Universitat Oberta de Catalunya

uoc.edu

Predicción de tiempo de fracaso de un tratamiento antirretroviral mediante algoritmos de machine learning de supervivencia

Javier Amado Bouza
Máster Universitario de Bioinformática y Bioestadística
Área 2

Nuria Pérez Álvarez
Marc Maceira Duch

12/1/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción de tiempo de fracaso de un tratamiento antirretroviral mediante algoritmos de machine learning de supervivencia*
Nombre del autor:	*Javier Amado Bouza*
Nombre del consultor/a:	Nuria Pérez Álvarez
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	12/1/2021
Titulación::	*Máster Universitario de Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM-Bioinformática y Bioestadística Area 2 aula 1*
Idioma del trabajo:	***Castellano***
Palabras clave	*VIH,Análisis de supervivencia, aprendizaje automático,R, imputación*

Resumen del Trabajo:

En este trabajo se realiza un estudio sobre el desempeño de los algoritmos de machine learning aplicados a los problemas de supervivencia. Se parte de una base de datos (la base de datos Lake) con multitud de datos faltantes, obtenida del ensayo clínico con el mismo nombre y en el que se ensayan dos tratamientos antirretrovirales. La problemática de los datos faltantes es tremendamente común en los estudios longitudinales, entre los que se encuentran los ansayos clínicos. Además se realiza una comparación de la eficacia de dos tratamientos antirretrovirales ensayados. Para ello se realiza todo un primer procesamiento de los datos para prepararlos para un procedimiento de imputación múltiple, llevado a cabo utilizando la librería MICE. Una vez realizada esta imputación, sobre la base de datos imputada se aplicaron 3 algoritmos de machine learning (bosque aleatorio de supervivencia, máquina de soporte vectorial de supervivencia, y boosting). Con el fin de comparar su desempeño, y utilizando las predicciones de los algoritmos se calculó el índice C de Harrell. Finalmente, mediante el test de Wilcoxon para muestras pareadas, se comparó la eficacia de los tratamientos antirretrovirales ensayados.

Abstract:

In this work, a study was carried out on the performance of machine learning algorithms applied to survival problems. The Lake database, obtained from a clinical trial named identically, was th chosen to carry on these protocol. In this clinical trial with antiretroviral treatments were tested. The problem of missing data is extremely common in longitudinal studies, wich include clinical trials. In addition, a comparison of the efficacy of the two antiretroviral treatments tested was made. Firstly, a data manegement process was carried out to prepare the data for a multiple imputation procedure, carried out using the MICE library. Once this imputation was made, 3 machine learning algorithms were applied to the imputed database (random survival forest, survival support vector machine, and boosting). In order to compare their performance, the predictions of the algorithms were used to calculate Harrell's C index. Finally, using the Wilcoxon test for paired samples, the efficacy of the antiretroviral treatments tested was compared.

Memoria final

Javier Amado Bouza

12 de enero, 2021

Tabla de contenidos

Palabras clave: VIH, análisis de supervivencia, aprendizaje automático, R, imputación

1. Introducción

1.1 Contexto y justificación del Trabajo

1.1.1 Contexto y descripción general

El análisis de supervivencia es un subcampo de la estadística, donde la meta es analizar y modelar los datos donde el resultado es el tiempo hasta la ocurrencia de un evento de interés(1).

Una de las problemáticas más importantes de estos estudios longitudinales es la pérdida de seguimiento a pacientes. Esta pérdida desemboca en datos faltantes, lo que puede introducir sesgos lo suficientemente importantes como alterar las conclusiones obtenidas del estudio.

En la actualidad, existe un floreciente campo de la estadística dedicado a desarrollar herramientas para minimizar el efecto de estos missings sobre la representatividad de los datos obtenidos. Esta especialidad se apoya en las capacidades que tienen los sistemas informáticos para realizar multitud de cálculos en un tiempo reducido.

Otra especialidad que se encuentra en un momento de explosión es el aprendizaje automático. En el contexto del análisis de supervivencia, el aprendizaje automático proporciona herramientas para obtener predicciones basadas en el aprendizaje de casos anteriores.

1.1.2 Justificación del Trabajo de Fin de Máster

El seguimiento de los ensayos clínicos para la obtención de una vacuna efectiva y segura frente al SARS-CoV-2, ha alcanzado una importante relevancia en los medios de comunicación de masas. De entre los millones de datos generados en los ensayos clínicos; a buen seguro que ha habido pacientes que se han mudado de domicilio, olvidos de citas para tomar muestras, tubos con muestras que se han roto o extraviado. En definitiva, problemas del mundo real.

En este trabajo se utiliza una base de datos extraída de un ensayo clínico en el que se compara la efectividad de dos tratamientos antirretrovirales frente al VIH. Sobre ella se realiza un procedimiento de imputación múltiple a fin de obtener una base de datos completa sobre la que poner a trabajar tres algoritmos de aprendizaje automático. Finalmente se comparará el desempeño de los algoritmos utilizados, y también se seleccionará el mejor tratamiento antirretroviral de entre los dos testeados.

1.2 Objetivos del Trabajo

1.2.1 Objetivos generales

1. Seleccionar la combinación del método de transformación de datos censurados y algoritmo/os de aprendizaje automático que alcance/n el mejor desempeño en el índice C de Harrell y en la puntuación de Brier.

2. Seleccionar el tratamiento antirretroviral de entre los dos testeados que nos proporcione una menor media de carga viral, o una mayor media en el contaje de linfocitos CD4.

1.2.2 Objetivos específicos

1. Seleccionar la combinación del método de transformación de datos censurados y algoritmo/os de aprendizaje automático que alcance/n el mejor desempeño en el índice C de Harrell y en la puntuación de Brier.

- 1.1-Transformar los datos censurados mediante el método de calibración.
- 1.2-Aplicar el algoritmo de máquina de soporte vectorial de supervivencia.
- 1.3-Aplicar el algoritmo de bosque de supervivencia aleatorio.
- 1.4-Aplicar el algoritmo de boosting.
- 1.5-Evaluar el desempeño de los algoritmos mediante la puntuación de Brier y el índice C de Harrell.
- 1.6-Seleccionar el o los algoritmos con mejor desempeño para esta tarea.

2. Seleccionar el tratamiento antirretroviral de entre los dos testeados que nos proporcione una menor media de carga viral, o una mayor media en el contaje de linfocitos CD4.

- 2.1-Comprobar la normalidad de los datos de la columna "*CargaViral_48*", y aplicar el T-Test o el test de Shapiro-Wilk según haya normalidad o no. Cuanto menor sea la media, mejor.
- 2.2-Comprobar la normalidad de los datos de la columna "*CD4A_48*", y aplicar el T-Test o el test de Shapiro-Wilk según haya normalidad o no. Cuanto mayor sea la media, mejor.
- 2.3-Seleccionar el mejor tratamiento.

1.2.3 Posibles interacciones con la legislación

Los objetivos y las actividades realizadas en este TFM no suponen menoscabo para los derechos de terceras personas:

- No existe ningún tratamiento de datos privados de personas identificadas o identificables, ya que la base de datos utilizada es totalmente anónima. Sólo aparecen en ella los nombres de los doctores que colaboraron en el ensayo clínico, pero sus nombres también aparecen asociados a esta base de datos en las publicaciones realizadas en revistas científicas.

- No existe riesgo para la seguridad de ningún ser humano, animal, ni vegetal. El Trabajo ha sido realizado completamente *in silico*.

- En este proyecto no existe riesgo sobre propiedad intelectual de terceras personas. Todo el software utilizado es gratuito, y no se libera ningún tipo de código informático protegido por copyright.

1.3 Enfoque y método seguido

La estrategia elegida para llevar a cabo el Trabajo es la de optimizar al máximo los datos disponibles con una importante labor de gestión de datos. El principal problema de la base de datos utilizada es que tiene un elevado porcentaje de datos faltantes. Ésto, unido al bajo número de observaciones dificulta enormemente la tarea de obtener resultados fiables con de los que extraer conclusiones.

Para subsanar este problema se realizará una imputaciòn múltiple de los datos. A continuación, se aplicarán tres algoritmos de aprendizaje automático para conseguir predecir el tiempo al fracaso en el tratamiento antirretroviral. Tras esto, se medirá el desempeño de los mismos para elegir al que mejores métricas ofrezca con los datos disponibles.

Desde el punto de vista clínico el objetivo que se quiere conseguir con este TFM es determinar cuál de los tratamientos que se estudian tiene menor tasa de fallo. Para ello se comparará el número de fracasos en el tratamiento de cada uno de los antirretrovirales, y si el número de fracasos es el mismo, se utilizará el incremento en el contaje de linfocitos CD4 absolutos como métrica alternativa para realizar la comparación.

1.4 Planificación del Trabajo

En la PEC1 se definió una lista de objetivos y de actividades, además de una serie de acciones de mitigación para subsanar los inconvenientes que pudiesen surgir durante la realización del proyecto.

Esos objetivos y actividades fueron finalmente modificados debido a que hubo que poner en marcha las acciones de mitigación. A pesar de la aparición de problemas durante el TFM, las acciones de mitigación cumplieron perfectamente su función, y finalmente todos los objetivos (tanto generales como específicos) se cumplieron.

1.4.1 Descripción de los recursos necesarios para realizar el trabajo

Los recursos utilizados para este TFM fueron los siguientes:

- La persona encargada del proyecto. Yo mismo.
- Un ordenador MacBook Pro, con macos versión 10.13.6
- El navegador Mozilla Firefox, que es software gratuito.
- El software de programación estadística R, en su versión 4.0.3. Que se trata de un software gratuito.
- El entorno de desarrollo integrado RStudio, que es software gratuito. Siendo las diferentes librerías utilizadas en este TFM también gratuitas. La versión utilizada fue la 1.3.1093.
- El software de planificación de tareas GanttProject, que es software gratuito.
- El software para gestionar referencias bibliográficas Mendeley Desktop, que es software gratuito.
- El entorno de GIT para escritorio GitHub Desktop, que es software gratuito.

