Aplicación de métodos de aprendizaje automático para el estudio del análisis de supervivencia en pacientes infectados por el VIH

Luis Navarrete Bellot
Máster en Bioinformática y Bioestadística
Área 2, subárea 2: Análisis de datos

Nuria Pérez Álvarez
Carles Ventura Royo

Junio de 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Aplicación de métodos de aprendizaje automático para el estudio del análisis de supervivencia en individuos infectados por el VIH*
Nombre del autor:	*Luis Navarrete Bellot*
Nombre del consultor/a:	*Núria Pérez Álvarez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	07/2020
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 2 – Subárea 2: Análisis de datos*
Idioma del trabajo:	*Español*
Palabras clave	*VIH/SIDA, Análisis de supervivencia, Aprendizaje Automático (ML)*

Resumen del Trabajo (máximo 250 palabras):

El análisis de supervivencia tiene como objetivo analizar y modelar datos donde el resultado es el tiempo hasta que ocurre un evento de interés. Uno de los principales desafíos en este contexto es la presencia de instancias cuyos resultados de eventos se vuelven inobservables después de un cierto momento, sea porque no se hace un seguimiento suficientemente largo o porque no presentaran el evento estudiado (llamado censura). Actualmente se están desarrollando muchos algoritmos de aprendizaje automático adaptados para analizar datos censurados. Así pues, se define como objetivo del TFM estudiar los métodos existentes de aprendizaje automático en el contexto descrito, buscar el/los métodos/s más adecuado/s para estudiar el tiempo hasta fracaso al tratamiento antirretroviral en una cohorte de pacientes infectados por el VIH.

Este proyecto busca crear herramientas de descripción que permitan resumir la información que contienen los datos y generar un modelo que nos ayude a clasificar a los pacientes según la distribución del tiempo hasta el fracaso virológico. Para ello, se tiene la base de datos “dsurv” del estudio “Instinct”, que contiene 342 variables de 995 pacientes que han sido seguidos y evaluados en diferentes momentos de tiempo, dando lugar a una base desbalanceada y heterogénea.

En la memoria analizaremos cómo el tipo de tratamiento que recibe el paciente y la realización de las pruebas de mutaciones en la integrasa, la proteasa y la transcriptasa

inversa pueden influir en el fracaso virológico, o en el aumento de la probabilidad de supervivencia a lo largo del tiempo.

Abstract (in English, 250 words or less):

Survival analysis aims to analyze and model data where the result is the time until an event of interest occurs. One of the main challenges in this context is the presence of instances whose event results become unobservable after a certain moment, either because there is not a long enough follow-up or because they did not present the studied event (called censorship). Many adapted machine learning algorithms are currently being developed to analyze censored data. Thus, the objective of the TFM is defined to study the existing methods of machine learning in the described context, to find the most suitable method / s to study the time to failure of antiretroviral treatment in a cohort of patients infected with HIV.

This project seeks to create description tools that allow summarizing the information contained in the data and generating a model that helps us classify patients according to the distribution of time until virological failure. To do this, we have the “dsurv” database from the “Instinct” study, which contains 342 variables from 995 patients who have been followed and evaluated at different points in time, giving rise to an unbalanced and heterogeneous base.

In the project, we will analyze how the type of treatment that the patient receives and the performance of tests for mutations in integration, protection and reverse transcription can influence virological failure, or increase the probability of survival over time.

INDICE

…/…/…

Lista de figuras

Lista de tablas

1. INTRODUCCIÓN

En este primer apartado se da información relevante acerca del VIH, sus orígenes, cómo evoluciona la enfermedad en la actualidad y últimos avances médicos y legales.

1.1. Historia del VIH/SIDA

1.1.1. El comienzo de la pandemia del VIH/SIDA

Según StopVIH, debido a la forma explosiva con que apareció la epidemia en 1981, muchos científicos pensaron que se enfrentaban a una "nueva enfermedad" provocada por un nuevo agente infeccioso inédito. Hubo quienes dijeron que se trataba de un "virus maquinado" por la CIA o la KGB, sin embargo, los orígenes del SIDA, pese a que no están totalmente claros, son muy diferentes.

La epidemia es reciente, los primeros casos notificados en África y Europa datan de 1959, y en Norteamérica de 1968. Los virus que causan el SIDA, el VIH-1 y el VIH-2, podrían constituir una invención antigua de la naturaleza, según sospechan los biólogos; pertenecen a la familia de los retrovirus, virus cuya información genética está contenida en una molécula de ARN en vez de ADN, como es habitual en la mayoría de los seres vivos.

En los últimos años, los científicos han identificado en simios africanos numerosos virus emparentados con los VIH, aunque no provocan ninguna enfermedad grave salvo en el caso de los macacos. Desde el punto de vista genético, los dos virus que causan el SIDA están más próximos a algunos de ellos que entre sí mismos.

El VIH-2, endémico en África occidental y que se ha propagado principalmente a la India y Europa, está emparentado con el virus *SIVsm* del mono *mangabey* y el *SIVmac* del macaco. Y el pariente simiesco más próximo del VIH-1, endémico en África central, desde donde se ha extendido a América del Norte y Europa, es el virus *SIVcpz* del chimpancé. El estudio de los genes de estos virus y de otros hallados más recientemente,

como el *SIVmnd* del mandril y el *SIVagm* del mono verde, apunta a que todos ellos se separaron al mismo tiempo de un antecesor común.

El descubrimiento en 1989 del *VIH-2 ALT*, una variante del *VIH-2* más remota que algunos *SIV*, parece indicar que el virus que causa el SIDA existía en el continente africano casi un siglo antes de que surgiera la epidemia.

1981. En junio, el Centro para el Control de Enfermedad de Atlanta, Estados Unidos (CDC), publica el primer reportaje sobre un tipo raro de neumonía: «Pneumocistis Carinii» en cinco jóvenes, todos homosexuales activos residentes en Los Ángeles. No se frecuentaban entre ellos, no tenían amigos comunes y no tenían conocimiento de enfermedades similares entre sus compañeros sexuales. Dos de ellos informaron haber mantenido relaciones homosexuales con diversas personas.

La creencia inicial de limitar esta enfermedad solo a homosexuales llevó a algunos autores a denominarla «Síndrome de Inmunodeficiencia relacionada con los homosexuales» o «Peste Rosa». En agosto del mismo año, 111 casos similares fueron reportados al CDC, lo que llevó a organizar un registro nacional de casos. Luego, nuevos casos de SIDA fueron descritos en drogadictos haitianos, hemofílicos, pacientes transfundidos, hijos de madres en riesgo, parejas heterosexuales y trabajadores de la salud, lo que hizo a la comunidad médica y a la sociedad, tomar conciencia de la existencia de una nueva epidemia sin precedentes en la historia de la medicina.

1982. En julio se le da un nombre al nuevo “flagelo”. Las autoridades sanitarias de EE. UU. comienzan a utilizar el término: AIDS, siglas en inglés de «Acquired immuno-deficiency syndrome»; SIDA en castellano, Síndrome de Inmunodeficiencia Adquirida. En este mismo año se definieron las vías de transmisión: sexual, sanguínea y materno-infantil. En diciembre la primera infección del virus que causa el SIDA por una transfusión lleva al gobierno de EE. UU. a advertir que los suministros de sangre podrían estar infectados.

1983. En enero, en Francia, en el hospital parisino de la Pitié, el equipo del profesor W. Rozenbaum extirpa un ganglio cervical a F. Brugière, un paciente de 33 años del que se sospechaba estuviera infectado con el virus que causa el SIDA. El profesor Luc Montagnier del Instituto Pasteur, examina el ganglio enfermo y determina que existían rastros de la actividad bioquímica de un retrovirus.

En febrero del mismo año, el profesor Charles Daguet, obtiene la imagen de un nuevo virus sirviéndose de un microscopio electrónico, la fotografía es tomada en la misma muestra extirpada al joven francés. En marzo a los homosexuales, drogadictos intravenosos y otros considerados de alto riesgo, se les recomienda que se abstengan de donar sangre. En mayo, la revista «Science» publica el descubrimiento del profesor Montagnier, quien informa que ha aislado el virus que causa el SIDA y denomina el virus como LAV. En el artículo se señala que el paciente aún no posee los signos característicos del SIDA. El joven francés F. Bruguière fallecerá en 1988.

1984. Muere Gaetan Dugas, quien será considerado durante mucho tiempo como el «paciente cero», el que habría comunicado la enfermedad a sus diferentes parejas homosexuales.

1985. Primera prueba de anticuerpos contra el VIH en EE.UU. que se utiliza para examinar el suministro de sangre de la Nación. Además de la búsqueda de anticuerpos circulantes antiVIH en las personas, se abre un nuevo campo para la prevención, los estudios epidemiológicos y clínicos.

1986. Se identifica en París una variante del virus causante del SIDA en un paciente originario de Cabo Verde, la variedad es denominada HIV-2. El mundo médico y las autoridades piensan que la epidemia se limita al África Occidental.

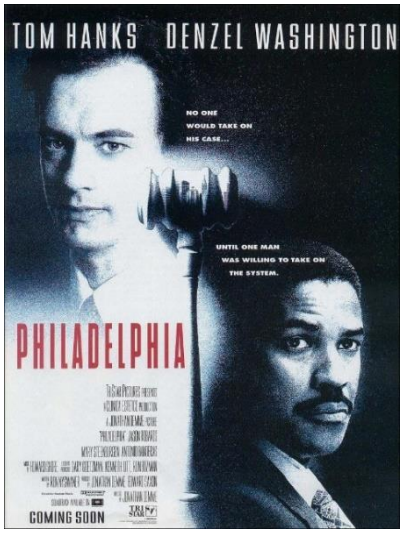
En mayo, la Comisión de Nomenclatura de Virus decide que el agente del SIDA se llamará definitivamente Virus de Inmunodeficiencia Humana (VIH). La Organización Mundial de la Salud, OMS, decreta el día 1 de diciembre **Día Mundial del SIDA**.

1987. En marzo se aprueba el primer fármaco contra el SIDA: AZT. En 1986 los ensayos con AZT, dieron las primeras evidencias acerca de la posibilidad de obtener un tratamiento para esta patología, el que si bien insuficiente para curar, resultaba apto para mejorar la calidad de vida y probablemente para prolongar la sobrevida de las personas afectadas.

1994. El 1 de diciembre de 1994, día mundial de la lucha contra el SIDA, los Jefes de Estado o representantes de 42 países reunidos en Francia, firmaron la Declaración de la Cumbre sobre SIDA en París que declara a la humanidad amenazada por la pandemia del SIDA, y compromete a los países firmantes a implementar estrategias adecuadas para enfrentar la emergencia sanitaria.

Los científicos desarrollan el primer esquema de tratamiento para mujeres embarazadas que sirvió para prevenir la transmisión materno-infantil.

- **Philadelphia**. En 1994, el actor Tom Hanks es premiado con el Oscar por su actuación en la película Philadelphia, una conmocionarte historia acerca de las tribulaciones sufridas por un abogado víctima del SIDA. La película, de trascendencia mundial, sirve para compartir la problemática psicológica de los afectados y favorecer fundamentalmente el debate social sobre el problema exponiendo en forma dramática la injusticia de la discriminación.



1996. Las Naciones Unidas crean el Programa Conjunto de las Naciones Unidas sobre el VIH y SIDA (ONUSIDA). Las características de la enfermedad hacen que escape a las competencias de la OMS. Se trata de acentuar que la epidemia es más que un problema de salud y requiere una aproximación diferente para darle respuesta efectiva.

En julio, en la 11ª Conferencia Internacional sobre el SIDA en Vancouver, Canadá, se presentan resultados de la terapia triple que sugieren que es posible hacer frente al virus (aparición de los Inhibidores de la Proteasa).

1.1.2. El VIH/SIDA en la actualidad

Según la Organización Mundial de la Salud (OMS), "hay una ralentización en el ritmo al cual se van reduciendo las nuevas infecciones por VIH, se va aumentando el acceso al tratamiento y se va terminando con las muertes relacionadas con el SIDA".

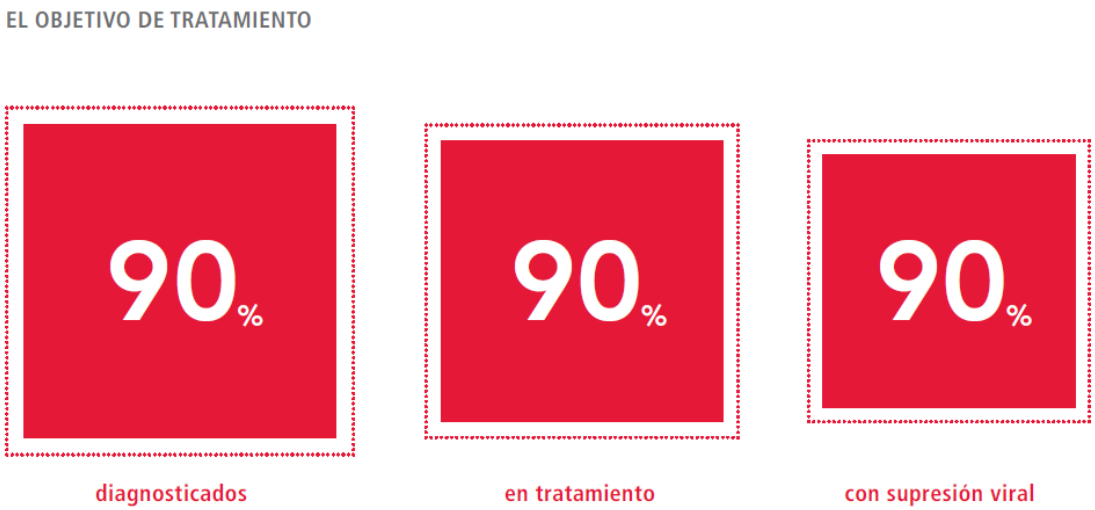
El pasado 1 de diciembre de 2019, con motivo del Día Mundial del SIDA, la OMS publicó las últimas estadísticas registradas (al cierre de 2018). De lo extraído en ONUSIDA, se hace hincapié en los siguientes datos:

- 24,5 millones de personas tenían acceso a la terapia antirretrovírica (al cierre de junio 2019).
- 37,9 millones de personas vivían con el VIH en todo el mundo.
- 1,7 millones de personas contrajeron la infección.
- 770.000 personas fallecieron a causa de las enfermedades relacionadas con el SIDA.
- 74,9 millones de personas contrajeron la infección por el VIH desde el comienzo de la epidemia.
- 32 millones de personas fallecieron a causa de enfermedades relacionadas con el SIDA desde el comienzo de la epidemia.

Además, la ONU ha puesto en marcha el proyecto 90-90-90 para el año 2020, cuyos objetivos son los siguientes:

- Que el 90% de las personas que tienen VIH conozcan su estado serológico respecto al VIH.
- Que el 90% de las personas diagnosticadas con el VIH reciban terapia antirretrovírica continuada.

- Que el 90% de las personas que reciben terapia antirretrovírica tengan supresión viral.



Según ONUSIDA, "cuando se alcance este triple objetivo, al menos el 73% de las personas que viven con el VIH en todo el mundo tendrá supresión viral; un número de dos a tres veces mayor que las estimaciones actuales. La modelización sugiere que lograr estos objetivos antes de 2020 permitirá a la comunidad mundial acabar con la epidemia de sida en 2030, lo que a su vez generará grandes beneficios económicos y sanitarios"

1.1.3. El VIH/SIDA en España

Desde el pasado 1 de noviembre de 2019, según el Ministerio de Sanidad, Consumo y Bienestar Social, "la indicación de profilaxis pre-exposición, de ahora en adelante, PrEP, estará financiada por el SNS. […] La dispensación de este medicamento en el ámbito del SNS se realizará a través de los servicios de farmacia hospitalaria o en centros asistenciales autorizados, sin coste directo para el paciente.

La PrEP como intervención biomédica supone otra opción preventiva y eficaz dentro de la estrategia sanitaria de prevención combinada frente a la infección por el VIH y, en la que el uso del preservativo supone la herramienta principal. Este medicamento, emtricitabina/tenofovir disoproxilo, está indicado en combinación con prácticas sexuales más seguras para reducir el riesgo de infección por el VIH-1, adquirido sexualmente.

Tanto en los ensayos clínicos, como en los estudios de factibilidad o en los estudios observacionales realizados en países en los que se implementa la PrEP, ésta se hace en el marco de la prevención combinada; de esta forma, se promueve fundamentalmente el diagnóstico precoz del VIH y el cribado de infecciones de transmisión sexual, la sensibilización para la reducción de conductas de riesgo o el consumo de drogas y la promoción del uso del preservativo.

La prescripción de la PrEP debe ser realizada por profesionales expertos en el VIH e Infecciones de Transmisión Sexual (ITS) que trabajen en centros del SNS, con garantías de calidad asistencial para el seguimiento de las personas subsidiarias de este medicamento. El Plan Nacional sobre el SIDA lidera actualmente un Grupo de Trabajo que incluye a todos los agentes implicados, representantes de las CC.AA., sociedades científicas y sociedad civil, para elaborar y planificar la estrategia de seguimiento de las personas a las que se les prescriba este medicamento".

