD<- ifelse(base$DROP_Y == "1", 1, ifelse(base$DEATH_Y== "1", 2, 1))
FFF<- if_else(base$DROP_Y == "1",
              base$DROP_D, if_else(base$DEATH_Y=="1", base$DEATH_D, base$L_ALIVE))

base$status <- DDD
base$FFF <- FFF
```

```
tiempo<- base$FFF - base$CEP_D
base$tiempo <- tiempo

#Calculo modelo de Kaplan-Meier
ggsurvplot(surv_fit(Surv(tiempo, status) ~ eventosrecod1, data=base), pval = TRUE,
           conf.int = TRUE, risk.table = TRUE, risk.table.col = "strata",
           linetype = "strata", surv.median.line = "hv",
           ggtheme = theme_bw(), ncensor.plot = TRUE,)
```

```
#Calculo modelos de Cox
res.cox <- coxph(Surv(tiempo, status) ~ il8+neop+dd+mda+cd40+edad, data = base)

res.cox2 <- coxph(Surv(tiempo, status) ~ eventosrecod1, data = base)

coef.vector <- list(exp(res.cox$coefficients), exp(res.cox2$coefficients))
```

### 8.2.3 Arboles de clasificación.

#### 8.2.3.1 Evento o no.

```
#Eliminamos las variables que no queremos incluir en el arbol.
train1<- train[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                  -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
                  -81,-72,-73,-86, -78,-85, -121, -113, -49, -98,-89, -109, -36,-41,-93,
                  -104, -123, -120, -117,-107, -88, -60)]

test1<- test[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
                -81,-72,-73,-86, -78, -85, -121,-113,-49, -98,-89, -109,-36, -41,-93,
                -104, -123, -120, -117,-107,-88,-60)]

#Calculamos el arbol.
set.seed(1234)
arbol<- rpart(as.factor(evenono) ~ .,method='class' , data= train1)

#Podamos el arbol.
arbol<- prune(arbol, cp =0.016332)

#Evaluamos capacidad clasificadora.
pred <- predict(arbol, newdata = test1, type="class")
mat<- confusionMatrix(as.factor(test$evenono), as.factor(pred))
```

#### 8.2.3.2 Eventos 1.

```
#Eliminamos registros no deseados.
train3 <- train[ which(train$eventosrecod1!="others"),]
test3 <- test[ which(test$eventosrecod1!="others"),]

#Eliminamos variables que no queremos incluir en el arbol.
train2<- train3[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                   -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
```

```
                  -81,-72,-73,-86, -78, -85,-35, -121,-113,-49, -98,-89, -109,-36, -41,
                  -93,-104, -120, -117, -43,-45)]

test2<- test3[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                 -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
                 -81,-72,-73,-86, -78, -85,-35, -121,-113,-49, -98,-89, -109,-36, -41,
                 -93,-104, -120, -117, -43,-45)]

#Calculamos el arbol
set.seed(1234)
arbol2<- rpart(as.factor(eventosrecod1) ~ .  ,method='class' , data= train2)

#Evaluamos capacidad clasificadora.
pred <- predict(arbol2, newdata = test2, type="class")
mat<- confusionMatrix(as.factor(pred), as.factor(test2$eventosrecod1))
```

#### 8.2.3.3 Eventos 2.

```
#Eliminamos variables que no queremos incluir en el modelo
train2<- train[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                  -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
                  -81,-72,-73,-86, -78, -85,-35, -121,-113,-49, -98,-89, -109,-36, -41,
                  -93,-104, -120, -117, -43,-45,-118, -107)]

test2<- test[,c(-1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-70,-11, -112, -116 ,-44, -56, -54, -48,
                -34, -97, -105, -108, -92, -111, -103, -71, -106, -115, -50, -52, -38,
                -81,-72,-73,-86, -78, -85,-35, -121,-113,-49, -98,-89, -109,-36, -41,
                -93, -104, -120, -117, -43,-45,-118, -107)]