1.4.2 Tareas a realizar

1. PEC0 - Definición de los contenidos del trabajo
- 1.1 Lectura artículo aprendizaje automático for Survival Analysis
- 1.2 Revisión protocolo del ensayo clínico antirretrovirales
- 1.3 Escritura y entrega de la definición de los contenidos del trabajo
2. PEC1 - Redacción del plan de trabajo

- 2.1 Búsqueda información (artículos, libros)
- 2.2 Lectura artículos seleccionados
- 2.3 Escritura y entrega de la redacción plan de trabajo

3. PEC2 - Desarrollo del trabajo - Fase 1

- 3.1 Gestión de datos: búsqueda y eliminación datos incongruentes
- 3.2 Aprendizaje método calibración
- 3.3 Aprendizaje imputación múltiple
- 3.4 Gestión de los datos previa a imputación múltiple
- 3.5 Imputación múltiple con la librería MICE
- 3.6 Aprendizaje máquina de aprendizaje extremo
- 3.7 Aprendizaje bosque de supervivencia aleatorio
- 3.8 Aprendizaje del algoritmo aprendizaje activo
- 3.9 Redacción y entrega del informe desarrollo trabajo fase 1

4. PEC3 - Desarrollo del trabajo - Fase 2

- 4.1 Aprendizaje y aplicación del algoritmo de Boosting
- 4.2 Aprendizaje puntuación de Brier
- 4.3 Evaluación mediante puntuación de Brier
- 4.4 Aprendizaje algoritmo máquina de soporte vectorial de supervivencia
- 4.5 Aplicación algoritmo máquina de soporte vectorial de supervivencia
- 4.6 Aplicación bosque de supervivencia aleatorio
- 4.7 Aprendizaje índice C de Harrell
- 4.8 Evaluación mediante índice C de Harrell
- 4.9 Conclusiones generales tras aplicación algoritmos y evaluación del desempeño
- 4.10 Comparación de las medias de los biomarcadores elegidos
- 4.11 Elección del mejor tratamiento
- 4.12 Redacción informe desarrollo trabajo fase 2

5. PEC4 - Cierre de la memoria

- 5.1 Redacción y entrega de la memoria definitiva

6. PEC5a - Elaboración de la presentación

- 6.1 Aprendizaje herramienta presenta
- 6.2 Elaboración y entrega de la presentación

7. PEC5b - Defensa pública

- 7.1 Respuesta a las preguntas del Tribunal

1.4.3 Planificación de las tareas mediante diagrama de Gantt

Las tareas planificadas para cada una de las entregas parciales de este TFM son las siguientes:

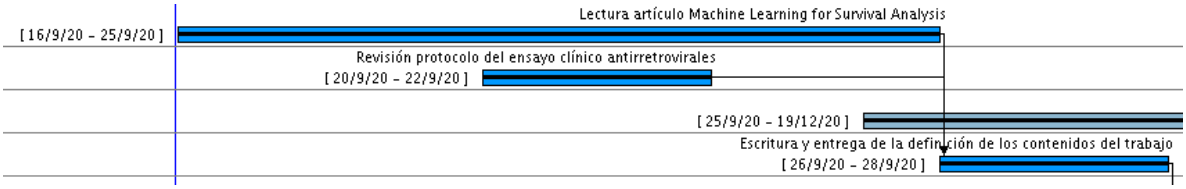

Figura 1: Tareas planificadas para la PEC_0

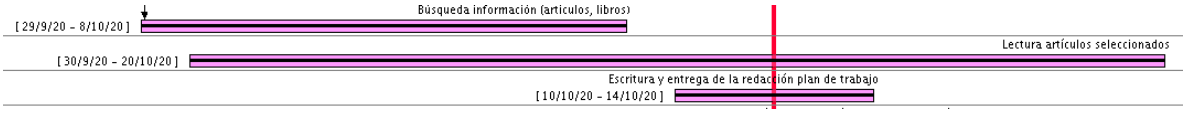

Figura 2: Tareas planificadas para la PEC_1

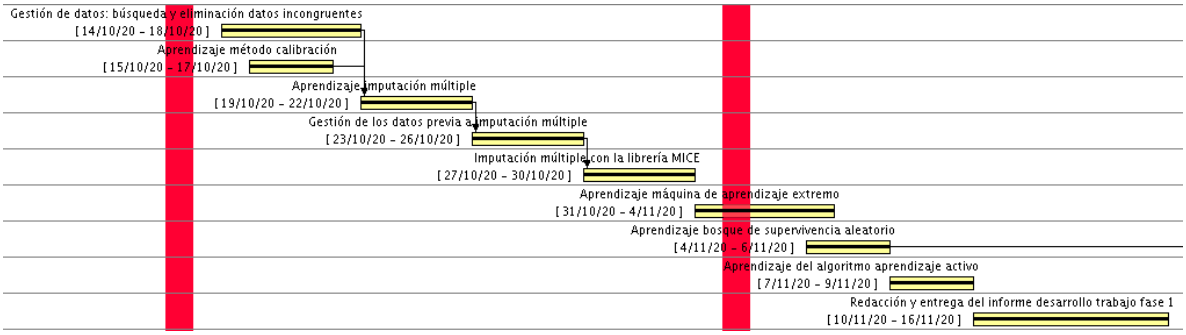

Figura 3: Tareas planificadas para la PEC_2

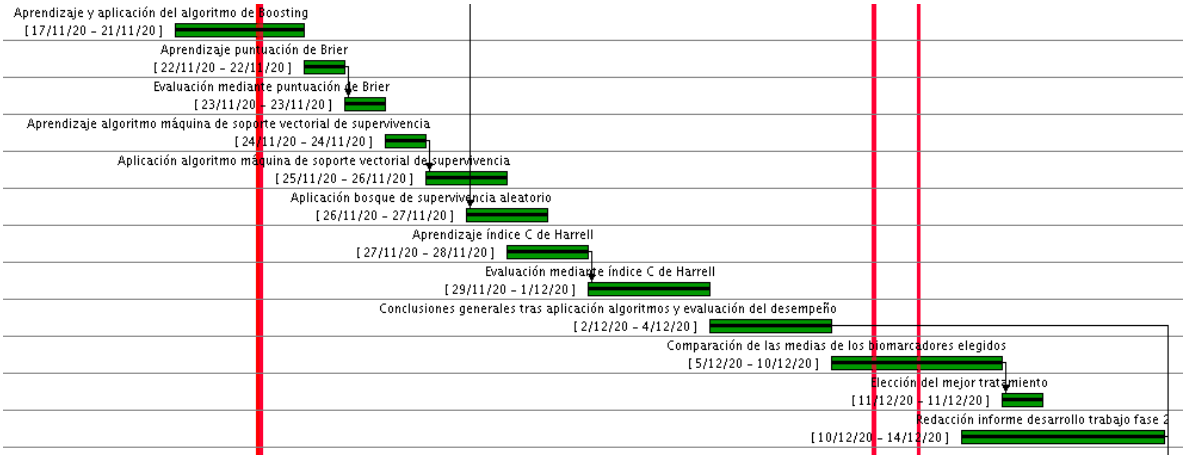

Figura 4: Tareas planificadas para la PEC_3

Figura 5: Tareas planificadas para la PEC_4

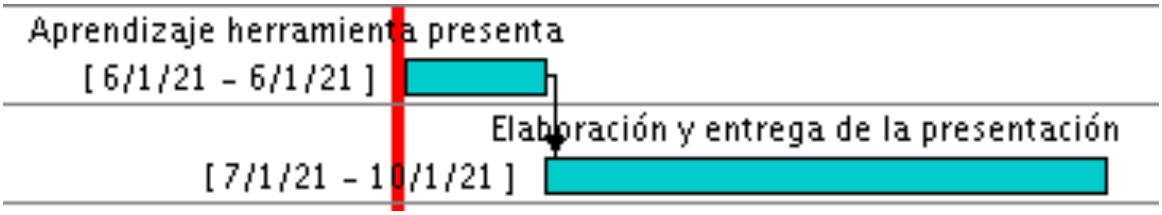

Figura 6: Tareas planificadas para la PEC_5a

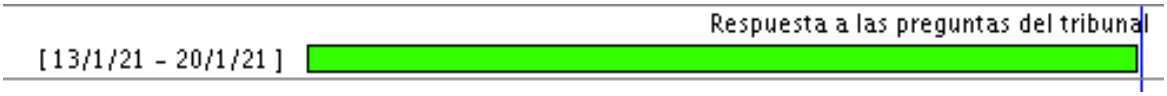

Figura 7: Tareas planificadas para la PEC_5b

Fue imposible una acotación de trabajo a unas horas del día en particular, debido a la rotación de horarios en el trabajo. Los fines de semana también se consideraron hábiles para el TFM, aunque no en todos los fines de semana se dedicó tiempo a la realización de tareas para el TFM.

Con respecto a los recursos disponibles para realizar este TFM, éstos fueron los siguientes:

Nombre	Función
Javier Amado Bouza	Encargado del proyecto
MacBook Pro	Indefinido
Mozilla Firefox	Indefinido
RStudio	Indefinido
GanttProject	Indefinido
Mendeley Desktop	Indefinido
RMarkdown	Indefinido
Bibliografía	Indefinido
GitHub Desktop	Indefinido
LibreOffice	Indefinido

Figura 8: Recursos disponibles

1.5 Breve sumario de productos obtenidos

Se ha obtenido un pipeline completo, que con las pertinentes adaptaciones a la base de datos concreta, sirve para trabajar con bases de datos obtenidas en ensayos clínicos.

Este pipeline abarca desde la gestión de datos al inicio del proceso, pasando por una imputación múltiple de los datos faltantes. Hasta la utilización de un algoritmo de aprendizaje automático,que utilizando esos datos como base, prediga con la mayor precisión posible el tiempo al fracaso de un tratamiento antirretroviral.

Finalmente también se obtendrá un código de análisis para dilucidad cualquier comparación entre tratamientos. En este caso antirretrovirales.

Este pipeline está escrito íntegramente en lenguaje R. El código completo ha sido liberado, y está en el siguiente repositorio de GitHub:

https://github.com/amadobouza/TFM.git

También se ha obtenido un diagrama de Gantt en formato .gan, que refleja las tareas realizadas durante el desarrollo del TFM.

Esta memoria definitiva es el tercer producto obtenido.

El cuarto producto es la base de datos completa tras la imputación, almacenada en formato .csv.

El último de los productos obtenidos es la presentación que se colgará en la aplicación present@, de la UOC.

Todos estos productos, salvo la presentación, se adjuntarán junto con la memoria definitiva para su evaluación por parte del Tribunal.

1.6 Breve descripción de los otros capítulos de la memoria

El punto 2 trata del análisis de supervivencia. Tratando sobre la censura y los pricipales algoritmos de aprendizaje automático para datos de supervivencia..

El punto 3 trata sobre los datos faltantes, su clasificación y cómo tratarlos mediante imputación.