Según Isabel Valdés, en relación con las CC.AA., "la Comunidad de Madrid será la segunda región, después de Cataluña, en dispensarla. Lo hará a partir de la segunda quincena de este mes, en el Centro Sandoval, con cita previa y para tres grupos de riesgo: mujeres que ejercen la prostitución y que refieran no usar el preservativo de forma habitual, hombres que tienen sexo con hombres y transexuales".

Las últimas novedades vienen dadas por la noticia referida a que "Sanidad financiará la PReP, la pastilla de prevención del VIH" (Clara Roca). En este artículo publicado por *El Diario*, se afirma que "la directora del Plan Nacional sobre el SIDA, Julia del Amo, ha anunciado la inclusión de la profilaxis pre-exposición (PReP) en la cartera de los servicios básicos de la Seguridad Social".

1.2. Contexto y justificación

Los científicos trabajan para desarrollar una cura; mientras tanto, los avances en biotecnología y la mayor concienciación, prevención y acceso a la sanidad, salvan muchas vidas y son la clave para exterminar la enfermedad, aunque todavía queda mucho trabajo por hacer.

Uno de los principales desafíos para estos datos de tiempo hasta el evento es que generalmente existen instancias censuradas, es decir, el evento de intereses no se observa para estas instancias debido a la limitación de tiempo del período de estudio o la pérdida de seguimiento durante el período de observación. Más precisamente, ciertas instancias han experimentado un evento (o etiquetado como evento) y la información sobre la variable de resultado para las instancias restantes solo está disponible hasta un momento específico en el estudio. Por lo tanto, no es adecuado aplicar directamente algoritmos predictivos utilizando los enfoques estadísticos y de aprendizaje automático estándar para analizar los datos de supervivencia. El análisis de supervivencia proporciona varios mecanismos para manejar los problemas de datos censurados que surgen al modelar datos tan complejos (también conocidos como datos de tiempo hasta el evento cuando modelar un evento particular de interés es el objetivo principal del problema) que ocurre de manera ubicua en varios dominios de aplicaciones del mundo real.

Varios investigadores han desarrollado nuevos algoritmos computacionales para manejar eficazmente desafíos tan complejos. Algunos trabajos relacionados han adaptado varios métodos de aprendizaje automático para resolver los problemas de análisis de supervivencia y los investigadores del aprendizaje automático han desarrollado algoritmos más sofisticados y efectivos que complementan o compiten con los métodos estadísticos tradicionales.

En este sentido, un tema muy novedoso es la aplicación de los diferentes métodos de Machine Learning al Análisis de supervivencia en pacientes infectados de VIH, ya que apenas hay información sobre esta metodología. Actualmente, aunque se están desarrollando muchos algoritmos de aprendizaje automático adaptados a analizar datos

censurados, los principales problemas de esta metodología es que las bases de datos contienen mucha información relacionada, es decir, disponemos de un gran número de variables que se explican entre ellas, pero no hay suficiente volumen de individuos sobre los que estudiar toda esta información.

1.3. Objetivos

El objetivo principal de esta memoria es proporcionar una visión general integral y estructurada de varios métodos de aprendizaje automático para el análisis de supervivencia junto con los métodos estadísticos tradicionales.

1) Estudiaremos los métodos existentes de Machine Learning en el contexto del Análisis de Supervivencia:

    a. Buscaremos los métodos más adecuados para estudiar el tiempo hasta el fracaso al tratamiento antirretroviral en un grupo de pacientes infectados por el VIH.
    b. Estudiaremos los algoritmos de Machine Learning aplicados al Análisis de supervivencia. Como ya hemos comentado, se están desarrollando muchos algoritmos de aprendizaje automático adaptados a analizar datos censurados. El trabajo publicado por Ping et al. hace una buena revisión de los métodos estadísticos que se utilizan habitualmente y las técnicas de aprendizaje automático desarrolladas para el análisis de supervivencia, y ofrece algunas pautas para su aplicación. Toda esta información la encontramos en el artículo científico “Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. 2017 Aug 15.”.

2) Este proyecto busca crear herramientas de descripción que permitan resumir la información que contiene nuestro conjunto de datos de estudio, mediante los métodos estudiados de Machine Learning aplicados al análisis de Supervivencia y mediante métodos propios del Análisis de Supervivencia:

a. Analizar nuestro conjunto de datos. Para ello, se tiene la base de datos "dsurv" del estudio "Instinct", que contiene 342 variables de 995 pacientes que han sido seguidos y evaluados en diferentes momentos de tiempo, dando lugar a una base desbalanceada y heterogénea. Estudiar qué tipo de datos son los que tenemos y qué métodos se pueden aplicar a los mismos.
b. Procesar la información del conjunto de datos (Data Management). Debemos analizar en profundidad el conjunto de datos, hacer una limpieza en el caso de que fuese necesario, estructurarlos y organizarlos de manera que obtengamos conclusiones clarividentes.
c. Generar un modelo, o modelos, que nos ayude a clasificar a los pacientes según la distribución del tiempo hasta el fracaso virológico, en función de los datos demográficos, clínicos y de mutaciones. Los algoritmos que apliquemos a nuestro conjunto de datos nos darán los resultados necesarios para conseguir el clasificador deseado.

3) Exponer y defender la memoria ante el tribunal. Realizar una presentación que comprenda todos los puntos especificados anteriormente, y hacer un video que contenga la defensa de la memoria.

1.4. Enfoque y metodología

Las metodologías que aplicaremos, en este orden, son:

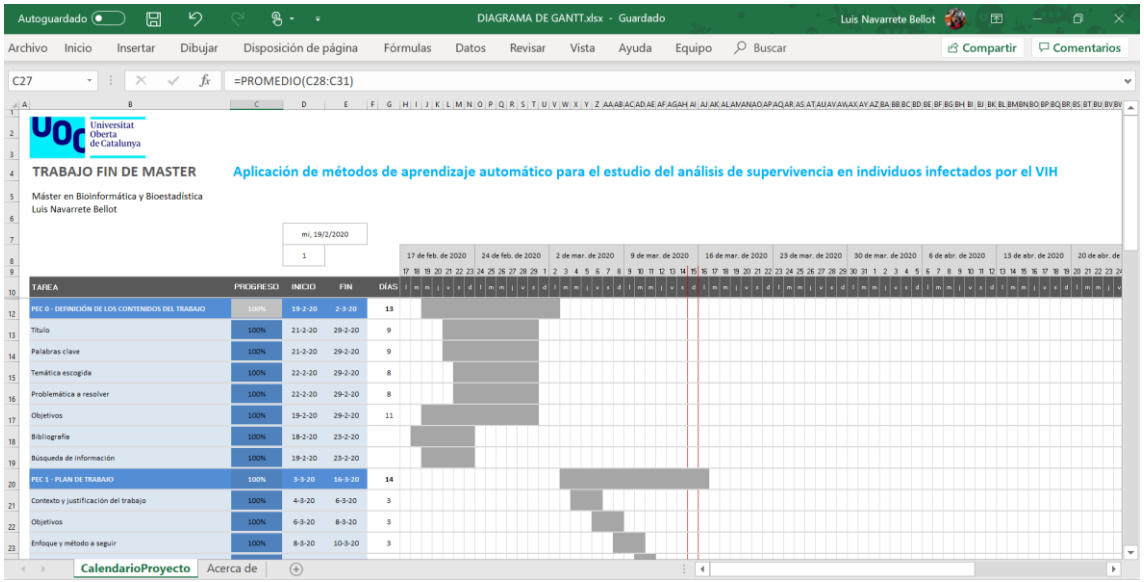
- Análisis descriptivo: en primer lugar, debemos realizar este análisis para estudiar el tipo de datos que tenemos a través de diferentes medidas de centralización y dispersión, con los correspondientes gráficos que nos ayuden a interpretar los datos. Aquí realizaremos análisis univariante y multivariante.
- Análisis de supervivencia: aplicaremos diferentes algoritmos para estudiar el tiempo hasta el fracaso al tratamiento antirretroviral en un grupo de pacientes infectados por el VIH.
- Análisis de Machine Learning: aplicaremos diferentes algoritmos de aprendizaje automático que nos ayuden a clasificar los datos. Esta parte es uno de los objetivos

planteados, generar un modelo de aprendizaje automático con aplicación directa en el análisis de supervivencia que nos ayude a clasificar cada individuo a lo largo de la curva de supervivencia estimada, en función del tratamiento y las mutaciones detectadas en los pacientes; es decir, a clasificar a los individuos según la distribución del tiempo hasta el fracaso virológico.

- Conclusiones y discusión: en base a todos los resultados obtenidos en los puntos anteriores, realizaremos una última parte donde se presentarán los resultados más relevantes del estudio, que nos ofrezcan la mejor bondad de ajuste de los modelos a los datos. Una vez tengamos todos resultados, podremos explicar las conclusiones a las que hemos llegado, y discutir sobre los aspectos de mejora que encontremos durante el análisis.

1.5. Planificación de la memoria: hitos, temporización e imprevistos

Para la realización de la memoria, su estructuración y organización, realizamos un diagrama de Gantt en formato MS Excel. Este diagrama cuenta, como estructura principal, las Pruebas de Evaluación Continua (PEC) a entregar durante el cuatrimestre. Cada una de las PEC cuenta con una serie de tareas, con fecha de inicio y fecha de finalización, y porcentaje de consecución de cada una de las tareas.



Además del factor tiempo y su alcance como factores de riesgo a la hora de realizar la memoria, podemos encontrarnos con otros factores o imprevistos y que, además, se han cumplido a lo largo de la realización de la memoria, como son:

- Dificultad a la hora de ejecutar los algoritmos: puede ser que desconozcamos la aplicación de algún algoritmo o, aunque sepamos aplicarlo, no nos ofrezca los resultados que nosotros esperemos. Esto podría hacer que perdamos tiempo por no saber si lo estamos aplicando correctamente. Para ello, hemos consultado en foros como "*Github*" o "*Stack overflow*", y hemos realizado búsquedas exhaustivas en internet para dar con ejemplos de los algoritmos utilizados, diferentes de los ejemplos que se proporcionan en la ayuda del software de R.
- Ejecución del informe dinámico con los resultados obtenidos: no siempre se ejecuta correctamente el informe dinámico en RMarkdown, y puede deberse a características de la programación que hagan que no funcione correctamente el informe. Una vez finalizado el informe dinámico procedemos a su ejecución, y hemos obtenido errores de que nomenclaturas en los "R chunk" (cuadros de código). Una vez resuelto esto, comprobamos que la ejecución funciona sin problema.
- Diseño del informe dinámico: aunque hayamos ejecutado correctamente nuestro informe de resultados, puede que no nos guste el diseño de determinados gráficos, o cuadros de código, los cuales deberemos ajustar para que el informe quede lo mejor posible. Una vez ejecutado el informe dinámico, analizamos cada página para comprobar el resultado, y realizamos los ajustes necesarios para que el informe quede ordenado y estructurado.

1.6. Resultados esperados

Al finalizar el proyecto, obtendremos los siguientes ficheros que compondrán la totalidad de la memoria.

- Memoria, que contendrá los siguientes archivos:

  - Memoria en formato PDF.
  - Fichero ".RData", que contiene los datos del estudio "Instinct".
  - Fichero "survkernel_reduced.csv", que contiene los datos del estudio "Instinct" depurados y validados.
  - Fichero ".Rmd" que contendrá el informe dinámico RMarkdown y el código utilizado en el software RStudio para la realización del estudio de los datos. Para su ejecución, se requiere la versión de R 3.6.2, y la instalación y carga de las librerías indicadas al comienzo del archivo.
  - Fichero ".bib" que contendrá la información de la bibliografía de las funciones utilizadas en el informe dinámico RMarkdown.
  - Informe dinámico generado por RMarkdown en formato PDF.

- Planificación de la memoria en un fichero MS Excel con el Diagrama de Gantt, que contempla los hitos y tareas de la memoria.
- Presentación MS PowerPoint.
- Video con la defensa de la memoria.

1.7. Estructura de la Memoria

Según el Plan Docente, la estructura de la memoria se organiza en los siguientes puntos:

- La primera parte consiste en presentar el tema sobre el que vamos a tratar. Realizamos una introducción sobre el VIH/SIDA, su historia y cómo evoluciona la enfermedad en la actualidad.
- La segunda parte consta del desarrollo del estudio: data management, análisis exploratorio de los datos (medidas de centralización, medidas de dispersión, gráficos, etc.), aplicación de los algoritmos de Machine Learning y análisis de supervivencia para obtener clasificadores e indicadores que mejor se ajusten a nuestro caso de estudio.
- La tercera, y última parte contendrá la totalidad de la memoria, su defensa pública con su correspondiente presentación audiovisual.

2. DESARROLLO DE LA MEMORIA

2.1. Lectura, exploración de datos y obtención de las muestras

2.1.1. Base de datos

La base de datos "dsurv" del estudio "Instinct" contiene 342 variables de 995 pacientes que han sido seguidos y evaluados en diferentes momentos de tiempo, y han dado lugar a una base desbalanceada y heterogénea.

La base de datos se estructura en 5 vistas diferentes, divididas entre ellas por 24 semanas:

- Vista 0: se recogieron datos de 1003 individuos, de los cuales 8 fueron descartados por diferentes motivos clínicos.
- Vista 24: se recogieron datos de 992 individuos.
- Vista 48: se recogieron datos de 988 individuos.
- Vista 72: se recogieron datos de 828 individuos.
- Vista 96: se recogieron datos de 659 individuos.

En cada vista, vemos el número descendiente de individuos; esto es debido a que los individuos han sido incluidos en el estudio en diferentes momentos de tiempo. La fecha de cierre de la base de datos con la que se trabaja fue 03/09/2018.

Además, en la base de datos encontramos los siguientes tipos de variables:

- Demográficas.
- Clínicas.
- De relación coste-efectividad.
- Inmuno-virológicas.
- Parámetros hematológicos y bioquímicos.

2.1.2. Data Management

En primer lugar, cargamos el dataset ".RData" y visualizamos el conjunto de datos en el software RStudio. Haciendo una primera revisión, detectamos muchas variables que contienen datos "NA". Para trabajar mejor con ello, vamos a guardar el dataset en un fichero csv, para realizar filtros de una manera más rápida.

Si abrimos el dataset en MS Excel, vemos más claramente las columnas que contienen datos NA para todos los individuos de la muestra. Como estas columnas no nos aportan información, prescindimos de ellas, y dejamos el dataset con 96 variables de las 342 iniciales.

Revisamos las variables del dataset una a una y detectamos que, en las variables de fechas, hay algunas fechas mal escritas para algunos individuos; más concretamente, el error está al haber insertado el año que, en vez de estar escrito 2016 o 2017, venía 0016 o 0017. Los individuos son {363, 380, 627, 732, 787, 804, 843, 848, 858, 863, 904}, por lo que procedemos a corregir estas erratas para que el dataset esté correcto y podamos realizar el mejor análisis sin necesidad de eliminar variables por errores a la hora de informar la fecha.

Otro tratamiento de datos importante que debemos realizar es en las variables "censura". En nuestro dataset, la censura se representa con el "1", pero el software R toma como valor de censura el "0". Debemos alternar el orden en el dataset para que los datos en R sean correctos.

Como también vamos a utilizar las variables de las diferentes mutaciones (integrasa, proteasa y transcriptasa inversa), aquí detectamos que hay valores "NA". Lo que haremos es que, en vez de eliminarlos de nuestro estudio ya que perderíamos información, vamos a recategorizar los "NA" como "Sin información". Así, para nuestros análisis, mantendremos información de las otras variables de estudio.

Por último, podemos cargar de nuevo nuestro dataset en R, para dar por finalizado el proceso ETL (Extract – Treatment – Load).