#Calculamos el arbol
set.seed(1234)
arbol2<- rpart(as.factor(eventosrecod1) ~ .  ,method='class' , data= train2)

#Evaluamos capacidad clasificadora.
pred <- predict(arbol2, newdata = test2, type="class")
mat<- confusionMatrix(as.factor(pred), as.factor(test2$eventosrecod1))
```

### 8.2.4 Redes neuronales

#### 8.2.4.1 Evento o no

```
#Selección de variables y limpieza de las mismas
set.seed(1234)
base1<-na.omit(base[,c(13:14,17, 20,23, 26, 110, 99, 31,119,84,87,118)])


train <- sample_frac(base1, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base1[-sample_id,]

train1<-data.frame(scale(train[,c(-13)]))
test1<-data.frame(scale(test[,c(-13)]))
```

```
train1<- data.frame(train1, train$evenono)
test1<- data.frame(test1, test$evenono)
#pcr+tnf+il6+il8+neop+dd+isopros+mda+cd14+c163+cd40


#Calculo red neuronal
n <- names(train1)
f <- as.formula(paste("medv ~", paste(n[!n %in% "medv"], collapse = " + ")))

set.seed(1234)
nn <- neuralnet(train.evenono ~  . ,data=train1,hidden=c(5,5),linear.output=F,
                learningrate = 0.1)

#Evaluación red neuronal
pr.nn <- compute(nn,test1)

res<-ifelse(pr.nn$net.result >0.5, 1,0)

confusionMatrix(as.factor(res), as.factor(test1$test.evenono))
plot(nn)
```

```
#Selección de variables y limpieza de las mismas
base1<-na.omit(base[,c(13:14,17, 20,23, 26, 110, 99, 31,119,84,87,39,118)])

set.seed(1234)
train <- sample_frac(base1, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base1[-sample_id,]

train1<-data.frame(scale(train[,c(-14)]))
test1<-data.frame(scale(test[,c(-14)]))
train1<- data.frame(train1, train$evenono)
test1<- data.frame(test1, test$evenono)

#Calculo red neuronal
n <- names(train1)
f <- as.formula(paste("medv ~", paste(n[!n %in% "medv"], collapse = " + ")))

set.seed(1234)
nn <- neuralnet(train.evenono ~  . ,data=train1,hidden=c(5,2, 2),linear.output=F,
                learningrate = 0.1)

#Evaluación red neuronal
pr.nn <- compute(nn,test1)

res<-ifelse(pr.nn$net.result >0.5, 1,0)

confusionMatrix(as.factor(res), as.factor(test1$test.evenono))
plot(nn)
```

#### 8.2.4.2 Muerte

```
#Selección de variables y limpieza de las mismas
set.seed(1234)

base1<-na.omit(base[,c(13:14,17, 20,23, 26, 110, 99, 31,119,84,87,54)])


train <- sample_frac(base1, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base1[-sample_id,]

train1<-data.frame(scale(train[,c(-13)]))
test1<-data.frame(scale(test[,c(-13)]))
train1<- data.frame(train1, train$DEATH_Y)
test1<- data.frame(test1, test$DEATH_Y)
#pcr+tnf+il6+il8+neop+dd+isopros+mda+cd14+c163+cd40

#Calculo red neuronal

n <- names(train1)
f <- as.formula(paste("medv ~", paste(n[!n %in% "medv"], collapse = " + ")))

set.seed(1234)
nn <- neuralnet(train.DEATH_Y ~  . ,data=train1,hidden=c(2),linear.output=F,
                learningrate = 0.1)

#Evaluación red neuronal
pr.nn <- compute(nn,test1)

res<-ifelse(pr.nn$net.result >0.5, 1,0)

confusionMatrix(as.factor(res), as.factor(test1$test.DEATH_Y))
plot(nn)
```

#### 8.2.4.3 Tipo Evento