El punto 4 trata los antirretrovirales, ya que los datos utilizados en este TFM vienen de un ensayo clínico que compara dos tratamientos antirretrovirales.

El punto 5 es el procedimiento utilizado en este TFM, donde se describe linealmente los pasos dados para alcanzar los objetivos generales y específicos.

El punto 6 son las conclusiones y la autoevaluación. En este punto se razonan y evalúan los resultados obtenidos, se analiza el camino llevado a cabo en el TFM para medir el rendimiento, se identifican las causas de éxito y fracaso, y se extraen conclusiones para enfocar mejor los proyectos posteriores.

El punto 7 es el glosario, en el que se definen los términos y acrónimos más relevantes utilizados dentro de la Memoria.

El punto 8 es la bibliografía, en la que quedan resumidas las referencias utilizadas en este TFM.

2. El análisis de supervivencia

2.1 Introducción

El principal objetivo del análisis de supervivencia es obtener una estimación del tiempo transcurrido hasta un evento de interés. Bien sea el desarrollo de una enfermedad, la rotura de una pieza, o la consecución de la meta de financiación para un proyecto de crowdfunding. Para ello se miden una serie de covariables o predictores que colaborarán en la estimación de este tiempo hasta el evento.

Por otra parte, el principal desafío del análisis de supervivencia es la existencia de instancias en las cuales la ocurrencia del evento se vuelve inobservable, debido a causas aleatorias independientes del evento de interés. A esas instancias se les llama censuradas. Por lo general, en las tareas de minería de datos censuradas las instancias censuradas se eliminan o los valores faltantes se imputan para convertir los datos censurados en datos sin censura(1).

Esta censura hace que las herramientas estadísticas utilizadas para el análisis de supervivencia sean muy específicas de este campo. Hasta los algoritmos de aprendizaje automático deben ser adaptados para trabajar con datos de supervivencia.

Para finalizar con la fotografía actual del campo del análisis de supervivencia, indicar que se trata de un campo de investigación profundamente fragmentado entre varias disciplinas que lo estudian. Eso ha provocado que existan pocos artículos sobre la utilización de aprendizaje automático en el análisis de supervivencia. Y no hay ningún *review* sobre los últimos avances de aprendizaje automático aplicados al análisis de supervivencia (2).

2.2 Tipos de censura, función de supervivencia y función

Según en qué momento se produce la censura, ésta se puede dividir en tres tipos:

- Censura por la derecha - Es la más común. En ella, el tiempo de las instancias censuradas es mayor o igual que el de la última observación realizada.
- Censura por la izquierda - En este tipo de censura, el tiempo de las instancias censuradas es menor o igual al del momento de la observación.
- Censura de intervalo - En la cual sólo se sabe que el evento ha sucedido durante un determinado intervalo de tiempo.

Se llama T al tiempo de supervivencia, que es el valor del tiempo en el que sucedió el evento. T sólo se conoce para aquellos sujetos del estudio que han sufrido el evento.

Se llama C al tiempo censurado. Su valor será el del momento de la pérdida de observación, el abandono del estudio, o la finalización del mismo.

Existen dos funciones básicas en el análisis de supervivencia:

La *función de supervivencia*, denotada por S, que se utiliza para representar la probabilidad de que el tiempo al evento de interés no sea menor que un tiempo *t* dado. Su valor es 1 cuando t = 0 lo que indica que en ese momento el 100% de los sujetos no han sufrido el evento, a partir de ese momento desciende monótonamente (2).

$$S(t) = Pr(T \geq t).$$

Figura 9: Función de supervivencia

La *función de riesgo*, denotada por h, es la tasa de eventos en el momento *t* dado que no ha sucedido el evento antes del tiempo *t* (2).

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)}$$

Figura 10: Función de riesgo

Siendo F(t) la función de distribución acumulada de muertes, y f(t) la función de densidad de muertes.

2.3 Métodos tradicionales utilizados en el análisis de supervivencia

Existen tres tipos diferentes de métodos estadísticos utilizados para estimar las funciones de supervivencia y riesgo (2).

Tipo	Ventajas	Desventajas	Métodos específicos
No Paramétricos	Son más eficientes cuando no se conoce la distribución teórica de los datos	* Difíciles de interpretar * Producen estimaciones imprecisas	* Kaplan-Meier * Nelson-Aalen * Tablas de vida
Semiparamétricos	No es necesario el conocimiento de la distribución subyacente de los tiempos de supervivencia	* La distribución de los resultados es desconocida * No son fáciles de interpretar	* Modelo de Cox * Cox regularizado * CoxBoost * Cox dependiente de tiempo
Paramétricos	*Interpretación sencilla * Más eficientes y precisos cuando los tiempos de supervivencia siguen una distribución particular	Cuando se viola la condición de la distribucion, pueden volverse inconsistentes y producir resultados subóptimos	* Tobit * Buckley-James * Regresión penalizada * Tiempo de fallo acelerado

Figura 11: Métodos utilizados en el análisis de supervivencia

2.3.1 Modelos no paramétricos

Dentro de estos modelos, el método más utilizado es el de Kaplan y Meier. Éstos desarrollaron la curva de Kaplan y Meier o estimador producto limite para estimar la función de supervivencia utilizando la duración real del tiempo estimado.

Sea $T_1 < T_2 < \ldots < T_K$ un conjunto ordenado de distintos tiempos de evento observados para N(K =< N) instancias. Además de esos tiempos de evento, también hay tiempos de censura para las instancias sin evento observado. Para un tiempo de evento específico T_j(j = 1,2,…,K), el número de eventos observados es d_j>= 1, y r_j instancias se consideran “en peligro” ya que sus tiempos hasta el evento o sus tiempos de censura son mayores que T_j. El valor de r_j se calcula como $r_j = r_{j-1} - d_{j-1} - c_{j-1}$. Por lo que la probabilidad condicional de sobrevivir más allá del tiempo T_j queda definida como (2):

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

Figura 12: Ecuación de Kaplan y Meier

2.3.2 Modelos semiparamétricos

Son un híbrido entre los modelos paramétricos y los no paramétricos.

Si los comparamos con los modelos paramétricos, pueden obtener un estimador más consistente bajo un rango más amplio de condiciones. Y si los comparamos con los no paramétricos, pueden obtener un estimador más preciso (2).

El método de Cox es el método semiparamétrico más utilizado. En él la función de riesgo basal, $h_0(t)$ no está especificada.

En este método, la función de riesgo presupone riesgos proporcionales dados por:

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

Figura 13: Función de riesgo de Cox

Siendo $h_0(t)$ (función de riesgo basal) una función arbitraria no negativa de tiempo. $X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})$ es el vector de covariables para la instancia *i*, y $\beta^T = (\beta_1, \beta_2, \ldots, \beta_p)$ es el vector de coeficientes.

Y la función de supervivencia:

$$S(t) = exp(-H_0(t)exp(X\beta)) = S_0(t)^{exp(X\beta)}$$

Figura 14: Función de supervivencia de Cox

El éxito del método de Cox hace que se hayan desarrollado modelos de Cox con diferentes funcionalidades (2):

- Modelos de Cox regularizados (Lasso-Cox, Ridge-Cox, EN-Cox, OSCAR-Cox) : Para los casos de datos con un número de variables mayor que el de indivíduos.
- CoxBoost: Para los casos en los que haya que incluir obligatoriamente alguna variable a la hora de crear el modelo, ya que el resto de modelos no tienen esta capacidad.
- Modelo de Cox tiempo dependiente: Para casos en los que existen covariables cuyos valores cambian con los valores de t para una misma instancia.

2.3.3 Modelos paramétricos

Los modelos de regresión paramétricos censurados asumen que los tiempos de supervivencia, o el logaritmo de los tiempos de supervivencia de todas las instancias de los datos siguen una distribución teórica particular. Estos modelos suponen una alternativa importante a los modelos semiparamétricos basados en Cox. Las distribuciones teóricas más utilizadas son: normal, exponencial, weibull, logística, log-logística, y log-normal (2).

Si los tiempos de supervivencia en todas las instancias siguen esas distribuciones, el modelo se conoce como modelo de regresión lineal. Aunque no se puede aplicar directamente la regresión lineal con el método de mínimos cuadrados a los problemas de supervivencia, debido a la existencia de instancias censuradas. Se han propuesto modelos lineales para manejar instancias censuradas. Aquí se listan algunos de ellos (2):

- Regresión Tobit: Se introduce una variable latente y^* y se asume que ésta depende linealmente de X de la siguiente manera: $y^* = X\beta + \epsilon$, $\epsilon \sim N(0,\sigma^2)$. Siendo ϵ el error con distribución normal.

- Regresión Buckley-James: Estima el tiempo de supervivencia de las instancias censuradas como el valor de respuesta basado en el método de estimación de Kaplan-Meier. Y entonces ajusta un modelo lineal de tiempo de fallo acelerado, considerando los tiempos de supervivencia de las instancias no censuradas y los tiempos aproximados de supervivencia de las instancias censuradas al mismo tiempo.

- Regresión penalizada: Puede proporcionar mejores resultados de predicción en la presencia tanto de multicolinearidad de las covariables como de alta dimensionalidad.

- Regresión ponderada: Tiene su aplicación cuando se incumple la asunción de varianzas constantes de los residuos para la regresión de mínimos cuadrados (heterocedasticidad). El efecto producido es el de minimizar la suma de los resíduos al cuadrado ponderados.

- Regularización estructurada: La habilidad para inferir efectivamente conocimiento latentea través de jerarquías basadas en árboles y relaciones basadas en grafos es extremadamente crucial en análisis de supervivencia. La aproximación estructurada es más robusta comparada con el los métodos

estadísticos estándar y aquellos basados en Cox, ya que ésta puede adaptarse automáticamente a diferentes distribuciones de eventos e instancias censuradas.

Si es el logaritmo de los tiempos de supervivencia de todas las instancias el que sigue esas distribuciones, el problema puede ser analizado utilizando el modelo de tiempo de fallo acelerado (AFT), en el cual se asume que la variable puede afectar al tiempo hasta el evento de interés de una instancia por un factor constante (2).

2.4 Algoritmos de aprendizaje automático utilizados en análisis de supervivencia

Algunos métodos de aprendizaje automático como redes neuronales, bosques aleatorios, y máquinas de soporte vectorial han sido utilizados para realizar predicciones utilizando datos censurados. La particularidad de estos métodos para los datos censurados es que pueden manejar relaciones no lineales entre las covariables(1).