2.1.3. Exploración de los datos

Una vez hemos hecho la lectura de los datos, y hemos realizado las modificaciones pertinentes, vamos a estudiar el tipo de las variables demográficas y clínicas que comprenden el dataset:

- **Edad:** estamos interesados en saber cómo se distribuye la edad entre los pacientes de nuestro conjunto de datos. Para ello, generamos un histograma y un boxplot:

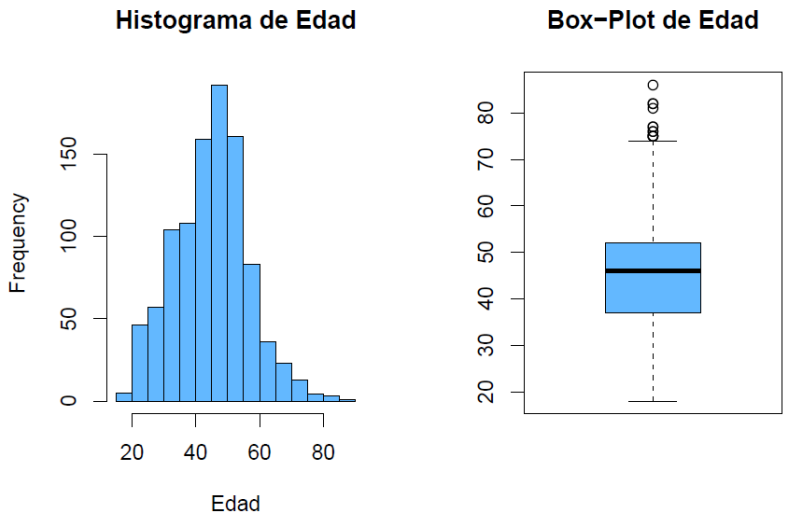


Ilustración 1: Histograma y Boxplot de la variable Edad

A primera vista, según apreciamos en el histograma, parece que la variable "Edad" se distribuye como una Normal. Vemos, además, que la mayoría de los individuos del estudio tienen edades comprendidas entre los 40 y los 55 años, pero también existe otro conjunto amplio entre los 20 y los 39 años.

El gráfico de Boxplot nos confirma lo que comentamos en el párrafo anterior, el conjunto de individuos mayoritario está entre los 37 y los 52 años, aproximadamente. Por último, detectamos que existen valores atípicos, es decir,

a partir de los 72 años hay unos individuos que podrían considerarse como datos atípicos.

- **Sexo:** para esta variable, vemos que nuestro conjunto de datos está formado por más hombres que mujeres; concretamente, el 80,2% de los pacientes de la muestra son hombres.

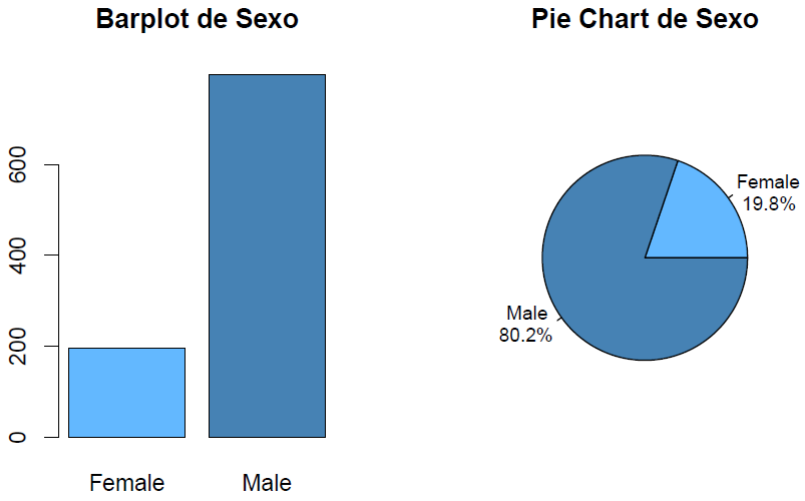


Ilustración 2: Gráfico de barras y de sectores de la variable Sexo

- **NADIR CD4**: estamos interesados en saber la distribución de los pacientes según el número mínimo de linfocitos CD4 de cada individuo.

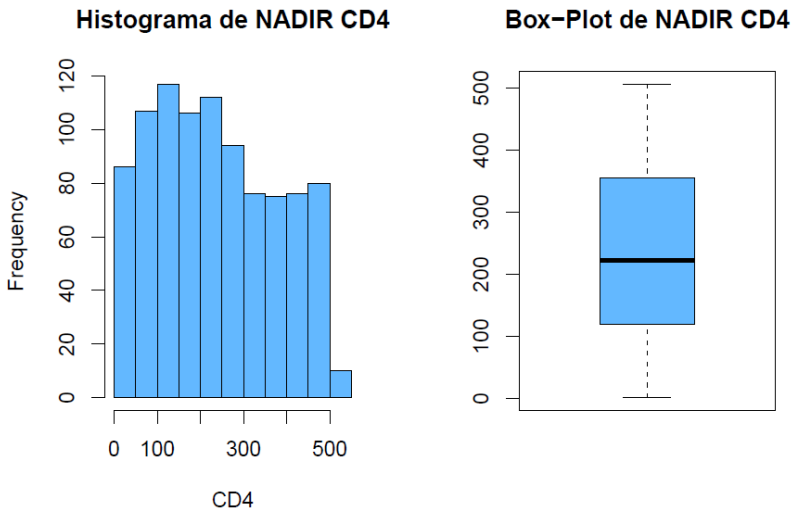


Ilustración 3: Histograma y Boxplot de la variable NADIR CD4

En los gráficos relacionados con la variable NADIR_CD4, que miden el mínimo de linfocitos CD4 al inicio del estudio, vemos que esta variable no sigue una distribución concreta, y tampoco existen datos atípicos. Sí se puede apreciar que, al parecer, existen más individuos en el estudio con un nivel de CD4 < 200.

- **Motivo de inicio**: para conocer más acerca de la información de los pacientes, estamos interesados en saber cuál es el motivo de inicio de su tratamiento:

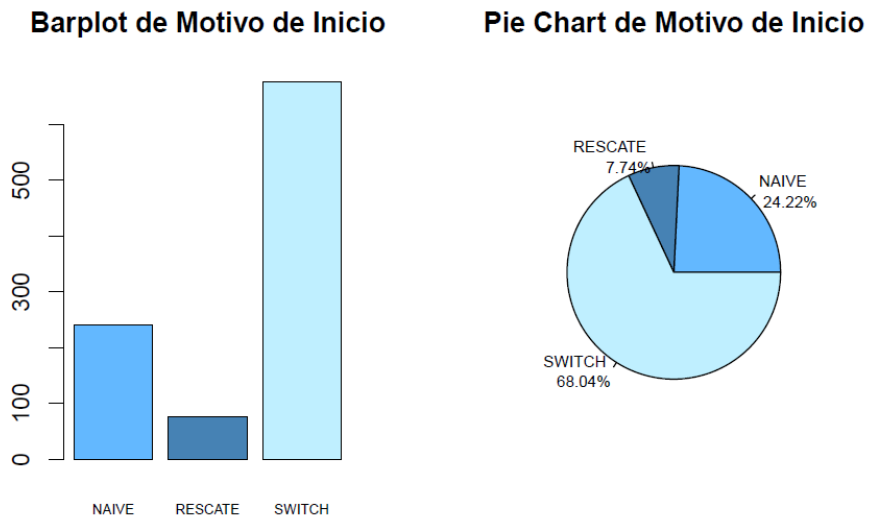


Ilustración 4: Gráfico de barras y de sectores de la variable Motivo de inicio

Nos encontramos con que la mayoría de los individuos del conjunto de datos comienzan el estudio desde el escenario de estrategias de cambio ("SWITCH"), concretamente, el 68% de los individuos. De los individuos que conforman el 32% restante del dataset, el 22% son de primera línea de tratamiento ("NAIVE"), y casi el 8% restante comienzan en el estudio desde un escenario de "RESCATE".

- **Inhibidores de Integrasa**: estamos interesados en saber cuántos pacientes están tomando los tratamientos *Dolutegravir*, *Elvitegravir* y *Raltegravir*:

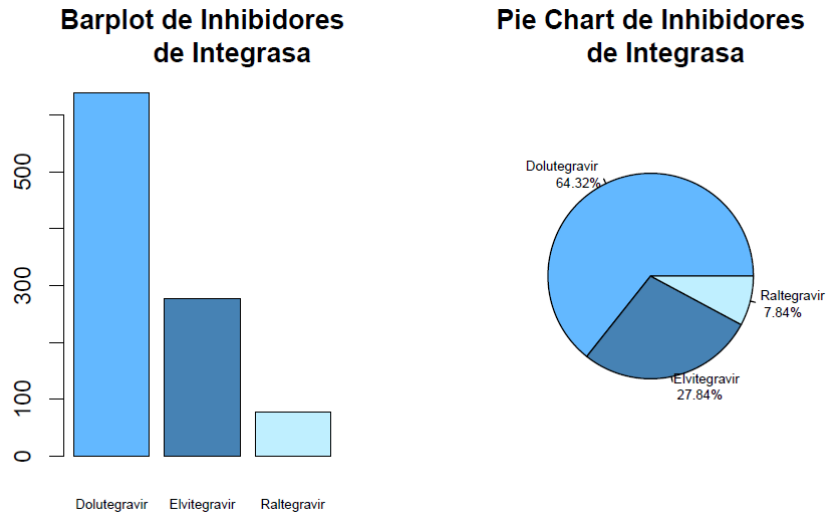


Ilustración 5: Gráfico de barras y de sectores de la variable Inhibidores de Integrasa

En los gráficos vemos que el 64% de los pacientes están tomando *Dolutegravir (DTG)*, casi el 28% de los pacientes está con *Elvitegravir (ELV)* y apenas el 8% de ellos toma *Raltegravir (RAL)*.

- **Tratamientos Antirretrovirales Previos (TAR) según los Inhibidores de Integrasa**: estamos interesados en saber si los pacientes venían o no de un tratamiento antirretroviral previo, según el tratamiento que están tomando.

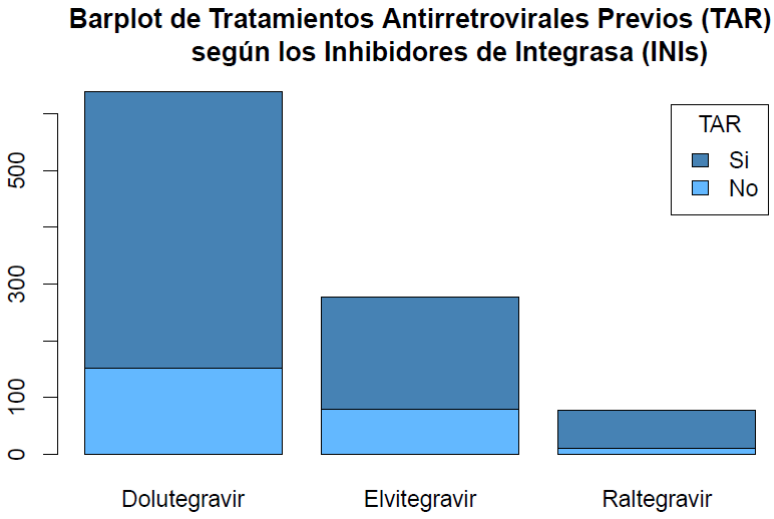


Ilustración 6: Gráfico de barras apiladas de los Tratamientos Antirretrovirales Previos (TAR) según los Inhibidores de Integrasa

En este gráfico vemos que:

- Del 64% de los pacientes que están tomando *Dolutegravir*, el 76% de los individuos del estudio venían de tomar un tratamiento antirretroviral previo.
- Casi el 28% de los pacientes que están tomando *Elvitegravir*, el 71% de los individuos del estudio venían de tomar un tratamiento antirretroviral previo.
- Casi el 8% de los pacientes que están tomando *Raltegravir*, cerca del 86% de los individuos venían de tomar un tratamiento antirretroviral previo.

El resto de los pacientes, poco más del 24% de ellos, no estaban tomando un tratamiento antirretroviral previo.

- **Inhibidores de Integrasa según el motivo de inicio**: estamos interesados en saber cuántos pacientes toman los tratamientos antirretrovirales según el motivo de inicio del tratamiento.

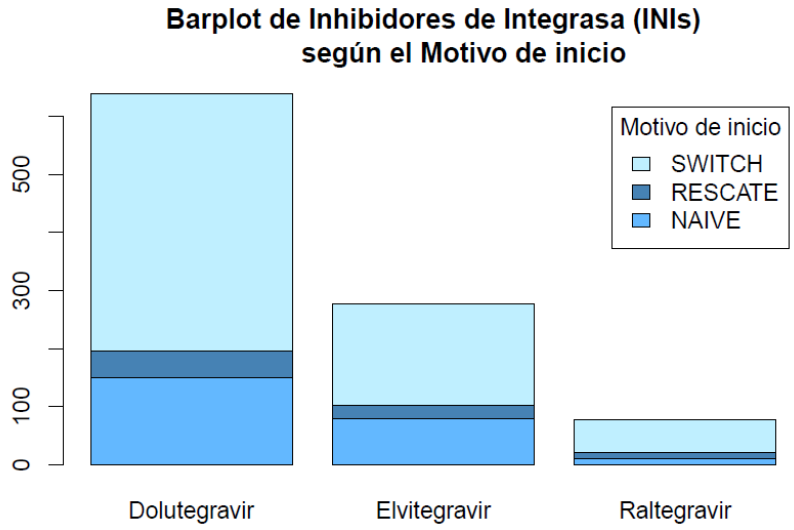


Ilustración 7: Gráfico de barras apiladas de los Inhibidores de Integrasa según el Motivo de inicio

En este gráfico vemos que:

- Del 64% de los pacientes que están tomando *Dolutegravir*, el 69% viene de una estrategia de cambio y el 23% son de primera línea de tratamiento.
- Casi el 28% de los pacientes que están tomando *Elvitegravir*, el 63% viene de una estrategia de cambio y casi el 29% son de primera línea de tratamiento.
- Casi el 8% de los pacientes que están tomando *Raltegravir*, el 74% viene de una estrategia de cambio y el 14% son de primera línea de tratamiento.

El resto de los pacientes, cerca del 8%, vienen de una estrategia de rescate.

- **Mutaciones en la integrasa, proteasa y transcriptasa inversa según el Inhibidor de integrasa**: estamos interesados en saber con qué tratamiento puede haber mayores mutaciones, o no haberlas.

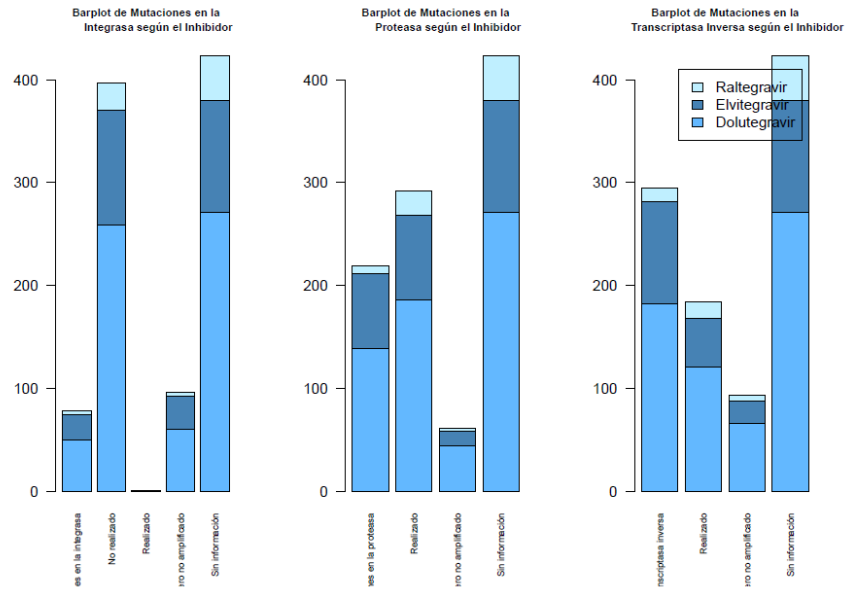


Ilustración 8: Gráfico de barras apiladas de las Mutaciones según el Inhibidor de Integrasa

Según vemos en los diagramas de barras en función del tipo de mutación se aprecia claramente que, si ha habido mutaciones en la integrasa, proteasa y transcriptasa inversa ha podido ser debido al *Dolutegravir*; esto puede deberse a

que el tratamiento fue retirado rápidamente después de la detección de fracaso virológico, mientras que, en la rutina clínica, donde muchos pacientes se siguen cada 6 meses (como es nuestro caso de estudio), la detección de un fracaso a INIs (incluyendo al DTG) podría retrasarse y conducir a la acumulación de resistencias a esta familia de fármacos.

Además, según vemos en los gráficos, parece ser que hay menos mutaciones en los fármacos *Raltegravir* y *Elvitegravir*, respectivamente.

2.1.4. Obtención de las muestras

A continuación, para proceder con el estudio de los datos a través de los diferentes algoritmos de Machine Learning y de Supervivencia, vamos a crear los subconjuntos *train* y *test*. Para crear estos subconjuntos, en primer lugar, establecemos una semilla aleatoria que nos ayudará a fijar los mismos subconjuntos de datos en cualquier software R de diferentes equipos informáticos.

Lo que hacemos es seleccionar una muestra aleatoria entre todo el tamaño muestral, y obtenemos como resultado el subconjunto de *train* con $n_{train} = 663$ individuos y el subconjunto *test* con $n_{test} = 332$ individuos.