```
#Selección de variables y limpieza de las mismas

set.seed(1234)

base1<-na.omit(base[,c(13:14,17, 20,23, 26, 110, 99, 31,119,84,87,123)])


library(fastDummies)

base1<-dummy_cols(base1, select_columns = "eventosrecod1")




train <- sample_frac(base1, 0.9)
sample_id <- as.numeric(rownames(train))
test <- base1[-sample_id,]
```

```
train1<-data.frame(scale(train[,c(-13:-18)]))
test1<-data.frame(scale(test[,c(-13:-18)]))
train1<- data.frame(train1, train$eventosrecod1_cardiovascular, train$eventosrecod1_liver,
                    train$eventosrecod1_neoplastic, train$eventosrecod1_neuropsychiatric,
                    train$eventosrecod1_others)
test1<- data.frame(test1, test$eventosrecod1_cardiovascular, test$eventosrecod1_liver,
                   test$eventosrecod1_neoplastic, test$eventosrecod1_neuropsychiatric,
                   test$eventosrecod1_others)
#pcr+tnf+il6+il8+neop+dd+isopros+mda+cd14+c163+cd40

colnames(train1)[13:17]<-c("cardio", "liver", "neo", "neuro", "otros")
colnames(test1)[13:17]<-c("cardio", "liver", "neo", "neuro", "otros")


#Calculo red neuronal
n <- names(train1)
f <- as.formula(paste("medv ~", paste(n[!n %in% "medv"], collapse = " + ")))

set.seed(1234)
nn <- neuralnet(cardio+liver+neo+neuro+otros ~  . ,data=train1,hidden=c(20, 9),
                linear.output=F, learningrate = 0.01)

#Evaluación red neuronal
pr.nn <- compute(nn,test1)

res<-ifelse(pr.nn$net.result >0.5, 1,0)

res1<-ifelse(res[,1]==1, "cardio", ifelse(res[,2]==1, "liver", ifelse(res[,3]==1,
                                  "neo", ifelse(res[,4]==1, "neuro", "otros"))))

res2<-ifelse(test1[,13]==1, "cardio", ifelse(test1[,14]==1, "liver", ifelse(test1[,15]==1,
                                    "neo", ifelse(test1[,16]==1, "neuro", "otros"))))