2.4.1 Algoritmos clásicos

Árboles de supervivencia

Los árboles de supervivencia son un tipo de árboles de clasificación y regresión hechos a medida para manejar datos censurados. La intuición básica tras los modelos de árbol es la de partir los datos basándose en un criterio particular de división. Y los objetos que son similares entre ellos sobre la base de un evento de interés, se ubicarán en el mismo nodo (2).

La principal mejora de un arbol de supervivencia sobre un árbol de decisión estándar es su habilidad para manejar los datos censurados utilizando la estructura de árbol (2).

Métodos bayesianos

Existen dos tipos de algoritmos de aprendizaje automático que utilizan el teorema de Bayes: Bayes ingenuo, y red bayesiana.

Los resultados experimentales al cualquiera de estos dos métodos sobre datos de supervivencia nos muestran que los métodos bayesianos tienen buenas propiedades de interpretabilidad y razonamiento de incertidumbre (2).

Bayes ingenuo es uno de los algoritmos más simples y al mismo tiempo más efectivos. Un defecto del algoritmo de Bayes ingenuo es que realiza la asunción de independencia entre todas las características, y eso puede no ser cierto en muchos de los problemas de análisis de supervivencia (2).

En cambio, en las redes bayesianas las características pueden estar relacionadas unas con las otras en varios niveles. Este algoritmo puede representar visualmente todas las relaciones entre las variables , los cual lo hace fácilmente interpretable para el usuario final (2).

Redes neuronales artificiales

Inspiradas en los sistemas neuronales biológicos, las redes neuronales artificiales han sido ampliamente utilizadas en el análisis de supervivencia (2).

Máquinas de soporte vectorial

Se trata de una aproximación muy exitosa de aprendizaje supervisado. Utilizadas en clasificación, pero que pueden ser modificadas para problemas de análisis de supervivencia (2).

2.4.2 Métodos combinados de aprendizaje

Una de las nuevas aproximaciones son los métodos combinados de aprendizaje. Éstos, generan un comité de clasificadores y luego predicen las etiquetas de clase para los nuevos puntos de datos venidero. Esto se hace tomando un voto ponerado entre los resultados de predicción de todos estos clasificadores (2).

Los **árboles de supervivencia con agregación de bootstrap** reducen la varianza de los modelos de base. El procedimiento que utilizan es el siguiente (2):

- 1 - Mediante *bootstrap*, se crean muestras de los datos de entrenamiento.
- 2 - Para cada una de esas muestras se construye un arbol de supervivencia, y se asegura el hecho de que para todos los nodos terminales el número de eventos es mayor o igual al umbral.
- 3 - Al promediar las predicciones de los nodos hoja, se calcula la función de supervivencia agregada de bootstrap. Para cada nodo hoja, la función de supervivencia se estima utilizando el estimador de Kaplan-Meier. Se asume que todos los indivíduos en el mismo nodo tienen la misma función de supervivencia.

Los **árboles aleatorios de supervivencia** extendieron el método de bosques aleatorios de Breiman mediante la utilización para la predicción de un bosque de árboles aleatorios de supervivencia. Los pasos dados en este caso son (2):

- 1 - Mediante *bootstrap* se crean muestras de los datos de entrenamiento. Esto se llama *Out Of Bag*(OOB), ya que aproximadamente el 37% de los datos son excluidos en cada muestra.

- 2 - Para cada muestra se construye un árbol de supervivencia mediante la selección aleatoria de características. Y se divide el nodo utilizando la característica candidata que puede maximizar la diferencia de supervivencia entre los nodos secundarios.
- 3 - Se hace crecer el árbol a su tamaño completo bajo la restricción de que un nodo terminal debe tener no menos de 0 muertes únicas.
- 4 - Utilizando el estimador no paramétrico de Nelson-Aalen, se calcula la función de riesgo acumulativo (CHF) de conjunto de los datos OOB tomando el promedio del CHF de cada árbol.

Este algoritmo de aprendizaje automático tiene varias ventajas (3):

- Cuando el objetivo es la predicción, los árboles aleatorios de supervivencia son un método atractivo para construir un modelo.
- Otra característica atractiva es que no imponen una estructura restrictiva sobre cómo se deben combinar las variables. Si la relación entre las variables predictoras y la variable de respuesta es compleja con patrones e interacciones no lineales, entonces la RF es capaz de incorporar esto.

Un bosque aleatorio de supervivencia puede proporcionar una medida de la importancia de una variable, y se denomina medida de importancia de la variable (VIMP). Esta medida nos indica cuánto empeoraría la predicción si esa variable no estuviera disponible. En el resultado también se proporciona el error de predicción. Para otros modelos de predicción esto se denomina índice C y está relacionado con el área bajo la curva ROC. El error de predicción es la fracción de veces que para un par de sujetos la persona que se predijo que viviría más tiempo en realidad murió antes, es decir, la predicción clasificó incorrectamente a estas dos personas. Por tanto, una tasa de error de predicción pequeña es buena y sería deseable tasas de error inferiores al 25%. Las tasas de error de predicción del 50% o más son inútiles porque no son mejores que lanzar una moneda (3).

El método de *boosting* es uno de los métodos de conjunto más utilizados, diseñado para combinar a varios “aprendices de base” en una suma ponderada que representa el resultado final del “aprendiz fuerte”. Se ajusta de forma iterativa a los residuos definidos de forma adecuada según el algoritmo de descenso de gradiente.

2.4.3 Aprendizaje activo

El aprendizaje activo basado en los datos con observaciones censuradas puede ser muy útil para el análisis de supervivencia, ya que las opiniones de un experto en el dominio pueden incorporarse a los modelos. Su mecanismo permite al modelo de supervivencia seleccionar un subconjunto de sujetos aprendiendo primero de un

conjunto limitado de sujetos etiquetados y luego consultar al experto para obtener la etiqueta del estado de supervivencia antes de considerarlo en el conjunto de entrenamiento. La retroalimentación del experto es particularmente útil para mejorar el modelo en muchos dominios de aplicación del mundo real. El objetivo del aprendizaje activo para los problemas de análisis de supervivencia es construir un modelo de regresión de supervivencia utilizando las instancias censuradas por completo sin eliminar o modificar la instancia (2).

Dentro del aprendizaje activo. Un algoritmo muy utilizado es el algoritmo de regresión de Cox regularizado activo (ARC). Está basado en una estrategia de muestreo de gradiente discriminativo mediante la integración del método de aprendizaje activo con el modelo de Cox. El marco ARC es un algoritmo basado en iteraciones con tres pasos principales (2):

- 1 - Se construye una regresión de Cox regularizada utilizando los datos de entrenamiento
- 2 - Se aplica el modelo obtenido anteriormente a todas las instancias en el grupo sin etiquetar
- 3 - Se actualizan los datos de entrenamiento y el grupo sin etiquetar, y se seleccione la instancia cuya influencia en el modelo sea mayor. Se marca esta instancia antes de ejecutar la siguiente iteración.

2.4.4 Transferencia de aprendizaje

La recopilación de información etiquetada en los problemas de supervivencia lleva mucho tiempo, es decir, uno tiene que esperar a que ocurra el evento de un número suficiente de instancias de entrenamiento para construir modelos sólidos. Para ello la transferencia de conocimiento entre tareas relacionadas generalmente producirá resultados mucho mejores en comparación con un enfoque de integración de datos. El método de aprendizaje por transferencia se ha estudiado ampliamente para resolver problemas estándar de regresión y clasificación. Se ha propuesto un modelo de daños proporcionales de Cox regularizado llamado Transfer-Cox, para mejorar el rendimiento de predicción del modelo de Cox en el dominio de destino a través de la transferencia de conocimiento desde el dominio de origen en el contexto de modelos de supervivencia construidos en múltiples conjuntos de datos de alta dimensión. El modelo Transfer-Cox emplea la norma $l_{2,1}$ para penalizar la suma de las funciones de pérdida (probabilidad logarítmica parcial negativa) tanto para los dominios de origen como de destino. Por lo tanto, el modelo, con complejidad de tiempo O (NP), no solo seleccionará características importantes, sino que también aprenderá una representación compartida en los dominios de origen y destino para mejorar el rendimiento del modelo en la tarea de destino (2).

2.4.5 Aprendizaje multitarea

El problema de la predicción del tiempo de supervivencia se reformula como un problema de aprendizaje de múltiples tareas. En los datos de supervivencia, la matriz de etiquetado de resultados está incompleta, ya que la etiqueta de evento de cada instancia censurada no está disponible después de su tiempo de censura correspondiente; por lo tanto, no es adecuado manejar la información censurada usando los métodos estándar de aprendizaje multitarea. Para resolver este problema, el modelo de aprendizaje multitarea para el análisis de supervivencia (MTLSA) traduce las etiquetas originales de eventos a una matriz de indicadores N x K, donde K = max (y_i) ($\forall_i$ = 1, …,N) es el tiempo máximo de seguimiento de todas las instancias en el conjunto de datos. El elemento I_{ij} (i = 1, …,N; j = 1, … ,K) de la matriz del indicador será 1 si el evento ocurrió antes del tiempo y_j por ejemplo i, de lo contrario será 0. Uno de Las principales ventajas del enfoque MTLSA es que puede capturar la dependencia entre los resultados en varios puntos de tiempo mediante el uso de una representación compartida entre las tareas relacionadas en la transformación, lo que reducirá el error de predicción en cada tarea. Además, el modelo puede aprender simultáneamente de instancias censuradas y no censuradas basándose en la matriz de indicadores. Una característica importante de los eventos no recurrentes, es decir, una vez que el evento ocurre, no volverá a ocurrir, se codifica mediante la restricción de estructura de lista no negativa y no creciente. En el algoritmo MTLSA, la penalización de la norma $l_{2,1}$ se emplea para aprender una representación compartida, con complejidad de tiempo O (NPK), a través de tareas relacionadas y, por lo tanto, calcular la relación entre los modelos individuales construidos para varios puntos de tiempo de eventos únicos (2).

2.5 Medidas de evaluación del desempeño

2.5.1 Índice C de Harrell

En el análisis de supervivencia, una forma común de evaluar un modelo es considerar el riesgo relativo de un evento para una instancia diferente, en lugar de los tiempos absolutos de supervivencia para cada instancia. Esto se puede hacer calculando la probabilidad de concordancia o el índice de concordancia (índice C). Los tiempos de supervivencia de dos instancias se pueden ordenar para dos escenarios: 1- ambos no están censurados; 2- el tiempo de evento observado de la instancia sin censura es menor que el tiempo de censura de la instancia censurada (2).