2.2. Aplicación de algoritmos de Supervivencia

"Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable "tiempo hasta que ocurre un evento" y su dependencia de otras posibles variables explicativas. Por ejemplo, en el estudio de enfermedades crónicas o tratamientos muy agresivos, el tiempo hasta que ocurre la muerte del enfermo (tiempo de supervivencia) y su dependencia de la aplicación de distintos tratamientos, pero en otras enfermedades, el tiempo hasta la curación, o el tiempo hasta la aparición de la enfermedad. En procesos de control de calidad se estudia el tiempo hasta que un cierto producto falla (tiempo de fallo), o el tiempo de espera hasta recibir un servicio (tiempo de espera), etc.

Debido a que la variable tiempo es una variable continua podría ser, en principio, estudiada mediante las técnicas de análisis de la varianza o los modelos de regresión. Hay, sin embargo, dos dificultades importantes para este planteamiento. En primer lugar, en la mayor parte de los estudios citados la variable tiempo no tiene una distribución normal, más bien suele tener una distribución asimétrica y aunque podrían intentarse transformaciones que la normalizaran, existe una segunda dificultad que justifica un planteamiento específico para estas variables, y es que para observarlas se tiene que prolongar el estudio durante un período de tiempo suficientemente largo, en el cual suelen ocurrir pérdidas, que imposibilitan la observación del evento.

Existen tres motivos por los que pueden aparecer estas pérdidas, en primer lugar, por fin del estudio. Supóngase, por ejemplo, que para evaluar una intervención quirúrgica arriesgada se sigue en el tiempo, durante un año, a dos grupos de pacientes. A los de un grupo se les practicó la intervención y a los de otro no, y se registró la duración del intervalo de tiempo entre la intervención (o la entrada en el estudio, para el grupo no intervenido) y la muerte. Al final del estudio puede haber individuos que no hayan muerto. Otra causa es la pérdida propiamente dicha, por ejemplo, se quiere evaluar la eficacia de un tratamiento preventivo para el SIDA, y se sigue durante cinco años a individuos VIH+. Algunos de los individuos, y puede ser un número importante, desaparecerán del estudio en algún momento de este por diversos motivos: cambio de domicilio, falta de interés, etc. Una última causa de pérdida es la ocurrencia de un evento competitivo, en los ejemplos anteriores puede ser muerte por alguna otra causa ajena al estudio. Aunque los ejemplos anteriores son del ámbito de Ciencias de la Salud, estos mismos problemas aparecen en cualquier estudio que necesite un largo tiempo de observación.

Hay que tener en cuenta también que la variable es el tiempo hasta que ocurre un evento, y está definida por la duración del intervalo temporal entre los instantes en que empieza la observación y ocurre el evento. En los ejemplos citados, la observación no comienza en el mismo instante para todos los individuos. En algunos textos se denomina pérdida por la izquierda a esta no coincidencia de los tiempos en que comienza la

observación, ya que, si el estudio está diseñado para acabar en un tiempo determinado, el efecto de esta no coincidencia es reducir, para los que empiezan más tarde, el tiempo de observación".

Siguiendo con V. Abraira, "si se quisiera aplicar un modelo de regresión lineal a un estudio de este tipo, habría que eliminar del mismo las observaciones perdidas, ya que para ellas no se conoce el valor de la variable; sin embargo, sí se tiene alguna información útil sobre la misma: se sabe que es mayor que el tiempo en el que se produjo la pérdida".

2.2.1. Modelo de Riesgos Proporcionales de Cox

2.2.1.1. Introducción al Modelo de Riesgos Proporcionales de Cox

Según López Montoya, en la investigación biomédica, el conocimiento de los factores que determinan el pronóstico de los pacientes es de gran importancia clínica. En la mayoría de los casos, la variable respuesta representa, en cierto sentido, un tiempo de supervivencia (por ejemplo, el tiempo que transcurre antes de la ocurrencia de un evento particular de interés), y por lo tanto se formula un modelo de regresión con el fin de determinar la relación entre el tiempo y un conjunto de covariables explicativas. El modelo de CPH, ver Cox (1972), es el modelo utilizado por la mayoría de las aplicaciones en el campo de la Bioestadística y generalmente, en los estudios de fiabilidad y supervivencia.

El modelo de Cox es el enfoque de análisis de regresión más utilizado para los datos de supervivencia y difiere significativamente de otros métodos, ya que se basa en el supuesto de riesgos proporcionales y emplea una probabilidad parcial para la estimación de parámetros. El método de regresión de Cox se describe como un método semiparamétrico, ya que la distribución del resultado sigue siendo desconocida, incluso si se basa en un modelo de regresión paramétrica.

La relevancia de este modelo depende fuertemente de que, paralelamente al desarrollo de los importantes resultados teóricos en estos últimos años, hay algoritmos

implementados en programas estadísticos gratuitos. La mayoría de los paquetes estadísticos cuentan con funciones para facilitar el ajuste del modelo de CPH en aplicaciones reales. El entorno estadístico R, es actualmente el software líder en este sentido y, en particular, el *survival package* que proporciona varias funciones y bases de datos para el análisis de la supervivencia.

2.2.1.2. Aplicación del Modelo de Riesgos Proporcionales de Cox

En primer lugar, para obtener el modelo de regresión de Cox, debemos cargar la librería {survival}, y utilizamos la función *coxph*. Según los resultados obtenidos, detectamos como covariables más significativas el tratamiento y las mutaciones en la integrasa, pero, en líneas generales, podemos concluir que el modelo de regresión de Cox es aceptable para cualquiera de estos tres criterios: test de razón de verosimilitud, test de Wald y test de Score o *logrank*.

Estos coeficientes estimados se consideran significativos cuando *z*, en valor absoluto, es superior a 2, ya que para muestras grandes este coeficiente se distribuye según la Ley Normal (prueba de Wald).

Si analizamos los resultados más detalladamente:

- El modelo supone que el efecto en la supervivencia del tratamiento *Elvitegravir* es 1.36 veces mayor que los otros dos tratamientos, es decir, a mayor efecto del *Elvitegravir*, menos probabilidad de supervivencia en los pacientes.
- En las mutaciones de integrasa, produce un mayor efecto el que no se haya realizado mutación alguna; esto es, existe mayor probabilidad de supervivencia si no hay mutaciones en la integrasa.

Realizamos el gráfico de supervivencia según el modelo de regresión de Cox:

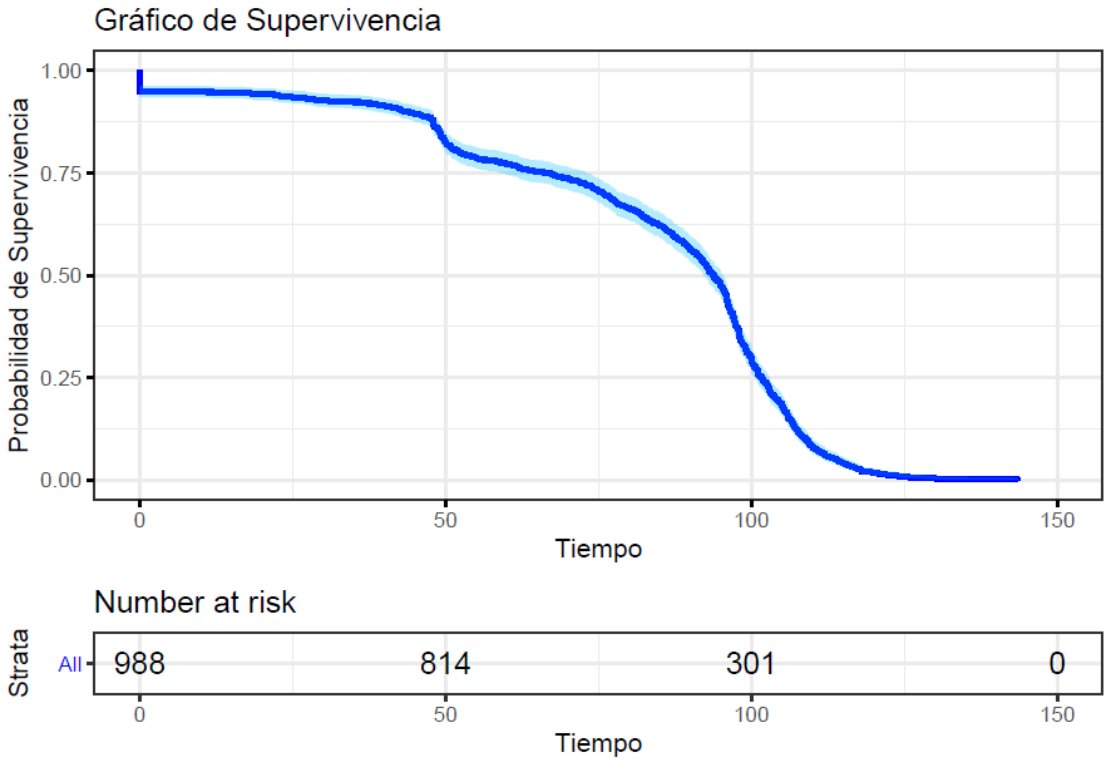


Ilustración 9: Gráfico de supervivencia del modelo de Riesgos Proporcionales de Cox

2.2.2. Modelo de tiempo de vida acelerada (AFT)

2.2.2.1. Introducción al modelo de tiempo de vida acelerada.

Según Kleinbaum, el modelo de Cox es el modelo de supervivencia más utilizado en las ciencias de la salud, pero no es el único modelo disponible. Existe una clase de modelos de supervivencia, llamados modelos paramétricos, en los que la distribución del resultado (es decir, el tiempo hasta el evento) se especifica en términos de parámetros desconocidos. Muchos modelos paramétricos son modelos de tiempo de falla de aceleración en los que el tiempo de supervivencia se modela en función de las variables predictoras.

Específicamente, Kleinbaum presenta modelos de supervivencia paramétricos y los supuestos que subyacen a estos modelos, examina el supuesto de tiempo de falla acelerado (AFT) y lo contrasta con el supuesto de riesgos proporcionales (PH).

La regresión lineal, la regresión logística y la regresión de Poisson son ejemplos de modelos paramétricos que se usan comúnmente en las ciencias de la salud. Con estos modelos, se supone que el resultado sigue alguna distribución, como la distribución normal, binomial o de Poisson. Por lo general, lo que realmente se quiere decir es que el resultado sigue a una familia de distribuciones de forma similar con parámetros desconocidos. Es solo cuando se conoce el valor de los parámetros que la distribución exacta está completamente especificada. Para los modelos de regresión paramétrica, los datos se usan típicamente para estimar los valores de los parámetros que especifican completamente esa distribución.

El modelo de riesgos proporcionales de Cox, por el contrario, no es un modelo totalmente paramétrico. Más bien es un modelo semiparamétrico porque incluso si se conocen los parámetros de regresión, la distribución del resultado sigue siendo desconocida. La función de supervivencia (o riesgo) de referencia no se especifica en un modelo de Cox.

Las estimaciones de supervivencia obtenidas a partir de modelos de supervivencia paramétricos generalmente producen gráficos más consistentes con una curva de supervivencia teórica. Si el investigador se siente cómodo con el supuesto de distribución subyacente, se pueden estimar los parámetros que especifican por completo las funciones de supervivencia y peligro. Esta simplicidad e integridad son los principales atractivos del uso de un enfoque paramétrico.

2.2.2.2. Aplicación del modelo de tiempo de vida acelerada

Para obtener los diferentes modelos de vida acelerada con los que vamos a trabajar, debemos utilizar la función *flexsurvreg* de la librería {flexsurv}. Esta función nos da la posibilidad de, a partir de nuestros datos de estudio, asumir que el efecto de las covariables es acelerar o desacelerar el curso de la vida del VIH en los pacientes.

Asumiremos que el efecto de las covariables se distribuye de las siguientes formas:

- Distribución Weibull.
- Distribución Gamma
- Distribución Exponencial.
- Distribución Log-Normal.
- Distribución Log-Logistic.

Una vez hemos generado los modelos de las distribuciones, mostramos en la siguiente tabla los resultados más relevantes:

Tabla 1: Resultados de los modelos de vida acelerada

Distribución	Criterio de Akaike (AIC)	Razón de verosimilitud
Weibull	8537	-4267
Gamma	8818	-4407
Exponencial	10 227	-5113
Log-Normal	8985	-4491
Log-Logistic	8860	-4428

Si atendemos al criterio de Akaike, la distribución que mejor se ajusta a la curva de supervivencia es la distribución Weibull, ya que es la distribución con menor valor para este estadístico. En el siguiente gráfico, podemos visualizar la bondad de ajuste de la distribución a la curva de supervivencia:

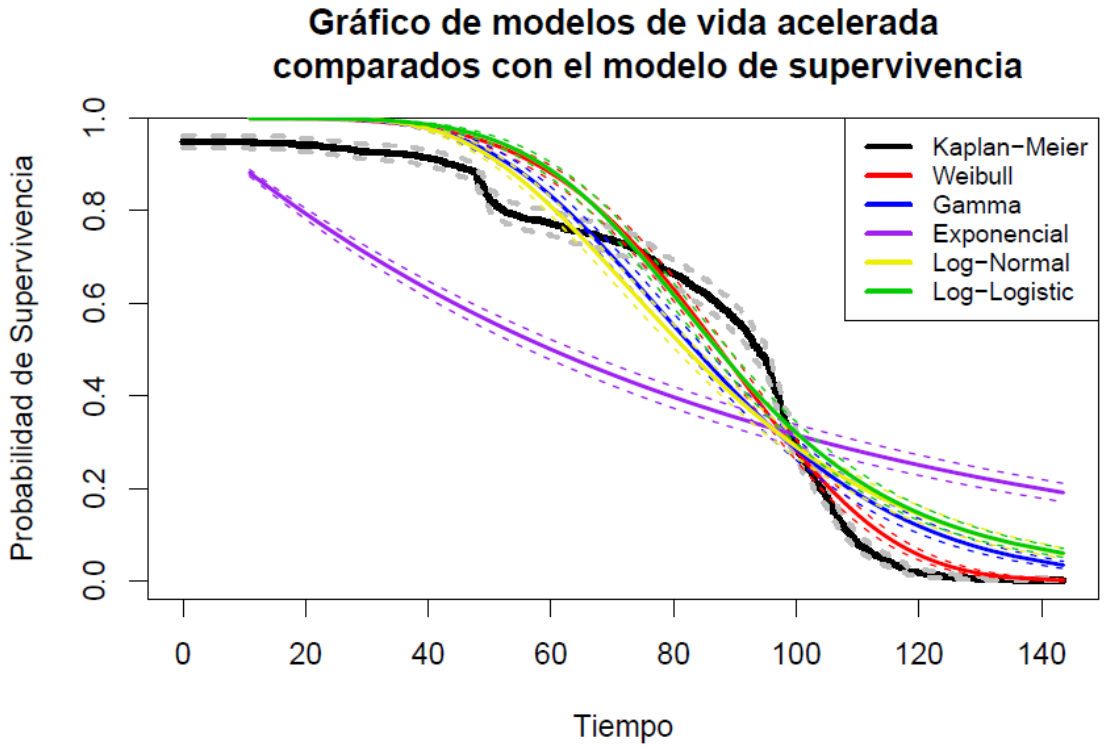


Ilustración 10: Gráfico de curvas de modelos de vida acelerada comparados con la curva del modelo de supervivencia

Cabe destacar que podemos descartar del modelo la distribución Exponencial, por ser la distribución que menos se ajusta a la curva de supervivencia, además de tener mayor valor en el criterio de Akaike.

2.3. Aplicación de algoritmos de Machine Learning

En el análisis de supervivencia, el principal desafío de los métodos de aprendizaje automático es la dificultad de manejar adecuadamente la información censurada y la estimación del tiempo del modelo. El aprendizaje automático es efectivo cuando hay una gran cantidad de instancias en un espacio de características dimensionales razonable, pero este no es el caso para ciertos problemas en el análisis de supervivencia.

Este punto se estructura en cuatro puntos importantes: Survival Trees, Métodos Bayesianos, Máquinas de Vector Soporte de Supervivencia (SSVM), Regresión Logística Multitarea (MTLR) y Redes Neuronales Recurrentes (RNN). Dentro de cada punto

explicamos las características de los algoritmos y los resultados obtenidos más relevantes para nuestro estudio.

2.3.1. Survival Trees

2.3.1.1. Introducción al modelo Survival Trees

Los árboles de supervivencia son una forma de árboles de clasificación y regresión que están diseñados para manejar datos censurados. La intuición básica detrás de los modelos de árbol es dividir recursivamente los datos en función de un criterio de división particular, y los objetos que son similares entre sí en función del evento de interés se colocarán en el mismo nodo.

La principal diferencia entre un árbol de supervivencia y el árbol de decisión estándar está en la elección del criterio de división. El método del árbol de decisión realiza particiones recursivas en los datos al establecer un umbral para cada característica, sin embargo, no puede considerar las interacciones entre las características ni la información censurada en el modelo. Los criterios de división utilizados para la supervivencia se pueden agruparse en dos categorías:

- Maximizar la heterogeneidad entre nodos: minimiza la función de pérdida utilizando el criterio de homogeneidad dentro del nodo.
- Minimizar la homogeneidad dentro de los nodos: se emplean estadísticas de prueba de rango logarítmico para medidas de heterogeneidad entre nodos.