confusionMatrix(table(as.factor(res1), as.factor(res2)))
plot(nn)
```

# References


[1] A. Nogales Espert A. Panero López. "Neopterina como marcador de la activación inmune en infecciones víricas y bacterianas." In: *AEPED* 45.6 (1996).

[2] *Acerca del VIH/SIDA*. URL: https://www.cdc.gov/hiv/spanish/basics/whatishiv.html.

[3] Marta Benito Gutíerrez Ana Pilar Roca Susana Riesco Riesco. "Utilidad del dímero D como marcador analítico en urgencias pediátricas". In: *Journal of the spanish Society of Emergency Medicine* (2008).

[4] José D. Martín Antonio J. Serrano Emilio Soria. "Redes Neuronales Artificiales". Universitat de València. 2009.

[5] LEO BREIMAN. *CLASSIFICATION AND REGRESSION TREES*. CRC Press, 2017.

[6] Juan Miguel Marín Diazaraque. "Introducción a las redes neuronales aplicadas". Universidad Carlos III de Madrid.

[7] *Diccionario de cáncer*. URL: https://www.cancer.gov/espanol/publicaciones/diccionario.

[8] Annette J. Dobson and Adrian G. Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[9] T. Español and S. Sauleda. "Respuesta inmunológica frente a la infección por el VIH". In: *Archivos de Bronconeumología* 28.1 (1992), pp. 27–31. DOI: 10.1016/s0300-2896(15)31385-5.

[10] Domingo Morales González. "Modelos lineales generalizados". 2018.

[11] Brian B. Hart et al. "Inflammation-Related Morbidity and Mortality Among HIV-Positive Adults". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 77.1 (2018), pp. 1–7. DOI: 10.1097/qai.0000000000001554.

[12] Fernando Gil Hernández. "El papel de los biomarcadores en Toxicología Humana". PhD thesis. Facultado de Medicina de la Universidad de Granada, 2011.

[13] *Hoja informativa - Últimas estadísticas sobre el estado de la epidemia de sida*. URL: https://www.unaids.org/es/resources/fact-sheet.

[14] Denise C Hsu, Irini Sereti, and Jintanat Ananworanich. "Serious Non-AIDS events: Immunopathogenesis and interventional strategies". In: *AIDS Research and Therapy* 10.1 (2013), p. 29. DOI: 10.1186/1742-6405-10-29.

[15] J.I. Pérez Calvo I. torres Courchoud. "Biomarcadores y práctica clínica". In: *Editorial* (2016).

[16] Joseph G. Ibrahim John P. Klein Hans C. van Houwelingen and Thomas H. Scheike. *Handbook of survival Analysis*. CRC Press, 2019.

[17] David G. Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text*. Springer, 2012.

[18] Mar Masiá et al. "Contribution of Oxidative Stress to Non-AIDS Events in HIV-Infected Patients". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 75.2 (2017). DOI: 10.1097/qai.0000000000001287.

[19] M L Munier and A D Kelleher. "Acutely dysregulated, chronically disabled by the enemy within: T-cell responses to HIV-1 infection". In: *Immunology and Cell Biology* 85.1 (May 2006), pp. 6–15. DOI: 10.1038/sj.icb.7100015.

[20] Clínica Universidad de Navarra. *Diccionario médico*. URL: https://www.cun.es/diccionario-medico.

[21] Abdelmalik Moujahid Pedro Larrañaga Iñaki Inza. "Redes Neuronales". Departamento de Ciencias de la Computación e Inteligencia Artificial.

[22] Max Roser and Hannah Ritchie. *HIV / AIDS*. Apr. 2018. URL: https://ourworldindata.org/hiv-aids.

[23] María Teresa Rugeles-López et al. "Biomarcadores inmunológicos de riesgo cardiovascular en la infección por el virus de inmunodeficiencia humana-1". In: *Revista Colombiana de Cardiología* 24.2 (2017), pp. 153–160. DOI: 10.1016/j.rccar.2016.05.012.

[24] Hidalgo Rivas A. y Sánchez Astorga M. Salvatierra Cáceres E. Salinas Rodríguez J. "Diagnostic capability of the salivary biomarkers interleukin 6 and 8 for diagnosis of oral squamous cell carcinoma". In: *Av Odontoestomatol* 33.2 (2017).

[25] Ronald M. Schrader. “Logistic Regression - Interpreting Parameters”.

[26] Sidalava. *La historia del VIH, en una línea de tiempo: hitos que marcaron la evolución de la enfermedad*. Mar. 2019. URL: https://www.sidalava.org/la-historia-del-vih-en-una-linea-de-tiempo-hitos-que-marcaron-la-evolucion-de-la-enfermedad/.

[27] Pedro Suarez. *Isoprostanos: química, clasificación, genética, inmunología.* Dec. 2018. URL: https://decs.es/compuestos-quimicos-y-drogas/isoprostanos.1/.

[28] Pang-Ning Tan et al. *Introduction to data mining*. Pearson, 2020.

[29] *VIH*. URL: http://gtt-vih.org/aprende/informacion_basica_sobre_el_vih/que_son_los_cd4.

[30] *VIH y sida: MedlinePlus en español*. URL: https://medlineplus.gov/spanish/hivaids.html.

[31] Alberto J. Núñez Sellés. Wilfredo Mañon Rossi Gabino Garrido. “Biomarcadores del estrés oxidativo en la terapia antioxidante”. In: *Journal of Pharmacy and Pharmacognosy Research* (Mar. 2016).