Para una instancia censurada, la comparación sólo se puede hacer con una instancia no censurada con menor valor de tiempo. Sin embargo, cualquier instancia no puede

ser comparada con con ninguna otra instancia, sea del tipo que sea (censurada o no censurada) tras su tiempo de censura (2).

Para los métodos de supervivencia los cuales permiten aprender directamente el tiempo de supervivencia, el índice C de Harrell debe ser calculado como (2):

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

Figura 15: Índice C de Harrell

2.5.2 Puntuación de Brier

Esta puntuación fue desarrollada para predecir la inexactitud de los pronósticos meteorológicos probabilísticos. Solo puede evaluar los modelos de predicción que tienen resultados probabilísticos; es decir, el resultado debe permanecer dentro del rango [0,1], y la suma de todos los resultados posibles para un determinado individuo debe ser 1 (2).

Inicialmente la puntuación de Brier no estaba adaptada a los probleas de supervivencia, pero se extendió el modelo, y ahora también cumple la función de ser una medida del desempeño para problemas de supervivencia con información censurada. Para evaluar esos modelos de predición el resultado debe ser binario o de naturaliza categórica. Esa puntuación de Brier puede ser calculada de la siguiente manera

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t)[\hat{y}_i(t) - y_i(t)]^2$$

Figura 16: Puntuación de Brier

Siendo w_i el peso de la iésima instancia, estimado mediante la incorporación del estimador Kaplan-Meier de la distribución de censura. Con una muestra de N instancias, y para cada X_i (i = 1,2,…,N), el resultado predicho al tiempo t es $\hat{y}_i(t)$ y el resultado real es y_i.

3. Los datos faltantes

3.1 Introducción

Cualquier toma de datos en el mundo real puede traer consigo sucesos que produzcan pérdida de datos. Desde la avería de la máquina que está midiendo unas muestras de sangre, al abandono de un estudio longitudinal por parte de un paciente (sea por las razones que sea), pasando por olvidos cometidos por el propio investigador.

En el caso particular que atañe a este TFM, Una proporción sustancial de sujetos a menudo abandona los estudios longitudinales antes de completar el estudio. Los abandonos pueden no ser despreciables, en el sentido de que los métodos que ignoran el mecanismo que conduce al abandono suelen producir resultados sesgados (4).

Durante casi un siglo, los científicos han estado lidiando con los datos faltantes eliminando o completando arbitrariamente los casos perdidos posthoc. Estas técnicas son propensas al sesgo en la medida en que los resultados del estudio carecen de significado, pero siguen utilizándose. Durante las últimas tres décadas, se han logrado grandes avances en el desarrollo de técnicas analíticas para estimar los efectos causales en presencia de datos faltantes. La mayor utilización de métodos como la ponderación de probabilidad inversa, la imputación múltiple y el análisis basado en la probabilidad mejoró enormemente el rigor sobre los métodos ad hoc (por ejemplo, última observación llevada adelante, análisis de caso completo) que anteriormente dominaban el panorama de los RCT. Aún así, es importante comprender que estos métodos son herramientas más que soluciones. Cuando faltan datos, el resultado de cualquier análisis estadístico se basa en los supuestos no verificables sobre la relación entre los datos no observados y las razones por las que faltan. En otras palabras, las conclusiones extraídas de los ensayos clínicos con datos faltantes pueden variar según las suposiciones realizadas y el método analítico elegido (5).

El principio de intención de tratar (ITT) requiere la inclusión completa de todos los datos de todos los pacientes aleatorizados en el análisis y se considera el criterio más apropiado para evaluar la utilidad de una nueva terapia. En un análisis ITT, todos los participantes asignados al azar tienen resultados evaluados y se analizan en el grupo en el que fueron asignados al azar (independientemente de la intervención real recibida). Cuando los participantes abandonan o pierden visitas, lo que genera datos faltantes, la capacidad de realizar un análisis de intención de tratar y sacar conclusiones sobre un vínculo causal se ve comprometida (5).

*En la literatura existente sobre el abandono en estudios longitudinales, se asume comunmente que sólo faltan los resultados en el momento del abandono, pero todas las covariables se observan por completo. Sin embargo, las covariables que varían en

el tiempo son comunes en los estudios longitudinales. Estas covariables, junto con la variable de resultado, generalmente no se observan en el momento del abandono. Por lo tanto, la suposición de covariables completamente observadas a menudo no es realista en presencia de covariables que varían en el tiempo (4).

En la base de datos utilizada en este TFM, existen mediciones repetidas cada 12 semanas hasta alcanzar la semana 48. Los valores obtenidos de estas mediciones son los de diferentes variables hematológicas, siendo las más importantes en este caso la carga viral y el contaje de linfocitos CD4.

3.2 Tipos de datos faltantes

Según la clasificación de Little y Rubin (6), existen tres categorías para clasificar cómo se generan los datos faltantes: datos faltantes completamente al azar (MCAR según el acrónimo inglés), datos faltantes al azar (MAR según el acrónimo inglés), datos faltantes no al azar (MNAR según su acrónimo en inglés).

Se utilizan los acrónimos en inglés ya que son la denominación estándar en cualquier idioma.

3.2.2 Datos faltantes completamente al azar

La definición de MCAR es que la probabilidad de que falten datos no está relacionada con ninguna variable observada o no observada. Es decir, la probabilidad de que falten datos es la misma para las personas en diferentes grupos de tratamiento y para aquellos que tienen una gravedad de la enfermedad o una respuesta al tratamiento diferentes al resto. Por ejemplo, un tubo de ensayo que se cae en un laboratorio o un fallo en el equipo de medición pueden provocar la pérdida de datos. Como es igualmente probable que esto ocurra en cualquier sujeto del estudio (es decir, independientemente del tratamiento recibido, la gravedad de la enfermedad, etc.), representa un proceso completamente aleatorio. Posteriormente, el efecto promedio del tratamiento será el mismo en aquellos con y sin datos faltantes (5).

Table 2. MCAR Example: Results from a hypothetical clinical trial evaluating the effect of treatment on improvement in body fat (Outcome).

a. Results from the whole study population.

	Outcome			
Treatment	**Y**	**N**		
				• Risk Ratio=(30/40)/(20/40)=1.5
A	30	10	40	• SE(Ln(RR))=0.18
B	20	20	40	• 95% CI =1.05,2.15
	50	30	80	• Chi-Square=5.33 P=0.02

b. Results from completers following 30% missing data from a Missing Completely At Random Mechanism (MCAR)*.

	Outcome			
Treatment	**Y**	**N**		
				• Risk Ratio=(21/28)/(14/28)=1.5
A	21	7	28	• SE(Ln(RR))=0.22
B	14	14	28	• 95% CI =0.98,2.30
	35	21	56	• Chi-Square=3.73 P=0.053

*Each cell from Table 2a is multiplied by 70% to obtain 30% missing data from an MCAR mechanism in Table 2b.

Figura 17: Ensayo con datos MCAR

Para ilustrar el mecanismo MCAR supóngase que durante un estudio que mide la mejora en la la cantidad de grasa corporal, la escala de bioimpedancia que se utilizó para medir la grasa corporal, no era fiable y producía errores de medida en el 30% de los sujetos participantes. También se asume que esta tasa de error no estaba relacionada con el tratamiento de los sujetos, o con si esos sujetos estaban mejorando su grasa corporal o no (5).

A modo de ilustración del mecanismo MCAR, la Figura 17 (Tabla 2a) presenta los resultados de un ensayo hipotético de pérdida de peso que compara el efecto de dos tratamientos (A y B) sobre el resultado, la mejora de la grasa corporal, que se clasifica como sí o no. Una medida común del efecto relativo de dos tratamientos es la razón de riesgo (RR; que es la proporción de aquellos en el tratamiento A con una mejoría, dividida por la proporción en el tratamiento B con una mejoría). En la población de estudio, la mejora en la grasa corporal fue 1,5 veces más probable para el tratamiento A en comparación con el B (es decir, RR = 1,5). El intervalo de confianza del 95 por ciento (1.05, 2.15) y el p-valor (0.02) generados a partir de la prueba de chi-cuadrado muestran una diferencia estadísticamente significativa que indica que el tratamiento A es superior a B (5).

Para ilustrar el mecanismo MCAR, supóngase que durante el estudio la escala de bioimpedencia utilizada para medir la grasa corporal no era fiable y falló en el 30 por ciento de los sujetos participantes . También asumimos que esta tasa de error no estaba relacionada con su tratamiento o si tuvieron o no una mejora en la grasa corporal (es decir, un mecanismo MCAR). La Figura 17 (Tabla 2b) muestra los resultados del estudio en aquellos sin datos faltantes (es decir, aquellos que completaron el estudio). La razón de riesgo en los que completaron se mantiene sin cambios en 1.5. Por lo tanto, la estimación del efecto del tratamiento no estuvo sesgada por los datos faltantes de MCAR. Sin embargo, debido al tamaño reducido de la muestra debido a la falta de datos, el error estándar es mayor, lo que resulta en un intervalo de confianza más amplio (0.98, 2.30) que incluye el valor nulo de 1.0 y un estadístico de chi-cuadrado que ya no es significativo en el Umbral de 0,05 (p = 0,053). Este ejemplo ilustra que el análisis de casos completo no da como resultado una estimación sesgada de la diferencia de tratamiento en ausencia de datos que surjan de un mecanismo MCAR; sin embargo, hay una pérdida de precisión en la estimación de la diferencia de tratamiento (es decir, errores estándar más grandes y límites de confianza más amplios), así como una pérdida de potencia en la prueba de significancia (es decir, estadísticas de prueba más conservadoras que producen p valores más elevados) (5).

3.2.3 Datos faltantes al azar

Cuando la probabilidad de que falten datos está relacionada con las variables observadas pero no con las variables no observadas, el mecanismo es de datos faltantes al azar (MAR). Si en un ensayo clínico el abandono es más probable para los hombres que para las mujeres, pero todos los hombres tienen la misma probabilidad de abandono y todas las mujeres tienen la misma probabilidad de abandono; el mecanismo de datos faltantes es MAR. Otros ejemplos del mecanismo MAR son datos faltantes causados por características del diseño del estudio (por ejemplo, proporcionar terapia de rescate cuando las condiciones no están suficientemente controladas de acuerdo con los criterios del protocolo), abandonos basados en efectos secundarios registrados o falta de eficacia, o abandonos basados en características básicas conocidas (5).