Otro aspecto importante de la construcción de un árbol de supervivencia es la selección del árbol final. Se pueden seguir procedimientos como la selección hacia atrás o hacia adelante para elegir el árbol óptimo. Sin embargo, un conjunto de árboles puede evitar el problema de la selección final de árboles con un mejor rendimiento en comparación con un solo árbol.

2.3.1.2. Aplicación del algoritmo Survival Trees

Entrenamos el modelo de Survival Trees con los datos del conjunto *train*. Para entrenar este modelo, usamos la función *rfsrc* del paquete {randomForestSRC}. El objeto devuelto por este algoritmo nos proporciona información sobre los parámetros de ejecución, como el tamaño de la muestra, el número de árboles, etc. Mostramos la importancia de cada variable según el modelo obtenido:

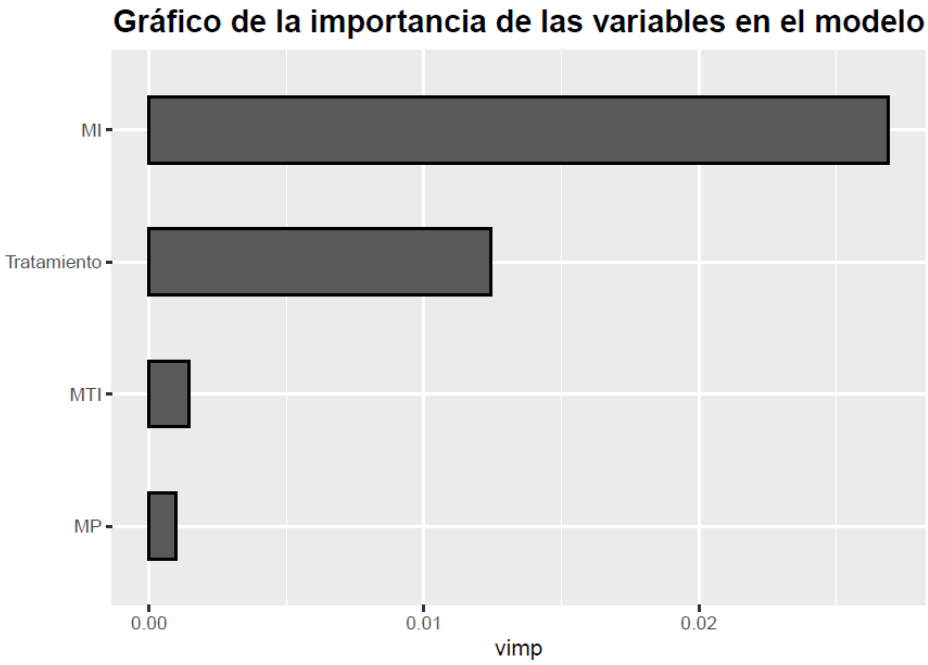


Ilustración 11: Gráfico de la importancia de las variables en el modelo Random Forest SRC

Según vemos en el gráfico, tenemos ordenadas las variables según la importancia que tiene cada una en el modelo: Mutaciones en la integrasa, Tratamiento, Mutaciones en la proteasa y, por último, las Mutaciones en la Transcriptasa Inversa.

Una vez hemos ejecutado el modelo con los datos del conjunto *train*, vemos que hemos obtenido error en el modelo de 46,5%. Además, el modelo cuenta con el objeto "time.interest", que contiene los tiempos de supervivencia en los que han ocurrido eventos. La predicción del modelo nos muestra un error del modelo del 41,7%. Además, obtenemos un índice de concordancia del 50,3%.

Por último, mostramos los gráficos de la predicción. En primer lugar, los gráficos de supervivencia y de riesgo acumulado de cada uno de los factores del modelo:

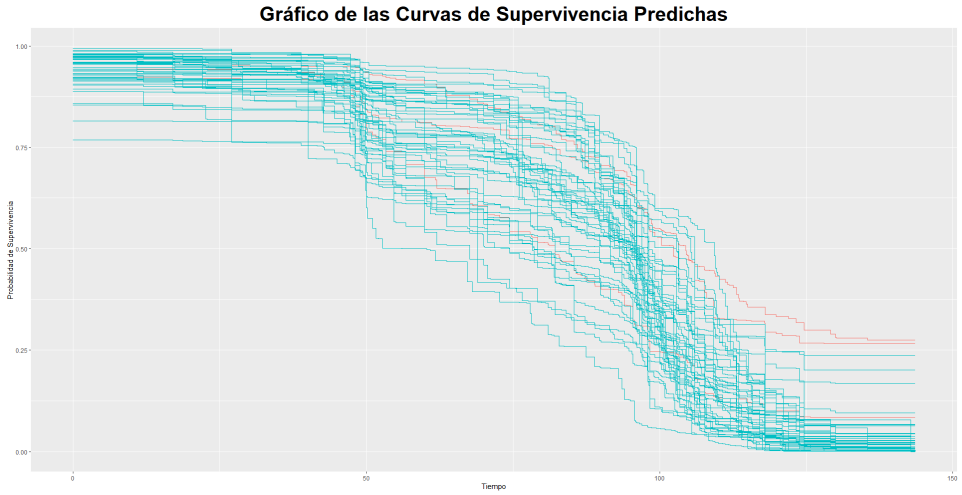


Ilustración 12: Gráfico de curvas de supervivencia predichas del modelo Random Forest SRC

En el gráfico de curvas de supervivencia predichas detectamos la estabilidad de los pacientes hasta, prácticamente, la semana 50. A partir de esta semana, disminuye la probabilidad de supervivencia hasta la semana 125 que, parece ser, las curvas se estabilizan. Además, observamos 3 curvas en color rojo que, posiblemente, representen a la variable de las Mutaciones en la Transcriptasa Inversa ya que, según el gráfico de importancia de variables, se nos indicaba como la menos importante del estudio.

Para terminar con el análisis de este modelo, graficamos las curvas de supervivencia de cada variable y sus factores:

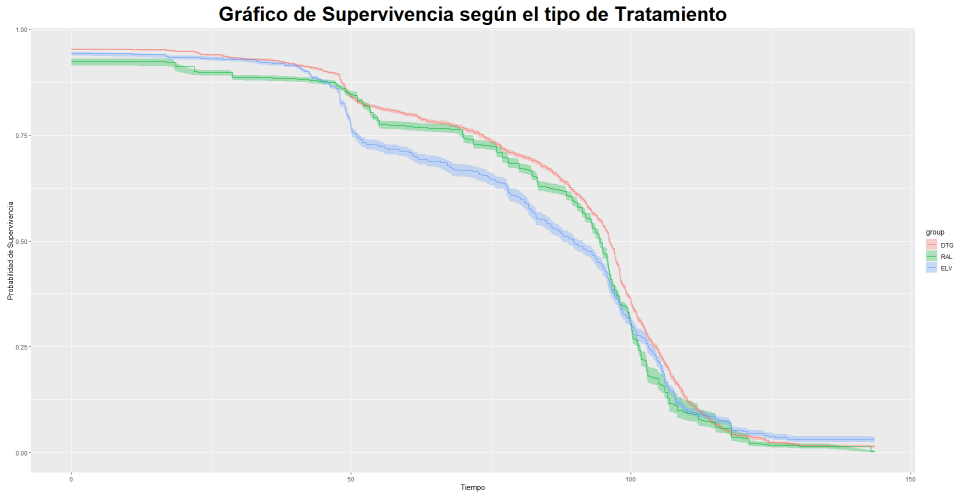


Ilustración 13: Gráfico de curvas de supervivencia del modelo Random Forest SRC según el Tratamiento

Según vemos en el gráfico, con el tratamiento *Elvitegravir* disminuye la supervivencia del paciente poco antes de la semana 50 en casi un 25%. Observamos que los pacientes que toman *Dolutegravir* y *Raltegravir* tienen mayores probabilidades de supervivencia.

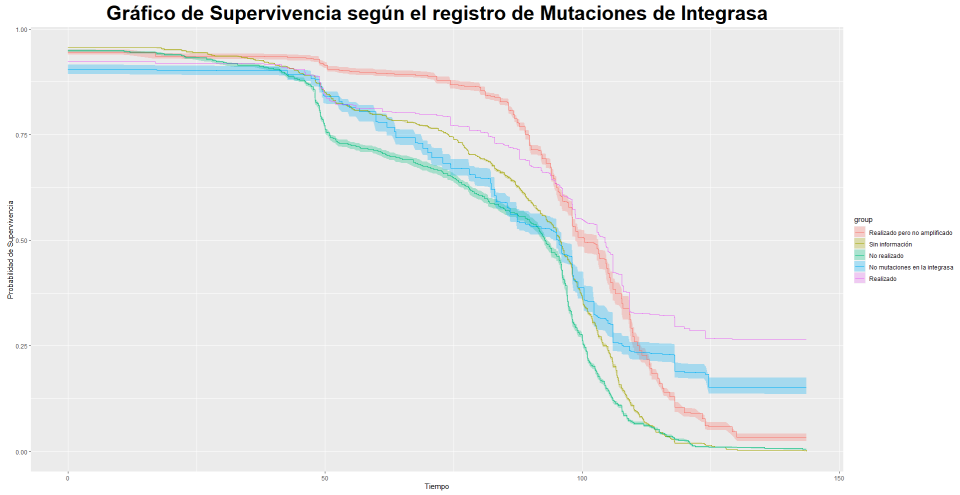


Ilustración 14: Gráfico de curvas de supervivencia del modelo Random Forest SRC según las Mutaciones en la Integrasa

Según vemos en el gráfico de las mutaciones de integrasa, existe mayor estabilidad de supervivencia en los pacientes en los que no ha habido mutaciones en la integrasa; pero se observa mayor probabilidad de supervivencia en los pacientes en los que se ha realizado el test de las mutaciones, aunque no se detecta nada porque no hay suficiente carga viral.

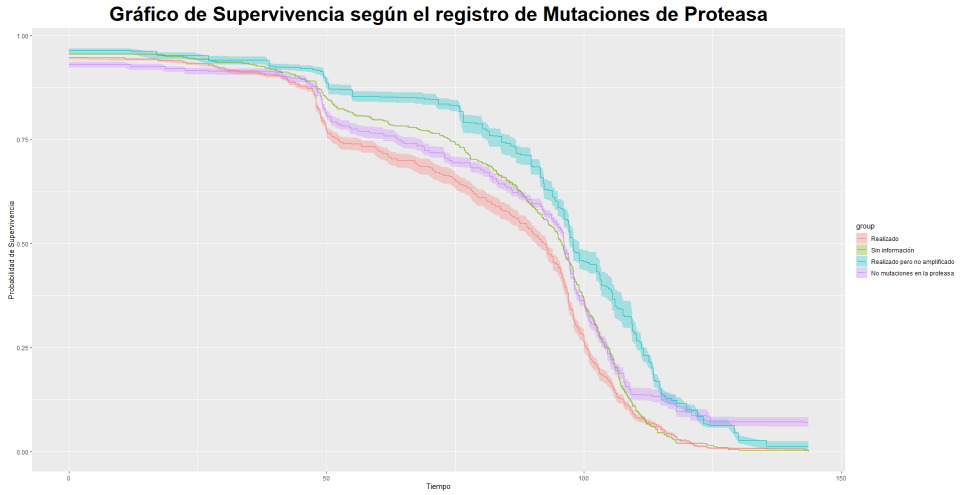


Ilustración 15: Gráfico de curvas de supervivencia del modelo Random Forest SRC según las Mutaciones en la Proteasa

Según el gráfico de las mutaciones de proteasa, ocurre algo similar a las mutaciones de integrasa, se observa mayor probabilidad de supervivencia en los pacientes en los que se ha realizado el test de las mutaciones, pero no se detecta nada porque no hay suficiente carga viral.

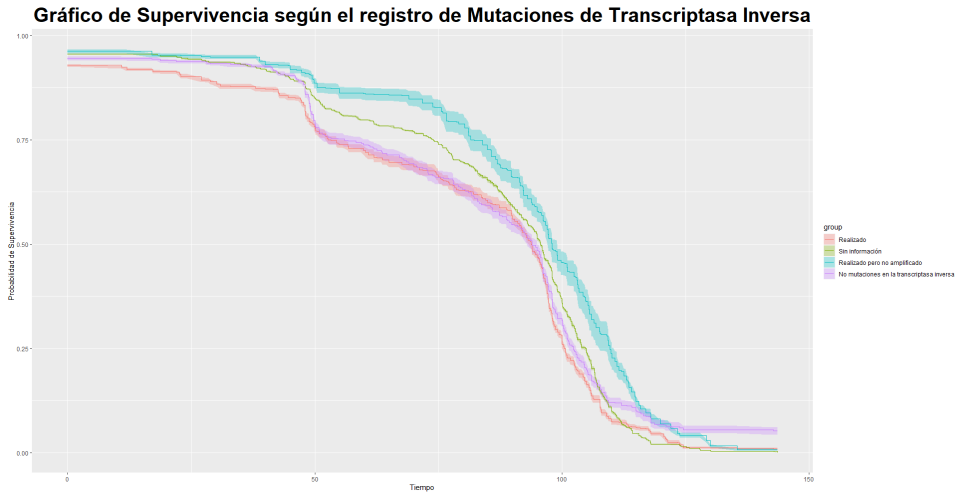


Ilustración 16: Gráfico de curvas de supervivencia del modelo Random Forest SRC según las Mutaciones en la Transcriptasa Inversa

Por último, en el gráfico de las mutaciones de transcriptasa inversa, observamos mayor probabilidad de supervivencia en los pacientes en los que se ha realizado el test de las mutaciones, pero no se detecta nada porque no hay suficiente carga viral.

2.3.2. Métodos Bayesianos

2.3.2.1. Introducción al modelo de Métodos Bayesianos

El teorema de Bayes es uno de los principios más fundamentales en la teoría de probabilidad y estadística matemática; proporciona un enlace entre la probabilidad posterior y la probabilidad previa, de modo que uno puede ver los cambios en los valores de probabilidad antes y después de contabilizar un evento determinado. Usando el teorema de Bayes, hay dos modelos, a saber, Naïve Bayes (NB) y red bayesiana (BN).

Ambos enfoques, que proporcionan la probabilidad del evento de intereses como sus resultados, se estudian comúnmente en el contexto de la predicción clínica. Los resultados experimentales del uso de métodos bayesianos en los datos de supervivencia muestran que los métodos bayesianos tienen buenas propiedades de razonamiento tanto de interpretabilidad como de incertidumbre.

Naïve Bayes, un método probabilístico bien conocido en el aprendizaje automático, es uno de los algoritmos de predicción más simples pero efectivos. En un sentido amplio, los modelos de Naïve Bayes son una clase especial de algoritmos de clasificación de Aprendizaje Automático que se basan en una técnica de clasificación estadística llamada “teorema de Bayes”.

Estos modelos son llamados algoritmos “Naïve”; en ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad. Lo consiguen proporcionando una forma de calcular la probabilidad ‘posterior’ de que ocurra un cierto evento *A*, dadas algunas probabilidades de eventos ‘anteriores’.

En la siguiente tabla se presentan las principales fortalezas y debilidades del modelo bayesiano:

Tabla 2: Fortalezas y debilidades del Método Bayesiano

Fortalezas	Debilidades
Una manera fácil y rápida de predecir clases para problemas de clasificación binarios y multiclase.	Aunque son unos clasificadores bastante buenos, los algoritmos Naive Bayes son conocidos por ser pobres estimadores. Por ello, no se deben tomar muy en serio las probabilidades que se obtienen.
En los casos en que sea apropiada una presunción de independencia, el algoritmo se comporta mejor que otros modelos de clasificación, incluso con menos datos de entrenamiento.	La presunción de independencia Naive muy probablemente no reflejará cómo son los datos en el mundo real.
El desacoplamiento de las distribuciones de las características condicionales de clase significa que cada distribución puede ser estimada independientemente como si tuviera una sola dimensión. Esto ayuda con problemas derivados de la dimensionalidad y mejora el rendimiento.	Cuando el conjunto de datos de prueba tiene una característica que no ha sido observada en el conjunto de entrenamiento, el modelo le asignará una probabilidad de cero y será inútil realizar predicciones. Uno de los principales métodos para evitar esto, es la técnica de suavizado, siendo la estimación de Laplace una de las más populares.

Un clasificador bayesiano se usa para hacer predicciones en medicina clínica mediante la estimación de varias probabilidades a partir de los datos. Recientemente, integra de manera efectiva los métodos bayesianos con un modelo AFT extrapolando la probabilidad de evento anterior para implementar la predicción de la etapa temprana en los datos de supervivencia para los puntos de tiempo futuros. Un inconveniente del método de Naïve Bayes es que hace la suposición de independencia entre todas las características, lo que puede no ser cierto para muchos problemas en el análisis de supervivencia.