Table 3. MAR example: Results from a hypothetical clinical trial evaluating the effect of treatment on improvement in body fat (Outcome).

a. Results from the whole study population.

Men	Outcome			
Treatment	**Y**	**N**		
A	225	75	300	• Outcome in A = 225/300 = 0.75
B	150	150	300	• Outcome in B = 150/300 = 0.50
	375	225	600	• Risk Ratio = 1.5

Women	Outcome			
Treatment	**Y**	**N**		
A	90	210	300	• Outcome in A = 90/300 = 0.30
B	60	240	300	• Outcome in B = 60/300 = 0.20
	150	450	600	• Risk Ratio = 1.5

Total	Outcome			
Treatment	**Y**	**N**		• Outcome in A = 315/600 = 0.525
A	315	285	600	• Outcome in B = 210/600 = 0.35
B	210	390	600	• Risk Ratio =1.5
	525	675	1200	• 95%CI=1.31, 1.71
				• Mantel Haenszel RR=1.5
				• 95% CI=1.33, 1.70

b. Results from completers following missing data from a Missing At Random Mechanism (MAR)*.

Men	Outcome			
Treatment	**Y**	**N**		
A	180	60	240	• Outcome in A = 180/240 = 0.75
B	120	120	240	• Outcome in B = 120/240 = 0.50
	300	180	480	• Risk Ratio = 1.5

Women	Outcome			
Treatment	**Y**	**N**		
A	81	189	270	• Outcome in A = 81/270 = 0.30
B	27	108	135	• Outcome in B = 27/135 = 0.20
	108	360	405	• Risk Ratio =1.5

Total	Outcome			
Treatment	**Y**	**N**		• Outcome in A = 261/510 = 0.51
A	261	249	510	• Outcome in B = 147/375 = 0.39
B	147	228	375	• Risk Ratio = 1.31
	408	477	885	• 95% CI=1.12, 1.52
				• Mantel Haenszel RR=1.5
				• 95% CI=1.30, 1.73

*Probabilities of missing are dependent on the combination of treatment and gender to mimic an MAR mechanism
Probability Missing for Men in Trt A = 0.20
Probability Missing for Men in Trt B = 0.20
Probability Missing for Women in Trt A = 0.10
Probability Missing for Women in Trt B = 0.55

Figura 18: Ensayo con datos MAR

Para ilustrar MAR, los resultados de otro hipotético ensayo de pérdida de peso se muestran en la Figura 18 (tabla 3a). Aunque la probabilidad general de mejora fue mayor en los hombres (62,5 por ciento = 375/600) que en las mujeres (25 por ciento = 150/600), los cocientes de riesgo fueron idénticos (RR = 1,5). La combinación de los datos específicos de género también da como resultado la misma razón de riesgo bruta y ajustada por género (es decir, RR de Mantel Haenszel = 1.5, IC del 95 por ciento = 1.33, 1.70). Supóngase que la probabilidad de que falten datos depende de la combinación del tratamiento y el género (es decir, un mecanismo MAR). Por ejemplo, al final del estudio, el 20 por ciento de los hombres en el tratamiento A, el 20 por ciento de los hombres en el tratamiento B, el 10 por ciento de las mujeres en el tratamiento A y el 55 por ciento de las mujeres en el tratamiento B tenían datos faltantes. Téngase en cuenta que las tasas de datos faltantes dependen de los datos observados pero no del resultado no observado. La Figura 18 (Tabla 3b) demuestra las tabulaciones cruzadas con las tasas de datos faltantes aplicadas a toda la población del estudio. A pesar de los diferentes porcentajes de datos faltantes, las proporciones de riesgo específicas de género siguen siendo las mismas, 1,5. Sin embargo, la combinación cruda conteniendo sólo los datos de los finalizadores de cada género da como resultado una razón de riesgo de 1,31 (IC del 95 por ciento = 1,12, 1,52) que es menor que la razón de riesgo real de 1,5. Después del ajuste por género mediante el método de Mantel-Haenszel, la razón de riesgo real de 1,5 se recupera en los que completaron el estudio, aunque con precisión reducida y límites de confianza ligeramente mayores (1,30, 1,73) en comparación con toda la población del estudio (5).

3.2.4 Datos faltantes no al azar

Cuando la probabilidad de que falten datos depende de los datos no observados, los datos faltantes se denominan perdidos no al azar (MNAR). Por ejemplo, en los ensayos por abuso de sustancias con la abstinencia como resultado, es habitual que el abandono sea mayor para los que han recaído. El problema es que aquellos que abandonan los ensayos no suelen obtener el estado de recaída. En este caso, la probabilidad de que falten datos depende de los datos no observados - estado de recaída. En otro ejemplo, considérese un estudio que evalúa tratamientos para reducir el consumo de cocaína en el que el resultado es el nivel de droga de una prueba de drogas en orina medida cada lunes por la mañana. Se espera que los participantes que consumen cocaína durante el fin de semana y no se presenten a la prueba de orina tengan niveles más altos de metabolitos de cocaína. Por lo tanto, la probabilidad de que falten los datos está directamente relacionada con el nivel de cocaína no observado (5).

Table 4. MNAR example: Results from a hypothetical clinical trial evaluating the effect of treatment on improvement in body fat.

a. Results from the whole study population.

	Outcome			
Treatment	**Y**	**N**		
				• Risk Ratio=(315/600)/(210/600)=1.5
A	315	285	600	• 95% CI =1.31, 1.71
B	210	390	600	• Chi-Square=37.3 P<0.001
	525	675	1200	

b. Results from completers following missing data from a Missing Not At Random Mechanism*.

	Outcome			
Treatment	**Y**	**N**		
				• Risk Ratio=(189/474)/(210/444)=0.84
A	189	285	474	• 95% CI =0.73, 0.98
B	210	234	444	• Chi-Square=5.14 P=0.02
	399	519	918	

*Probabilities of missing are dependent on the treatment and outcome to mimic an MNAR mechanism
Probability Missing for Outcome "Y" in Trt A = 0.40
Probability Missing for Outcome "N" in Trt A = 0.00
Probability Missing for Outcome "Y" in Trt B = 0.00
Probability Missing for Outcome "N" in Trt B = 0.40

Figura 19: Ensayo con datos MNAR

Continuando con el ensayo clínico hipotético que evalúa el efecto del tratamiento en la mejora de la grasa corporal, la Figura 19 (Tabla 4) demuestra las consecuencias de los datos faltantes que surgen de un mecanismo MNAR. La proporción de datos faltantes se estableció en 40 por ciento en aquellos que recibieron el tratamiento A y tuvieron una mejora en la grasa corporal y en 40 por ciento en aquellos que recibieron el tratamiento B y *no tuvieron* ninguna mejora. Por lo tanto, los datos faltantes se relacionaron con el resultado no observado (es decir, MNAR). El resultado neto fue una inversión completa de la razón de riesgo cuando se examinó toda la población del estudio (Figura 19 Tabla 4a, RR = 1,5) en comparación con los que completaron (Figura 19 Tabla 4b, RR = 0,84), y ambos alcanzaron significación estadística. Si se hubiera utilizado el análisis de casos completos, las conclusiones habrían sido opuestas al efecto real (5).

Una taxonomía alternativa se refiere a los datos faltantes que surgen de un mecanismo MNAR como no ignorables porque ignorar el proceso que conduce a los datos faltantes conducirá a resultados sesgados. Por el contrario, la probabilidad de que falten datos por falta de información ignorable (MCAR o MAR) en una variable en

particular no depende de los valores de esa variable dadas otras variables observadas. Estos datos aún pueden producir estimaciones no sesgadas sin la necesidad de un modelo que explique el mecanismo que falta. Para obtener diferencias de tratamiento no sesgadas cuando los datos faltantes son MNAR, es necesario modelar la relación entre el resultado de interés y la probabilidad de no respuesta. Determinar esta relación es una tarea difícil que resalta la importancia de obtener datos de resultado en cada paciente aleatorizado y recopilar datos auxiliares que pueden predecir el abandono (5).

3.3 Técnicas para tratar los datos faltantes

Debido a que es un problema con el que muchos estadísticos luchan a diario, a lo largo de los años se han desarrollado diferentes técnicas para lidiar con este tipo de datos. En la Figura 20 se puede observar para MCAR y MAR, cuáles de esas técnicas producen efectos insesgados y estimaciones correctas de los errores estándar y los p-valores. Se puede observar también cuáles producen únicamente efectos insesgados, y cuáles son directamente inaceptables. En el caso de MNAR se puede observar la única manera aceptable de manejar este tipo de datos.

MCAR		
Efectos insesgados y errores estándar	**Efectos insesgados**	**Inaceptable**
* Basados en la probabilidad * Imputación múltiple * Ponderación de la probabilidad inversa * Casos completos	* Imputación por media simple * Imputación por media condicional	* Última observación llevada adelante * La peor observación llevada adelante

MAR		
Efectos insesgados y errores estándar	**Efectos insesgados**	**Inaceptable**
* Basados en la probabilidad * Imputación múltiple * Ponderación de la probabilidad inversa	* Imputación por media condicional	* Última observación llevada adelante * La peor observación llevada adelante * Imputación por media simple * Casos completos

MNAR
Aceptable
Modelado conjunto del resultado, así como de la relación entre el resultado y la probabilidad de respuesta (por ejemplo, modelos de selección o mezcla de patrones)

Figura 20: Métodos para tratar datos faltantes

Como se puede observar, incluso con MCAR hay metodologías que no pueden ser utilizadas debido a que producen sesgos en los datos resultantes.

De todos modos, la existencia de estas técnicas de tratamiento de datos faltantes choca con el inconveniente de que no hay un análisis que indique cuál es el mecanismo de pérdida de datos que ha llevado a los datos faltantes observados al final de un estudio (5).

Por lo general, se asume que los datos faltantes son MAR, y aplicando el principio de intención de tratar (ITT), se procede con la imputación de los mismos.

Sin embargo, al faltar datos, no hay un análisis ITT inequívoco. Como tales, los métodos recomendados en condiciones MAR, a menudo proporcionan un enfoque sensato para el análisis primario. Pero se recomiendan los análisis de sensibilidad para comprender la solidez de las desviaciones del supuesto MAR (5).

3.3.1 Análisis de casos completos

Consiste en utilizar para el análisis sólo los datos de los sujetos que tienen el juego completo de datos observados. Así que los indivíduos con algún dato faltante se excluyen del análisis de los datos (5).

Su principal ventaja es su simplicidad, tanto estadística como computacional.

Tiene varias desventajas (5):

- Pérdida de potencia estadística y de precisión en la estimación de los efectos del tratamiento.
- Si los datos perdidos no lo son por MCAR, la estimación del efecto de la intervención va a estar sesgada.

3.3.2 Imputación simple

Utilizando una regla para asignar todos los datos perdidos a un valor, se crea un juego completo de datos para todos los sujetos randomizados. Existen varios tipos: última observación llevada adelante (LOCF), peor observación llevada adelante, y las imputaciones a la media (tanto la simple como la condicional) (5)

LOCF ha sido un método muy popular hasta hace poco, en parte porque se supone que produce una estimación conservadora de los valores. Sin embargo un pueden imaginarse fácilmente escenarios en los que introduce sesgos notorios. Supóngase un ensayo clínico en el que se mide el deterioro cognitivo de pacientes con enfermedad de Alzheimer. Si un paciente abandona el ensayo y sus datos se imputan mediante LOCF, se estaría produciendo un serio sesgo, que se podría incrementar si hay un grupo del ensayo en el que hay más abandonos que en el resto (5).

Además, imputar valores idénticos para un mismo individuo puede llevar a una infravaloración de la variabilidad, y a un p-valor reducido (5).

Con respecto a la imputación simple a la media.Ésta ignora la información de otras variables que puede ser relavante si los datos son MAR. En el caso de los datos MCAR puede conducir a estimaciones acertadas con estos datos; pero sin duda imputar

utilizando una constante provoca una infraestimación de la variabilidad subyacente en los datos faltantes. Sin embargo la imputación condicional a la media, acomoda las asociaciones con otras variables observadas al obtener el resultado mediante regresión sobre otras variables observadas en los participantes que completaron el ensayo. Ésto mejora la validez de las estimaciones, incluso bajo condiciones MAR. Aunque como otros procedimientos de imputación simple, la utilización de un valor simple para reemplazar los datos faltantes no captura completamente la incertidumbre de si ese valor es correcto. Y su acción conlleva la infraestimación de la variabilidad del tratamiento y p-valores inapropiadamente bajos (5).

3.3.3 Ponderación de la probabilidad inversa

Este método tiene su origen en la investigación de encuestas por muestreo en las que las respuestas de los participantes de la encuesta se ponderan para adaptarse a las probabilidades desiguales de selección. El peso de la encuesta para cada participante es el inverso de su probabilidad de selección, por lo que los que tienen menores probabilidades de selección tienen un mayor peso en el análisis. En el caso de los datos faltantes, se estiman las probabilidades de ser observado y los análisis se ponderan por su inverso. En consecuencia, aquellos con una baja probabilidad de observación tendrán una mayor ponderación en el análisis. Las ponderaciones se pueden obtener de un modelo, como por ejemplo una regresión logística que incluya el grupo de intervención, los valores previos del resultado de interés y otras covariables que pueden ser predictivas del hecho de ser observado (5). esta técnica proporciona estimaciones no sesgadas de la diferencia entre tratamientos bajo condiciones MAR. Sin embargo, no utiliza directamente todos los datos disponibles, ya que solo los sujetos con datos completos se incluyen en el modelo ponderado final. Por tanto, la potencia puede verse atenuada (5).

3.3.4 Imputación múltiple

El método de imputación múltiple genera m conjuntos de datos completos (típicamente m está en el rango de 5 a 20). Cada uno de estos conjuntos contiene los valores observados y diferentes valores plausibles imputados para las observaciones faltantes. Después de crear los m conjuntos de datos completos, cada uno se analiza utilizando los métodos habituales (por ejemplo, ANCOVA, regresión), y los resultados se combinan entre los análisis. Es importante señalar que las imputaciones no pretenden ser observaciones reales para ese individuo, sino más bien un conjunto de valores estadísticamente plausibles basados en otra información para ese individuo. Por lo tanto, el análisis de conjuntos de datos *"rellenados"* proporciona resultados estadísticamente plausibles que se habrían obtenido si no hubiera datos faltantes (5).

El procedimiento de imputación múltiple se subdivide en 3 pasos: imputación, análisis, y agrupación (7)

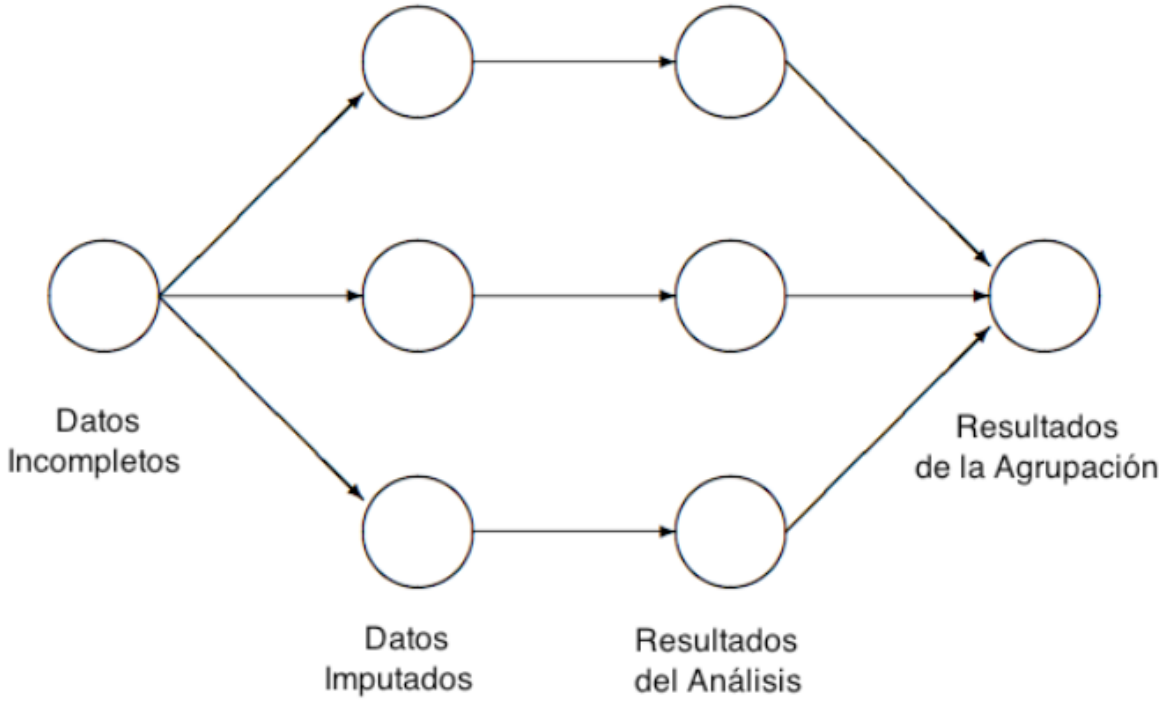

Figura 21: Pasos de la imputación múltiple

En condiciones de datos perdidos bajo MAR, la imputación múltiple produce estimaciones insesgadas del efecto de la intervención y p-valores correctos. Otras ventajas de la imputación múltiple incluyen la capacidad de manejar no solo datos faltantes sino también la información de covariables faltantes. Su implementación es relativamente fácil, y proporciona flexibilidad al separar la imputación del modelo analítico. Esto último permite una mayor complejidad del modelo de imputación para hacer más plausible el supuesto MAR. También proporciona un marco simple y atractivo para explorar la sensibilidad a los datos faltantes no aleatorios (5).

Los inconvenientes de la imputación múltiple incluyen la incapacidad de producir una estimación única del efecto del tratamiento (proporciona un resultado diferente cada vez que se usa); y el requisito de compatibilidad entre los modelos de imputación y análisis (por ejemplo, el modelo de análisis no puede contener variables, no linealidades o interacciones que no están en el modelo de imputación). El modelo de imputación puede ser más complejo que el modelo de análisis, pero este último no puede contener variables que no estén en el modelo de imputación (5).

3.3.5 Análisis basados en la probabilidad

La estimación de máxima verosimilitud (MLE) es un método de estimación común en estadísticas que depende de encontrar estimaciones de las diferencias de tratamiento que maximicen la probabilidad de los datos observados. Para ilustrar el enfoque de la MLE, supóngase que se hace un experimento en N personas donde la probabilidad de éxito para un individuo es p y la probabilidad de fracaso es 1-p. Si n personas tienen éxito y N-n personas fracasan, la probabilidad es proporcional al producto de las probabilidades de éxitos y fracasos o p_n $(1-p)_{N-n}$. El valor de p que maximiza la probabilidad es n/N (o proporción general de éxito). En la prueba de pérdida de peso sin datos faltantes, la máxima probabilidad producirá la mejor estimación de la diferencia en la grasa corporal entre los grupos de intervención que maximiza la probabilidad de observar los datos. El problema cuando aparecen datos faltantes es que sólo se puede observar un subconjunto de los datos, pero el objetivo es sacar conclusiones basadas en los datos completos. Bajo MAR, el análisis basado en la probabilidad permite lograr esto al promediar los datos faltantes de la probabilidad conjunta de los datos observados y faltantes. Esto es posible bajo MAR, porque el comportamiento estadístico futuro de un sujeto, condicionado a los datos observados, es el mismo si ese sujeto abandona o no (5).

Bajo el supuesto de MAR, MLE produce estimaciones insesgadas del efecto de la intervención y valores p correctos. A diferencia de la imputación múltiple, MLE proporciona una estimación única de la diferencia de tratamiento. MLE también requiere menos decisiones que MI y no depende de la compatibilidad de la imputación (ya que no hay imputación en MLE) y modelo de análisis. Las desventajas de MLE incluyen su dependencia de supuestos paramétricos (por ejemplo, normalidad) y que sólo es apropiado para datos de resultados faltantes (es decir, no puede acomodar datos de covariables faltantes) (5).