Una red bayesiana, en la que las características pueden relacionarse entre sí en varios niveles, puede representar gráficamente una distribución teórica sobre un conjunto de variables. Las redes bayesianas pueden representar visualmente todas las relaciones entre las variables, lo que lo hace interpretable para el usuario final. Puede adquirir información de conocimiento mediante el uso de procedimientos para estimar las estructuras y parámetros de red a partir de un conjunto de datos dado.

Más recientemente, se propuso un marco novedoso que combina el poder de la representación de red bayesiana con el modelo AFT mediante la extrapolación de las probabilidades anteriores a puntos de tiempo futuros. La complejidad temporal de estos enfoques bayesianos depende principalmente de los tipos de métodos bayesianos utilizados en los modelos.

2.3.2.2. Aplicación del algoritmo de Métodos Bayesianos

Para trabajar con el modelo bayesiano debemos utilizar la función *BayesSurv_HReg* que se encuentra en la librería {SemiCompRisks}. Esta función es la nueva versión de la antigua función *BayesSurv*, que se podía encontrar en versiones de R anteriores a 2.5.

Para trabajar con este modelo, lo primero de todo es establecer etiquetas a la variable explicativa INIs_ESTUDI, que contiene la información de los tratamientos de los pacientes. Debemos etiquetar cada tratamiento con un número porque la función con la que vamos a trabajar solo acepta valores numéricos.

El siguiente paso es establecer unos parámetros para que la función compile correctamente. Estos parámetros son:

- *hyperParams*: una lista que contiene listas o vectores para valores de hiperparámetros en modelos jerárquicos.
- *mcmcParams*: Una lista que contiene las variables requeridas para el muestreo Markov Chain Monte Carlo.

Una vez establecidos los parámetros, podemos aplicar el modelo a nuestros datos. Según los resultados obtenidos, el modelo nos da unas ratios de riesgo de:

- Tratamiento: $e^{\beta} = e^{0.065} = 1.067$.
- Mutaciones en la Integrasa: $e^{\beta} = e^{-0.049} = 0.952$.
- Mutaciones en la Proteasa: $e^{\beta} = e^{0.092} = 1.097$.
- Mutaciones en la Transcriptasa Inversa: $e^{\beta} = e^{-0.021} = 0.98$.

Si analizamos los resultados vemos que, prácticamente, para todas las covariables explicativas de nuestro modelo se produce un fracaso virológico a la semana, lo que supondría alrededor de 100 fracasos al finalizar el estudio, es decir, hemos obtenido fracaso virológico en 100 pacientes.

Por último, graficamos a curva de supervivencia predicha para el modelo de Bayes:

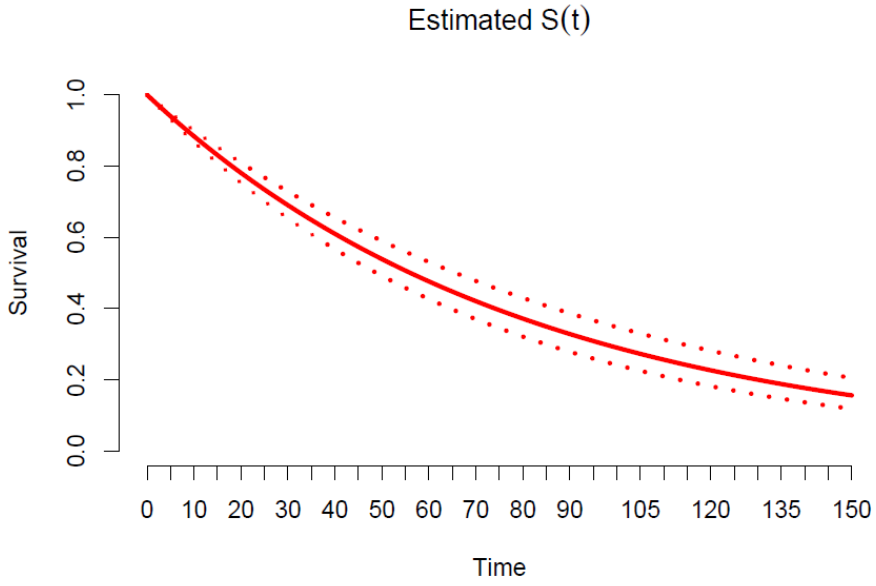


Ilustración 17: Gráfico de la curva de supervivencia predicha para el modelo bayesiano

2.3.3. Survival Support Vector Machines (SSVM)

2.3.3.1. Introducción al modelo Survival Support Vector Machines

Una máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria o regresión. Las máquinas de vectores de soporte son muy populares en aplicaciones como el procesamiento del lenguaje natural, el habla, el reconocimiento de imágenes y la visión artificial.

Una SVM construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo.

Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión. Las SVM pertenecen a una clase de algoritmos de Machine Learning denominados métodos kernel y también se conocen como máquinas kernel.

El entrenamiento de una SVM consta de dos fases:

1) Transformar los predictores (datos de entrada) en un espacio de características altamente dimensional. En esta fase es suficiente con especificar el kernel; los datos nunca se transforman explícitamente al espacio de características. Este proceso se conoce comúnmente como el truco kernel.
2) Resolver un problema de optimización cuadrática que se ajuste a un hiperplano óptimo para clasificar las características transformadas en dos clases. El número de características transformadas está determinado por el número de vectores de soporte.

A continuación, se muestran las ventajas y desventajas de las SVM:

Tabla 3: Fortalezas y debilidades del modelo SVM

Fortalezas	Debilidades
Efectivo en espacios de altas dimensiones.	Si el número de características es mucho mayor que el número de muestras, hay que evitar el ajuste excesivo al elegir las funciones del núcleo y el término de regularización es crucial.
Sigue siendo efectivo en casos donde el número de dimensiones es mayor que el número de muestras.	Los SVM no proporcionan directamente estimaciones de probabilidad, estas se calculan utilizando una costosa validación cruzada de cinco veces.
Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también es eficiente en la memoria.	
Versatilidad: se pueden especificar diferentes funciones de Kernel para la función de decisión. Se proporcionan núcleos comunes, pero también es posible especificar núcleos personalizados.	

También se ha adaptado con éxito a problemas de análisis de supervivencia. La principal desventaja de este enfoque es que la información del pedido incluida en las instancias censuradas será completamente ignorada. Otro posible enfoque para manejar los datos censurados es utilizar la clasificación de vectores de soporte utilizando el enfoque de clasificación de restricción que impone restricciones en la formulación de SVM para dos instancias comparables con el fin de mantener el orden requerido. Sin embargo, la complejidad computacional para este algoritmo es cuadrática con respecto al número de instancias. Además, solo se enfoca en el orden entre las instancias e ignora los valores reales de la salida.

2.3.3.2. Aplicación del algoritmo Survival Support Vector Machines

Para obtener los diferentes modelos de Máquinas de Vector Soporte de Supervivencia con los que vamos a trabajar, debemos utilizar la función *survivalsvm* de la librería {survivalsvm}. Aquí debemos notar que trabajaremos con tres tipos de kernel diferentes, y tres modelos de SVM diferentes.

Respecto a los diferentes tipos de kernel:

- Lineal.
- Additive.
- Radial (RBF).

Respecto a los diferentes modelos de SVM:

- Regresión.
- Vanbelle1.
- Vanbelle2.

A continuación, mostramos los resultados obtenidos mediante el índice de concordancia (función *conindex* de la librería {survivalsvm}) de cada model (kernel y tipo de SVM) en la siguiente tabla:

Tabla 4: Resultados del Índice de Concordancia obtenidos para cada modelo de Survival SVM y tipo de Kernel

C-index		**Modelo de Survival SVM**		
		Regresión	Vanbelle1	Vanbelle2
Kernel	Linear	0.5156	0.5498	0.5503
	Additive	0.5841	0.5574	0.5557
	RBF	0.5447	0.5647	0.5757

Según vemos en la tabla, el modelo que mejor índice de concordancia nos ofrece es el modelo de regresión con el tipo de kernel *additive*, con un valor de 0.5858.

2.3.4. Multitask Logistic Regression (MTLR)

2.3.4.1. Introducción al modelo MTLR

La Regresión Logística Multitarea (MTLR) fue diseñado específicamente para dar probabilidades de supervivencia en un rango de tiempos para observaciones individuales. Esto difiere de los modelos que producen puntajes de riesgo (como los dados por los riesgos proporcionales de Cox), los modelos de probabilidad de tiempo único (como el modelo Gail) y los modelos de población amplia (por ejemplo, curvas de Kaplan-Meier). La producción de probabilidades de supervivencia en un rango de veces brinda una visión más holística de la supervivencia a los pacientes y los médicos, lo que puede ser crítico para tomar decisiones de atención médica.

MTLR se introdujo por primera vez en 2011 en NIPS (Sistemas de Procesamiento de Información Neural) bajo el nombre de "Aprendizaje de distribuciones de supervivencia al cáncer específicas del paciente como una secuencia de regresores dependientes". Desde entonces, se ha trabajado mucho, incluido un sitio web que se puede utilizar para construir modelos MTLR en datos cargados.

Cuando se trata de predecir la función de supervivencia para una unidad específica, el modelo de riesgo proporcional de Cox suele ser el modelo de referencia. Sin embargo, presenta algunos inconvenientes importantes:

- Se basa en el supuesto de riesgo proporcional, que especifica que la función de peligro de dos individuos debe ser constante en el tiempo.
- La fórmula exacta del modelo que puede manejar vínculos no es computacionalmente eficiente, y a menudo se reescribe usando aproximaciones, como las aproximaciones de Efron o Breslow, para ajustarse al modelo en un tiempo razonable.
- El hecho de que el componente de tiempo de la función de peligro permanezca sin especificar hace que el modelo CoxPH no sea adecuado para las predicciones reales de la función de supervivencia.

Esa es la razón por la cual se desarrolló el modelo de Regresión logística multitarea (MTLR). Puede verse como una serie de modelos de regresión logística construidos en diferentes intervalos de tiempo para estimar la probabilidad de que el evento de interés ocurra dentro de cada intervalo.

Aunque el modelo MTLR proporciona resultados similares al modelo CoxPH sin tener que depender de los supuestos requeridos por este último, en su núcleo, todavía está impulsado por una transformación lineal. Por lo tanto, ambos modelos no logran capturar elementos no lineales de los datos y, en consecuencia, dejan de producir rendimientos satisfactorios. La regresión logística neuronal de tareas múltiples (N-MTLR) ayudará a resolver este problema.

2.3.4.2. Aplicación del algoritmo MTLR

Para llevar a cabo este análisis, debemos cargar la librería {MTLR}. En esta librería podemos encontrar información acerca de las regresiones logísticas multitarea.

En primer lugar, debemos encontrar el valor óptimo de validación cruzada. Este parámetro nos servirá para aplicarlo a nuestro modelo y que los cálculos sean óptimos, además de evitar que tengamos que indicar una malla de parámetros de validación cruzada en el modelo.

Una vez hemos obtenido el valor óptimo de la validación cruzada, estimamos nuestro modelo y obtenemos el siguiente gráfico:

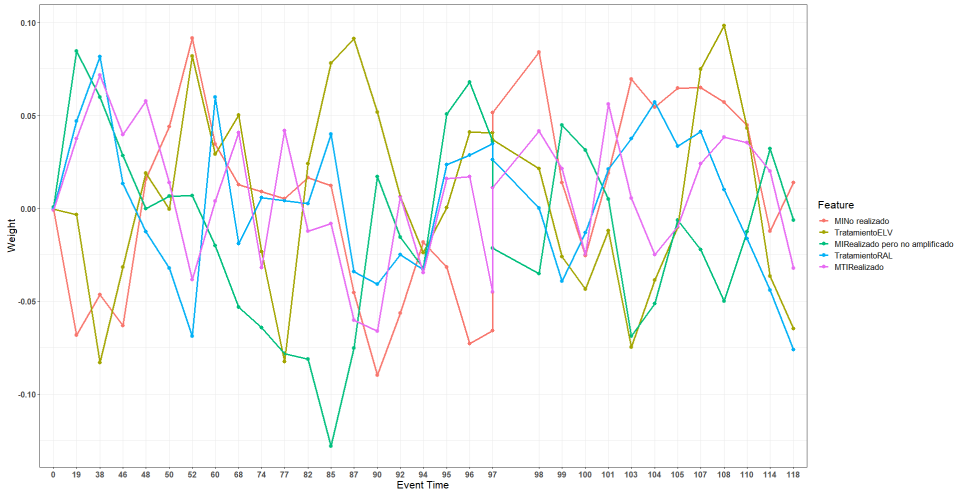


Ilustración 18: Gráfico de las cinco características con mayor suma de valores absolutos en el modelo MTLR

De forma predeterminada, en el gráfico obtenemos las cinco características que tuvieron mayor suma de valores absolutos a lo largo del tiempo, es decir, los factores de las variables con mayor influencia.

Al final del tiempo, en el gráfico podemos apreciar que el factor más influyente es que no se haya realizado el test de la mutación de integrasa, seguido de que sí se haya realizado dicho test, pero no se haya amplificado, y que se haya realizado el test de la mutación de la transcriptasa inversa.

A continuación, realizamos la predicción del modelo de regresión logística multitarea. Cuando usamos la función de predicción para curvas de supervivencia, se nos devolverá una matriz donde la primera columna (tiempo) es la lista de puntos de tiempo

que el modelo evaluó la probabilidad de supervivencia para cada observación (estos serán los puntos de tiempo utilizados por MTLR y unos 0 puntos adicionales).

Tabla 5: Tabla de los resultados de las predicciones individuales en el modelo MTLR

	time <dbl>	1 <dbl>	2 <dbl>	3 <dbl>	4 <dbl>	5 <dbl>
1	0.00000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
2	0.00000	0.9995856	0.9995769	0.9996216	0.9995769	0.9997693
3	18.56176	0.9560009	0.9551984	0.9598425	0.9551984	0.9756699
4	38.00000	0.9426993	0.9324848	0.9402548	0.9324848	0.9692087
5	45.99176	0.9323227	0.9084172	0.9148370	0.9084172	0.9636561

Cada columna siguiente corresponderá al número de fila de los datos pasados. En este caso, la columna 2 (llamada 1) corresponde a la fila 1 de la prueba. Cada fila de esta matriz da las probabilidades de supervivencia en el punto de tiempo correspondiente (dado por la columna de tiempo). Por ejemplo, la observación de prueba 1 tiene una probabilidad de supervivencia de 95.38% en el momento t = 18.56.

Dado que estas curvas pueden ser difíciles de interpretar al observar una matriz de probabilidades de supervivencia, también podemos elegir trazarlas. Graficamos las predicciones de las curvas de supervivencia para los 10 primeros individuos:

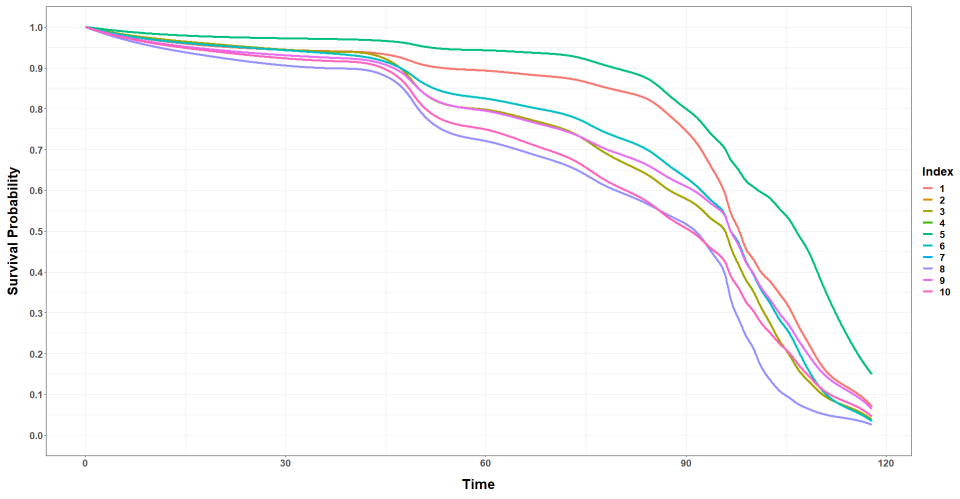


Ilustración 19: Gráfico de curvas de supervivencia predichas para cada individuo en el modelo MTLR

Aquí hemos especificado que queremos observar las curvas de supervivencia para las primeras 10 observaciones (correspondientes a las primeras 10 filas de prueba). Debemos notar que estas curvas se han suavizado, mientras que antes solo teníamos probabilidades para ciertos puntos de tiempo. Por último, realizamos una comparativa entre el gráfico de la curva de supervivencia real y las curvas de supervivencia predichas de 3 pacientes para el modelo MTLR:

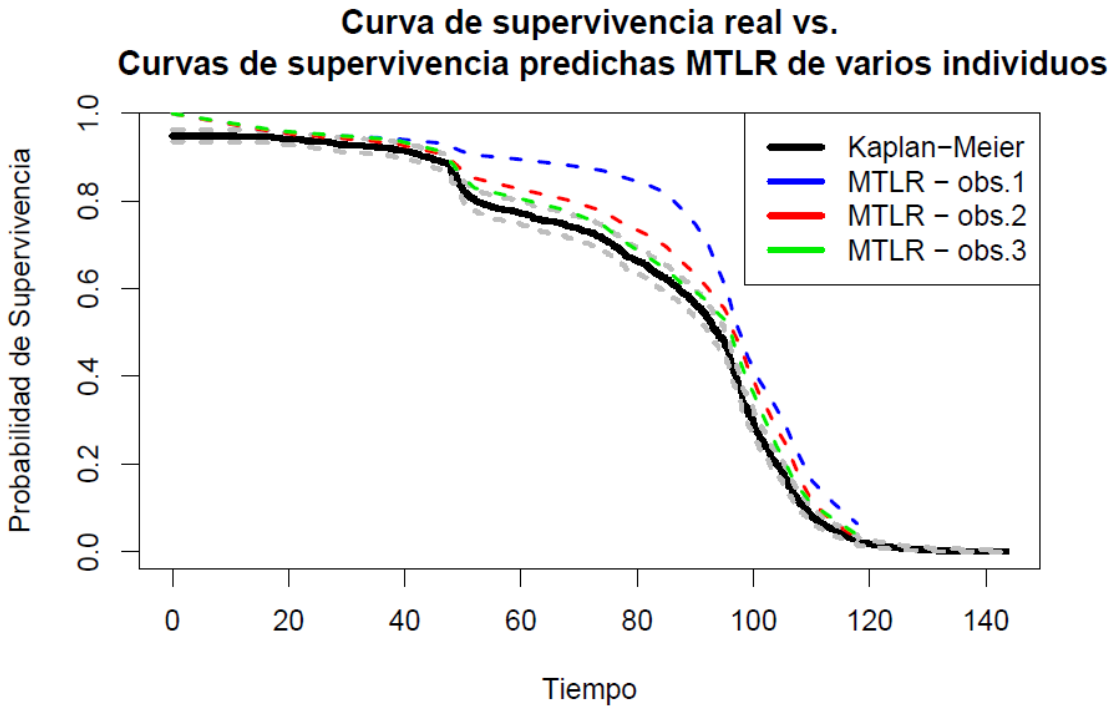


Ilustración 20: Gráfico de supervivencia real vs. Curvas de supervivencia predichas MTLR para varios individuos

2.3.5. Redes Neuronales Recurrentes (RNN)

2.3.5.1. Introducción al modelo de Redes Neuronales

Desde la primera mitad del siglo XX se han empezado a desarrollar modelos computacionales que han intentado emular el comportamiento del cerebro humano. Aunque se han propuesto una gran cantidad de ellos, todos usan una estructura en red en la cual los nodos o neuronas son procesos numéricos que involucran estados de otros nodos según sus uniones. Una clase de estos modelos computacionales son las Redes Neuronales Artificiales.

Las Redes Neuronales Artificiales (RNA) se han hecho muy populares debido a la facilidad en su uso e implementación y la habilidad para aproximar cualquier función matemática. Las Redes Neuronales Artificiales, con su marcada habilidad para obtener resultados de datos complicados e imprecisos, pueden utilizarse para extraer patrones y detectar tramas que son muy difíciles de apreciar por humanos u otras técnicas computacionales.

Una de las definiciones que se estima más certera de Red Neuronales Artificiales es la siguiente: "Las redes neuronales son conjuntos de elementos de cálculo simples, usualmente adaptativos, interconectados masivamente en paralelo y con una organización jerárquica que le permite interactuar con algún sistema del mismo modo que lo hace el sistema nervioso biológico".

Cabe destacar que las Redes Neuronales Artificiales, gracias al masivo paralelismo de su estructura, gozan de una serie de ventajas y desventajas:

Tabla 6: Fortalezas y debilidades de las Redes Neuronales Artificiales

Fortalezas	Debilidades
Almacenamiento de información en toda la red: la información, como en la programación tradicional, se almacena en toda la red, no en una base de datos. La desaparición de algunas piezas de información en un lugar no impide que la red funcione.	**Dependencia del hardware:** las redes neuronales artificiales requieren procesadores con potencia de procesamiento en paralelo, de acuerdo con su estructura. Por esta razón, la realización del equipo depende.
Capacidad para trabajar con conocimiento incompleto: después del entrenamiento de ANN, los datos pueden producir resultados incluso con información incompleta. La pérdida de rendimiento aquí depende de la importancia de la información que falta.	**Comportamiento inexplicable de la red:** este es el problema más importante de ANN. Cuando ANN produce una solución de sondeo, no da una idea de por qué y cómo. Esto reduce la confianza en la red.
Tener tolerancia a fallos: la corrupción de una o más celdas de ANN no impide que genere resultados. Esta característica hace que las redes sean tolerantes a fallas.	**Determinación de la estructura de red adecuada:** no existe una regla específica para determinar la estructura de las redes neuronales artificiales. Se logra una estructura de red adecuada a través de la experiencia y la prueba y error.
Tener una memoria distribuida: para que ANN pueda aprender, es necesario determinar los ejemplos y enseñar la red de acuerdo con la salida deseada mostrando estos	**Dificultad para mostrar el problema a la red:** los ANN pueden trabajar con información numérica. Los problemas deben traducirse en valores numéricos antes

ejemplos a la red. El éxito de la red es directamente proporcional a las instancias seleccionadas, y si el evento no se puede mostrar a la red en todos sus aspectos, la red puede producir resultados falsos.	de ser presentados a ANN. El mecanismo de visualización que se determinará aquí influirá directamente en el rendimiento de la red. Esto depende de la capacidad del usuario.
Corrupción gradual: una red se ralentiza con el tiempo y sufre una degradación relativa. El problema de la red no se corroe inmediatamente.	**Se desconoce la duración de la red:** la red se reduce a un cierto valor del error en la muestra significa que la capacitación se ha completado. Este valor no nos da resultados óptimos.
Capacidad para hacer aprendizaje automático: las redes neuronales artificiales aprenden eventos y toman decisiones comentando eventos similares.	
Capacidad de procesamiento en paralelo: las redes neuronales artificiales tienen fuerza numérica que puede realizar más de un trabajo al mismo tiempo.	

La idea detrás de las RNN es hacer uso de información secuencial. En una red neuronal tradicional suponemos que todas las entradas (y salidas) son independientes entre sí. Pero para muchas tareas es una muy mala idea. Si desea predecir la siguiente palabra en una oración, es mejor que sepa qué palabras llegaron antes. Las RNN se denominan recurrentes porque realizan la misma tarea para cada elemento de una secuencia, y la salida depende de los cálculos anteriores. Otra forma de pensar acerca de las RNN es que tienen una "memoria" que captura información sobre lo que se ha calculado hasta ahora. En teoría, las RNN pueden hacer uso de la información en secuencias arbitrariamente largas, pero en la práctica se limitan a mirar hacia atrás solo unos pocos pasos.

Entrenar a un RNN es similar a entrenar una red neuronal tradicional. Debido a que los parámetros son compartidos por todos los pasos de tiempo en la red, el gradiente en cada salida depende no solo de los cálculos del paso de tiempo actual, sino también de los pasos de tiempo anteriores. Por ejemplo, para calcular el gradiente en t = 4, tendríamos que volver a propagar 3 pasos y resumir los gradientes.

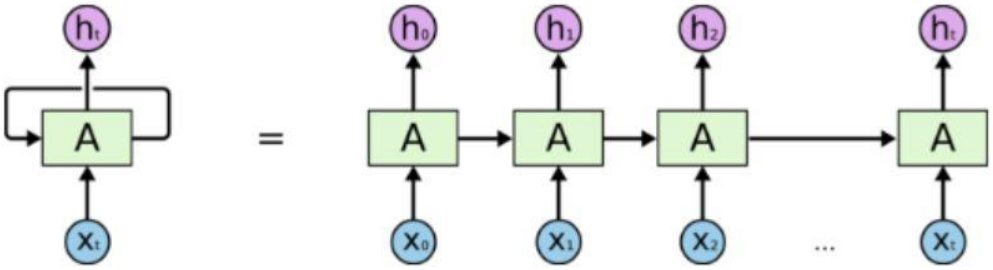


Ilustración 21: Modelo de Red Neuronal Recurrente. Extraído de https://medium.com/@purnasaigudikandula/

A continuación, se detallan las fortalezas y debilidades de las RNN:

Tabla 7: Fortalezas y debilidades de las Redes Neuronales Recurrentes

Fortalezas	Debilidades
Puede modelar la secuencia de datos (es decir, series de tiempo) para que se pueda suponer que cada muestra depende de las anteriores.	Problemas de desaparición y explosión de gradientes.
Se usa con capas convolucionales para extender la vecindad efectiva de píxeles.	Difíciles de entrenar.
	No puede procesar secuencias muy largas.

2.3.5.2. Aplicación del algoritmo de Redes Neuronales Recurrentes

Lo primero que hemos de notar es que no hemos encontrado evidencias de que este algoritmo se use para conjuntos de datos de supervivencia; se puede realizar un modelo, pero es difícil de interpretar.

Lo primero que debemos hacer es transformar los factores en números, y guardar cada variable de factores en variables independientes ($X_{Tratamiento}, X_{MI}, X_{MP}, X_{MTI}$). Hacemos lo mismo con la variable "censura" pero no tenemos en cuenta el tiempo, ya que para este algoritmo no compila correctamente la función *Surv* del paquete {survival}.

Una vez realizado este paso, debemos transformar de número entero a número binario cada variable creada. Por ejemplo, si tenemos la variable $X_{Tratamiento}$ y hemos indicado que el factor "*Dolutegravir*" sea un "1", cuando transformemos a número binario nos quedará (1, 0, 0, 0, 0, 0, 0, 0), y así con cada factor de cada variable. Por último, unificaremos todas las variables independientes en un array.

Entrenamos el modelo con la función *trainr* del paquete {rnn}, y realizamos la predicción con la función *predictr*. De la predicción podemos obtener un gráfico de sectores que nos indique la proporción de pacientes que produzcan fracaso en el tratamiento, como mostramos a continuación:

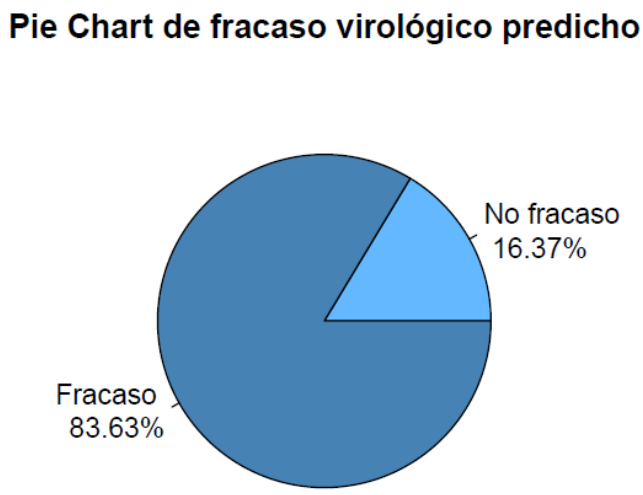


Ilustración 22: Gráfico de proporciones estimadas del modelo de Redes Neuronales Recurrentes

Parece ser que este modelo de red neuronal recurrente es capaz de aprender, en función del tratamiento y las mutaciones, si un paciente puede causar fracaso virológico o no. Al tratarse de una variable respuesta binaria (1 = fracaso, 0 = no fracaso), es fácil de interpretar mediante un gráfico de sectores la proporción de fracasos que podemos encontrar en un futuro.

Aun así, al ser un modelo difícil de interpretar en líneas generales, y no tener evidencias de que este algoritmo se use para conjuntos de datos de supervivencia, no podemos saber a través del modelo cuál es la covariable y factor más influyente a la hora de predecir fracaso virológico.

3. CONCLUSIONES, DISCUSIÓN Y TRABAJO FUTURO

A lo largo de este estudio hemos trabajado diferentes metodologías de análisis de supervivencia y de aprendizaje automático aplicado a ello. Partiendo de la base de una estructuración centrada en una discusión con sus resultados, una conclusión que abarque todo ello y las posibilidades de trabajo futuro que se ofrecen, proponemos una organización de este apartado basado en modelos. Esto es, de cada modelo propuesto para el análisis del tiempo de supervivencia de los pacientes y la posibilidad de fracaso virológico, se ofrece la información recabada sobre el modelo y su posible aplicación y, seguidamente, los resultados que ofrece el modelo concreto. Una vez descrito cada uno de estos apartados de manera breve se aporta, para finalizar, una valoración global de los análisis realizados.

En primer lugar, realizamos el análisis descriptivo de las variables demográficas y clínicas. Mediante varios gráficos, hemos visto que nuestro conjunto de datos está compuesto por individuos con edades comprendidas entre los 38 y los 52 años, y mayoritariamente hombres. El 64% de los pacientes del estudio toma "*Dolutegravir*" como tratamiento antirretroviral para el VIH. De este porcentaje, vemos que el 49% había recibido algún tratamiento antirretroviral previo y, el 44%, comenzó en este estudio desde un escenario de estrategia de cambio ("switch").

Una vez realizado el análisis descriptivo de las variables demográficas y clínicas, procedemos a la aplicación del Modelo de Riesgos Proporcionales de Cox a nuestros datos de estudio. De este análisis, hemos detectado que las covariables más significativas han sido el tratamiento "*Elvitegravir*" y las Mutaciones en la Integrasa; es decir, parece ser que el tratamiento "*Elvitegravir*" nos está influyendo en el tiempo de supervivencia. Concretamente, el efecto que produce es 1.36 veces mayor que los otros tratamientos, lo que hace que se reduzca el tiempo de supervivencia en los pacientes que toman esta medicación. Además, según los resultados, es bueno que no se hayan detectado mutaciones en la integrasa según los resultados de los tests realizados a los pacientes, porque esto nos da mayor probabilidad de supervivencia en los pacientes.

Como ya hemos mencionado, el modelo de Cox se describe como un método semiparamétrico, ya que la distribución del resultado sigue siendo desconocida, incluso si se basa en un modelo de regresión paramétrica. Es por ello por lo que realizamos el análisis de los modelos de vida acelerada para, posteriormente, comparar el modelo resultante con el supuesto de riesgos proporcionales de Cox.

Del análisis de vida acelerada, hemos presentado una tabla con los valores de los estadísticos resultantes de este análisis para cada modelo generado, y concluimos que la distribución Weibull es el modelo que mejor se ajustaría a la respuesta teórica de la supervivencia de los pacientes, frente a otros modelos como la distribución Exponencial, Gamma, Log-Normal y Log-Logistic:

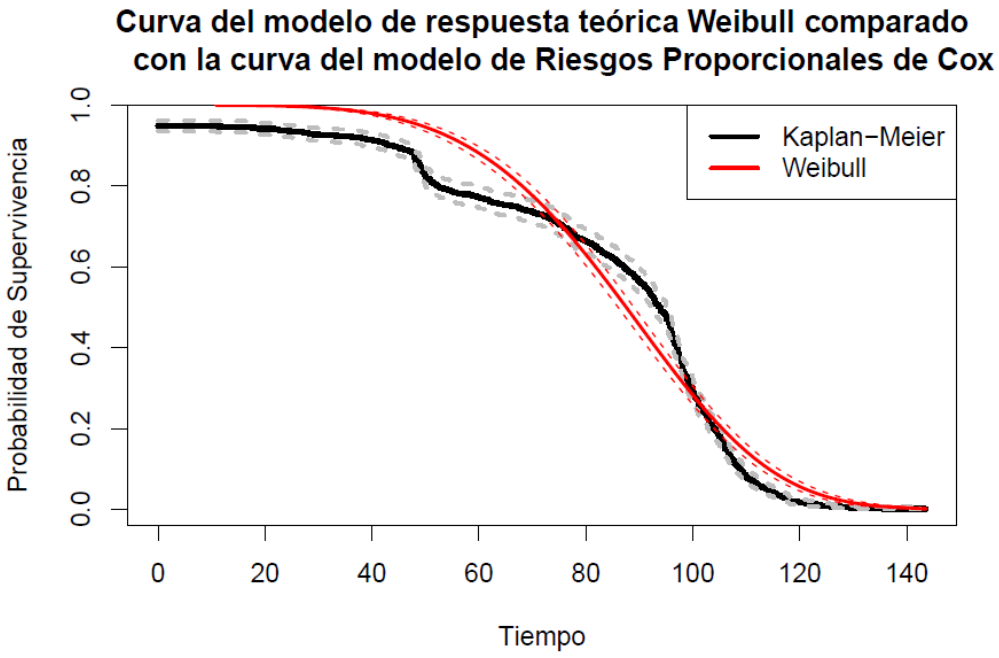


Ilustración 23: Curva del modelo de respuesta teórica Weibull comparado con la curva del modelo de Riesgos Proporcionales de Cox

Una vez hemos finalizado con el análisis de supervivencia, analizamos los algoritmos de Machine Learning aplicados a este análisis, y dividimos este apartado en cinco modelos de estudio.

Comenzamos por los Árboles de Supervivencia ("Survival Trees"). De este análisis vemos que, la covariable más importante, es la mutación en la integrasa seguido del tratamiento, resultado que se asemeja al obtenido en el modelo de Riesgos Proporcionales de Cox. Además, hemos obtenido un valor para el índice de concordancia de 50,3%, por lo que tenemos un buen punto de partida para mejorar este modelo.

De este modelo hemos obtenido los gráficos de las curvas de supervivencia predichas de cada covariable. Concretamente, en el gráfico que contiene las curvas de supervivencia del tratamiento, vemos que el tratamiento "*Elvitegravir*" es el primero que comienza a disminuir la probabilidad de supervivencia de los pacientes, conclusión a la que hemos llegado a través del modelo de Riesgos Proporcionales de Cox y, por tanto, ratificamos con el modelo de árboles de supervivencia. Por esta razón, el siguiente paso sería realizar este mismo modelo únicamente con estas dos variables para seleccionar el árbol final, y aquí podremos aplicar procedimientos como la selección "back" o "forward" para elegir el modelo de árbol óptimo que mejor se ajuste a nuestros datos.

En segundo lugar, hemos realizado un modelo bayesiano. Como resultados más relevantes hemos obtenido las ratios de riesgo de cada covariable, aunque sin mucha diferencia entre estas ratios. De hecho, prácticamente, para todas las covariables explicativas de nuestro modelo se produce un fracaso virológico a la semana, lo que supondría alrededor de 100 muertes al finalizar el estudio; es decir, si repitiésemos este modelo con nuevos individuos, podríamos obtener un fracaso virológico cada semana, por lo que debemos incluir nuevas covariables en el estudio, o decantarnos por las covariables del tratamiento y la mutación en la integrasa, como nos indican los modelos de árboles de supervivencia y riesgos proporcionales de Cox. Pero tampoco podemos fiarnos mucho de estos resultados, porque la información que nos ofrecen los estimadores bayesianos es pobre y no se debe tomar muy en serio las probabilidades que se obtienen.

A continuación, hemos estudiado nuestros datos mediante las Máquinas de Vector Soporte de Supervivencia. Hemos creado una tabla donde mostramos los resultados del índice de concordancia de este modelo, donde hemos tenido en cuenta el tipo de kernel y el tipo de máquina. De hecho, el mejor modelo obtenido es el compuesto por un kernel

del tipo “additive” en un tipo de máquina de regresión; concretamente, obtenemos un índice de concordancia de 58%. Posiblemente, obtengamos un resultado mayor de este índice al reducir el número de covariables y de factores, pero, aunque se aumente o disminuya el número de covariables con sus correspondientes factores, el modelo seguirá siendo igual de efectivo y fiable, independientemente de los recursos que necesite la máquina para ofrecer los resultados esperados. En este sentido, no importa el tiempo que necesite la máquina si sabemos que los resultados que obtendremos son fiables, pero no serán resultados reales dada la complejidad computacional de este algoritmo.

El siguiente modelo aplicado ha sido la Regresión Logística Multitarea (MTLR). La ventaja de este algoritmo es que nos permite hacer una búsqueda del valor óptimo de un parámetro que necesita el algoritmo en base al modelo que le planteamos para, así, obtener el mejor resultado posible. En este sentido, los primeros resultados nos ofrecen un gráfico con las cinco características que tuvieron mayor relevancia a la hora de calcular el modelo; de nuevo, obtenemos que la mutación en la integrasa y el tratamiento (concretamente, “*Elvitegravir*”) son los factores más importantes, como hemos verificado en modelos anteriores y al comienzo de nuestro estudio. Además, este modelo nos permite comparar la curva del modelo de riesgos proporcionales con la curva predicha de supervivencia de cada individuo, y la probabilidad de supervivencia en cada momento t. Lo malo de los resultados que obtenemos con este modelo es que el algoritmo debe realizar una transformación lineal; en este sentido, regresión logística neuronal de tareas múltiples (N-MTLR) ayudará a resolver este problema.

Como último modelo, hemos trabajado con el modelo de Redes Neuronales Recurrentes. Lo malo de trabajar con este modelo has sido la interpretación de los datos y la dificultad para mostrar el verdadero problema planteado en la red, ya que hemos tenido que transformar las covariables en variables binarias, y la variable censura exactamente igual. Este algoritmo no resuelve el problema de la supervivencia en el tiempo, ya que se usa en otros ámbitos, pero es interesante aplicarlo en datos de supervivencia ya que, como resultado final, nos puede ofrecer una predicción de la proporción de pacientes que puedan causar censura. Esto es importante para tener una estimación de, para futuros estudios, qué porcentaje de individuos podrían abandonar el

estudio y qué porcentaje podrían continuarlo hasta el final. Como decimos, no hay menciones de aplicación de este algoritmo en supervivencia, pero puede ser un paso importante desarrollar código de redes neuronales recurrentes aplicadas al análisis de supervivencia, aunque sigamos teniendo el problema de la difícil interpretación de los datos.

A modo de resumen de los resultados obtenidos en los modelos, hemos visto que el tratamiento y la mutación en la integrasa han sido las variables más importantes e influyentes a lo largo del estudio. Pero destacamos tratamiento con "*Elvitegravir*" que, según nos indican los modelos, disminuye la supervivencia en los pacientes por lo que habría que valorar, en función de otras variables, qué tratamiento le produciría menos efectos en su organismo para evitar el fracaso virológico, como podría ser el "*Dolutegravir*" que, según vemos en los resultados, hay menor probabilidad de obtener fracaso virológico. En tal caso, parece relevante la importancia de la realización de las pruebas de las mutaciones, como nos indican los modelos.

Por último, y para finalizar la memoria, la valoración personal va ligada a los conocimientos obtenidos en esta materia. Por la parte teórica, los contenidos adquiridos sobre esta enfermedad y las variables de estudio que se utilizan para la investigación. Este estudio ha sido coordinado por Núria Pérez, de la Fundación Lucha contra el SIDA. Gracias a esta memoria, hemos tenido la oportunidad de aprender más sobre este tema, sobre todo en lo que se refiere a los tratamientos y las pruebas de las mutaciones. Desgraciadamente, el VIH sigue siendo un tema tabú en la sociedad y, la persona que desee informarse debe investigar sobre este tema para aprender lo máximo posible y entender la lucha que se lleva batallando durante muchos años, de manera tanto clínica como humana.

Por la parte práctica, en relación con los algoritmos y la información obtenida de los mismos gracias a este conjunto de datos. El mayor problema que hemos tenido ha sido la ejecución de los algoritmos de Machine Learning aplicados al análisis de supervivencia. El Data Management, la aplicación de los datos a estos modelos y su posterior interpretación, todo basado en el desconocimiento que hemos tenido en estos

algoritmos, pero que han servido para enriquecernos profesional y personalmente. Gracias a estos conocimientos adquiridos en la materia práctica, el siguiente paso es seguir trabajando con estos algoritmos, proponer mejoras y explotar su uso hasta conseguir los resultados esperados basados en la investigación, y librar la batalla no sólo contra el VIH/SIDA, sino con otras enfermedades para conseguir vencerlas y, mientras tanto, mejorar la vida de las personas que las padezcan.

Cabría la posibilidad de que, los resultados obtenidos en esta memoria se compartiesen con la Fundación Lucha contra el SIDA y la comunidad científica en general. Nos planteamos la publicación de la memoria a través de un artículo de divulgación científica o metodológico-estadístico en revistas del sector, para hallar nuevos y mejores métodos en la batalla contra el VIH y las enfermedades infecciosas relacionadas con la enfermedad.

4. GLOSARIO

Sigla	Definición
ADN	Ácido desoxirribonucleico
AFT	Tiempo de Fallo Acelerado
ANN	Artificial Neural Network (Redes Neuronales Artificiales)
ARN	Ácido ribonucleico
AZT	Zidovudina
BN	Bayesian Network (Red Bayesiana)
CDC	Centros para el Control y Prevención de Enfermedades
CIA	Agencia Central de Inteligencia
DTG	Dolutegravir
ELV	Elvitegravir
ETS	Enfermedades de Transmisión Sexual
INI	Inhibidor de la Integrasa
IP	Inhibidor de la Proteasa
ITIAN	Inhibidor transcriptasa inversa análogo de nucleósido
ITINN	Inhibidor transcriptasa inversa no nucleósido
KGB	Comité para la Seguridad del Estado
MI	Mutaciones en la Integrasa
ML	Machine Learning
MP	Mutaciones en la Proteasa
MTI	Mutaciones en la Transcriptasa Inversa
MTLR	Multitask Logistic Regression
NADIR	Recuento hematológico (leucocitos, eritrocitos y plaquetas) más bajo para un paciente dado en un periodo determinado.
NAIVE	Se denomina así a aquéllos que no han tenido tratamiento antirretroviral previo.
NB	Naïve Bayes
NIPS	Neural Information Processing Systems (Sistemas de procesamiento de información neural).

OMS	Organización Mundial de la Salud
ONU	Organización de las Naciones Unidas
ONUSIDA	Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA
PEC	Prueba de Evaluación Continua
PH	Proportional Hazards (Riesgos Proporcionales)
PrEP	Profilaxis preexposición para el VIH
RLT	Raltegravir
RNN	Recurrent Neural Network (Redes Neuronales Recurrentes)
RVM	Relevance Vector Machine
SIDA	Síndrome de Inmunodeficiencia Adquirida
SNS	Sistema Nacional de Salud
SSVM	Survival Support Vector Machine
SVM	Support Vector Machine
VIH	Virus de Inmunodeficiencia Humana

5. REFERENCIAS BIBLIOGRÁFICAS

1. Algortimos De Machine Learning Para Clasificación (SVM). Available from: http://es.mathworks.com/discovery/support-vector-machine.html?s_tid=srchtitle.
2. ALBA CASTRO, J.L. Máquinas De Vectores Soporte (SVM). Universidad de Vigo. Available from: http://web.archive.org/web/20140801145654/http://www.gts.tsc.uvigo.es/~jalba/doctorado/SVM.pdf
3. BRITZ, D., 2015. Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs. Sep 17, Available from: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/.
4. Cesaire J. K. Fouodo, 2018. Survival Support Vector Analysis, Feb 5, vol. 0.0.5. Available from: https://cran.r-project.org/web/packages/survivalsvm/survivalsvm.pdf.
5. Césaire J. K. Fouodo, Inke R. König, Claus Weihs, Andreas Ziegler and Marvin N. Wright, 2018. Support Vector Machines for Survival Analysis with R, Jul, vol. 10/1. Available from: https://journal.r-project.org/archive/2018/RJ-2018-005/RJ-2018-005.pdf.
6. CESIDA. Coordinadora Estatal De VIH Y SIDA. Available from: https://www.cesida.org/.
7. Clara Roca., 2019. Sanidad Financiará La PrEP, La Pastilla De Prevención Del VIH. El Diario: , 10/10/, Available from: https://www.eldiario.es/sociedad/Sanidad-financiara-VIH-PrEP-adquisicion_0_951205277.html.
8. FLSIDA. Fundación Lucha Contra El SIDA. Available from: https://www.flsida.org/es.
9. FOTSO, S., 2018. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. Ene 17, Available from: https://arxiv.org/pdf/1801.05512.pdf.
10. FUNDACIÓN, S. Fundación SEIMC-GESIDA. Available from: http://fundacionseimcgesida.org/.

11. GESIDA. Grupo De Estudio Del SIDA-SEIMC. Available from: http://gesida-seimc.org/.
12. GESTAL POSE, M. Introducción a Las Redes De Neuronas Artificiales. Universidade da Coruña. Available from: http://sabia.tic.udc.es/mgestal/cv/RNAtutorial/rna.html.
13. GIL MARTÍNEZ, C., 2018. Árboles De Decisión Y Métodos De Ensemble. Jun, Available from: https://rpubs.com/Cristina_Gil/arboles_ensemble.
14. GÓMEZ MELIS, G. and CADARSO-SUÁREZ, C., 2017. El Modelo De Riesgos Proporcionales De Cox Y Sus Extensiones. Impacto En Estadística Y Biomedicina. La Gaceta De La RSME, vol. 20, pp. 513-538. Available from: https://grbio.upc.edu/en/shared/gacrsme203cox.pdf.
15. GUDIKANDULA, P., 2019. Recurrent Neural Networks and LSTM Explained, Mar27,. Available from: https://medium.com/@purnasaigudikandula/recurrent-neural-networks-and-lstm-explained-7f51c7f6bbb9.
16. HAIDER, H., 2019. Introduction to the MTLR Workflow. Jun 3, Available from: https://cran.r-project.org/web/packages/MTLR/vignettes/workflow.html.
17. Hemant Ishwaran and Udaya B. Kogalur, 2020. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), Jan 21, vol. 2.9.3. Available from: https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf.
18. IGL, W., 2018. Calculation of Hazard Ratios of Parametric Survival Models in R, Jan 7,. Available from: http://wilmarigl.de/wp-content/uploads/2018/01/tutorial_hr_parsurvmodels.pdf.
19. IMAGINA MÁS. ONG De Salud Sexual, Igualdad Y Diversidad. Available from: https://www.imaginamas.org/inicio/.
20. INFOSIDA. Infosida. Available from: https://www.infosida.es/.
21. Isabel Valdés., 2019. La Profilaxis Del VIH Empezará a Dispensarse En Madrid Este Diciembre. , 05/12/, Available from: https://elpais.com/ccaa/2019/12/04/madrid/1575450059_890719.html.
22. Jesús Herranz Valera, 2015. Análisis De Supervivencia. Alta Dimensionalidad. Instituto IMDEA Alimentación, Nov 6,. Available from:

https://docplayer.es/10845078-Analisis-de-supervivencia-alta-dimensionalidad.html.

23. KLEINBAUM, D.G., 1996. Survival Analysis. New York: Springer ISBN 0387945431.
24. LEE, K. The Function to Implement Bayesian Parametric and Semi-Parametric Regression Analyses for Univariate Time-to-Event Data in the Context of Hazard Regression (HReg) Models. Available from: https://www.rdocumentation.org/packages/SemiCompRisks/versions/3.3/topics/BayesSurv_HReg.
25. LÓPEZ MONTOYA, A.J., 2011. Comparación De Dos Modelos De Regresión En Fiabilidad. Universidad de Granada, Oct,. Available from: https://masteres.ugr.es/moea/pages/tfm1011/comparaciondedosmodelosderegresionenfiabilidad/!
26. MARTÍNEZ, J., 2017. Análisis De Supervivencia En R, May 22,. Available from: http://rstudio-pubs-static.s3.amazonaws.com/316989_83cbe556125645b698c9ff6cf88c4c1a.html#1_introducci%C3%B3n.
27. MIJWEL, M.M., 2018. Artificial Neural Networks. Advantages and Disadvantages. Jan, Available from: https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/#skip-link.
28. Ministerio de Sanidad, Consumo y Bienestar Social., 2019. El Sistema Nacional De Salud (SNS) Financia La PrEP Desde Mañana Como Medida De Prevención Del VIH En Personas De Alto Riesgo. , Oct 31, Available from: https://www.mscbs.gob.es/gabinete/notasPrensa.do?id=4708.
29. National Geographic. Ciencia - SIDA. Available from: https://www.nationalgeographic.es/ciencia/sida.
30. ORELLANA ALVEAR, J., 2018. Arboles De Decision Y Random Forest. Universidad de Cuenca, Nov,. Available from: https://bookdown.org/content/2031/arboles-de-decision-parte-i.html#que-son-los-arboles-de-decision.

31. ROMÁN, V., 2019. Algoritmos Naive Bayes: Fundamentos E Implementación. Apr 25, Available from: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f.
32. STOP SIDA. Salud Sexual LGTB+. Available from: https://stopsida.org/.
33. StopVIH. StopVIH. Available from: https://www.stopvih.org/.
34. UNAIDS - ONUSIDA. Programa Conjunto De Las Naciones Unidas Sobre El VIH/SIDA. Available from: https://www.unaids.org/es.
35. V. Abraira. Análisis De Supervivencia. Available from: http://www.hrc.es/bioest/Supervivencia_1.html.
36. WANG, P., LI, Y. and REDDY, C.K., 2017. Machine Learning for Survival Analysis: A Survey, Aug 15,. Available from: https://arxiv.org/pdf/1708.04649.pdf.
37. Yanying Yang., 2010. Neural Network Survival Analysis. Universiteit Gent, Jun 28,. Available from: https://lib.ugent.be/fulltxt/RUG01/001/458/812/RUG01-001458812_2011_0001_AC.pdf.