3.3.6 Aproximaciones aceptables bajo condiciones MNAR

El enfoque básico para manejar los datos faltantes ignorables (es decir, de un mecanismo MCAR o MAR) es ajustar todas las diferencias observables entre los casos perdidos y no perdidos y asumir que todas las diferencias restantes son asistemáticas, ignorando así el proceso por el cual los datos faltantes ocurren. Cuando se producen datos faltantes en un proceso MNAR, el análisis apropiado requiere el modelado conjunto del resultado junto con el mecanismo de datos faltantes. Esto puede ser muy complicado, dado que bajo MNAR el modelo, y por lo tanto el proceso de datos faltantes rara vez se conoce ; creando suposiciones no verificables para el análisis. Por ejemplo, se sospecha que la recaída da lugar a que falten datos en un estudio de tratamiento por abuso de sustancias, pero es poco probable que todos los datos

faltantes sean el resultado de una recaída. Para realizar un análisis MNAR, es necesario especificar la fuerza de esta relación, es decir, ¿cuál es la probabilidad de tener datos faltantes dada la recaída o de manera similar (pero no igual) cuál es la probabilidad de recaída en aquellos con datos faltantes? (5)

Paralelamente a estas preguntas relacionadas, hay dos enfoques de análisis bajo modelos MNAR: selección, y combinación de patrones. Los modelos de selección requieren la especificación de la relación entre el resultado y la probabilidad de que falte un dato. Por ejemplo, en un estudio de pérdida de peso, la probabilidad de que no se pierda una observación podría ser menor en aquellos que tienen un aumento reciente no observado de grasa corporal. Por otro lado, los modelos de mezcla de patrones especifican la distribución de resultados a través de los patrones de datos faltantes observados. Para el ejemplo de pérdida de peso, esto podría corresponder a indicar la probabilidad de varios perfiles de pérdida de peso para aquellos que abandonan después de su primera visita en comparación con aquellos que abandonan después de su segunda visita o aquellos que nunca abandonan. Si bien son arbitrarios, los perfiles de pérdida de peso elegidos podrían estar más informados por los datos registrados, como el motivo del abandono. Donde se podrían adoptar diferentes perfiles para los que abandonan debido a la migración en comparación con la falta de eficacia (5).

3.4 Patrones de datos faltantes

La librería MICE dispone de una función llamada *md.pattern* que realiza una gráfica en la que colorea cada una de las celdas de la base de datos según si contienen datos faltantes o no. Si la celda tiene datos, su color se muestra en azul, mientras que si en esa no tiene datos, el color es rojo.

Por razones teóricas y prácticas, se pueden distinguir varios patrones diferentes de datos faltantes (8), Figura 22:

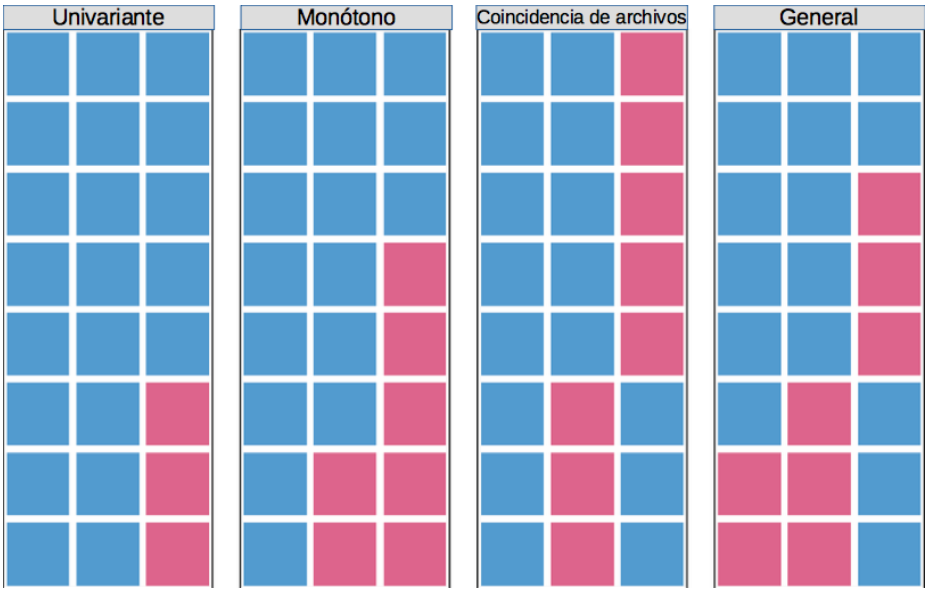

Figura 22: Patrones de datos faltantes

- 1 - Univariante y multivariante: Un patrón de datos faltantes es univariante si sólo existe una variable con datos faltantes.

- 2 - Monótono y no monónoto (o general): Si las variables $Y_j$pueden ordenarse de tal manera que si Y_j falta, todas las variables Y_k con k>j también faltan. Esto sucede, por ejemplo, en estudios longitudinales con abandono. Si el patrón es no monótono, se le llama no monótono o general.

- 3 - Conectado o no conectado: Se denominan así los patrones de datos faltantes en los que cualquier punto con datos observados, puede ser alcanzado desde cualquier otro punto de datos observados a través de una secuencia de movimientos horizontales o verticales (como los de la torre en el ajedrez). El patrón de coincidencia de archivos es un tipo específico de patrón no conectado en el que dos o más columnas no tienen ningún dato observado que coincida entre ellas. Eso hace difícil su mutua utilidad en la predicción de los valores a imputar.

4 Los antirretrovirales

4.1 Los antirretrovirales en la lucha contra el VIH

A pesar de que la replicación del virus de la inmunodeficiencia humana (VIH) puede ser suprimida con los tratamientos actualmente disponibles, la erradicación de la infección por el VIH es todavía un objetivo inalcanzable. Por ello, el tratamiento antirretroviral debe ser establecido de por vida en la mayoría de los infectados por el VIH (9).

La eficacia del tratamiento antirretroviral de gran actividad (TARGA) ha sido demostrada en varios ensayos clínicos. Aun así, una importante proporción de pacientes no consigue mantener una correcta supresión viral en la práctica clínica diaria (9).

La adherencia al tratamiento TARGA es crítica para obtener una supresión viral duradera. Por ello, factores que se relacionan con la adherencia como el elevado número de pastillas o de tomas, la complejidad del régimen antirretroviral, su tolerabilidad y las restricciones alimentarias pueden tener un efecto sobre la replicación viral (9)

5. El procedimiento seguido en este TFM

5.1 Gestión de los datos

5.1.1 El dataset “Lake”

Los datos que voy a utilizar para este TFM se encuentran almacenados en un archivo tipo CSV. Lo forman 116 filas con los datos de los indivíduos integrados en el ensayo, y cada una de las 219 columnas es una característica medida durante el ensayo.

Los datos son los extraídos del siguiente ensayo clínico de fase III: “Ensayo clínico multicéntrico, abierto, prospectivo. aleatorizado para evaluar la efectividad de abacavir 600 mg + lamivudina 300 mg en pauta QD + efarivenz 600 mg QD versus kaletra 400/100 g BID como tratamiento antirretroviral de inicio”. Este ensayo clínico fue promovido por la Fundació de Lluita Contra la SIDA. El objetivo principal del ensayo fue evaluar la equivalencia terapéutica entre las dos ramas de tratamiento en la respuesta virológica durante las 48 semanas de duración del estudio. Y evaluar el porcentaje de fracasos virológicos (9).

Como su nombre indica se trata de un ensayo aleatorizado, prospectivo, multicéntrico y abierto. En él participaron 126 pacientes con infección por VIH sin experiencia antirretroviral previa (63 por grupo de tratamiento)(9). Siendo los grupos los siguientes:

- Rama A: Kivexa (un comprimido al día), y Efarivenz 600 mg (una cápsula al día)
- Rama B: Kivexa (un comprimido al día), y Kaletra (tres cápsulas cada doce horas)

La duración aproximada del ensayo fueron 72 semanas (24 semanas de inclusión y 48 semanas de seguimiento de cada paciente) (9).

5.1.2 Carga de datos y preprocesamiento

La carga inicial de los datos se realiza mediante la función *read.csv2*, ya que el separador de las columnas en la base de datos es el punto y coma. Otra opción para realizar esta tarea sería utilizar la función *read.csv*, pero indicando que el separador es punto y coma.

Posteriormente se suma cuántos NA hay, con el fin de obtener el número global de datos faltantes. Siendo el resultado de 10423 celdas con datos faltantes, sobre un total de 25404 celdas; lo que da un 41,03% de datos faltantes. Y por si acaso hay algún tipo de carácter especial que R no considera datos faltantes, se realiza también un contaje de NAs con caracteres especiales. El número obtenido es el mismo que de NAs, por lo que no existen celdas que contengan caracteres especiales.

Un 41% de datos faltantes es muy elevado para que incluso el procedimiento de imputación de los datos funcione correctamente. Por lo que se hizo posteriormente una selección de las variables según su porcentaje de datos faltantes.

Es importante que cada columna se encuentre en el formato correcto según los datos que contiene. Como en la base de datos hay fechas, se convierten éstas a formato fecha de R. De la misma manera, todas las columnas que codifican factores se convierten a clase factor.

Tras una inspección visual de la base de datos, se observó que había fechas incongruentes con las reglas estipuladas en el protocolo del ensayo clínico cuya base de datos utilizamos. Esas fechas incongruentes son: fecha de nacimiento que dan edades de pacientes menores de 18 años, fecha de infección por VIH posterior al inicio del ensayo, y fecha de inclusión en el ensayo y/o fecha de la primera toma de muestras con valor posterior al de la finalización del ensayo. Se eliminó estas fechas codificando un NA en su lugar, para que se pudiesen imputar si se considerase necesario.

Se recodificaron las variables dicotómicas con valores diferentes de 0 y 1 a valores 0 y 1.

Se calcula el porcentaje de datos faltantes de cada variable, y con estos datos se procede a eliminar aquellas variables con más de un 50% de datos faltantes (salvo aquellas que contienen datos de carga viral o linfocitos CD4A). Además se eliminan aquellas que no contienen datos útiles para el Trabajo, y las fechas.

De las 10423 celdas iniciales con datos faltantes, tras esta eliminación de variables se pasa a 2796 celdas con datos faltantes. Y en porcentaje esto supone un 25,64% de celdas sin datos, lo que ya es un porcentaje más manejable.

Para tantear la posibilidad de utilizar la metodología de *complete cases*, que consiste en trabajar sólo con aquellas filas que tienen datos en todas las variables; se calcula cuántas filas de la base de datos actual tiene todas sus varibles completas. El número obtenido es 5, por lo que sólo 5 filas tienen sus datos completos. Ese es un número insuficiente a todas luces. Se necesita realizar una imputación de los datos para así entrenar y testear los diferentes algoritmos de aprendizaje automático.

Una buena manera de poder observar y clasificar la estructura de la pérdida de datos es mediante un gráfico de patrón de datos perdidos. Se obtiene ese gráfico pra todas las filas y columnas pertenecientes a la base de datos actual.

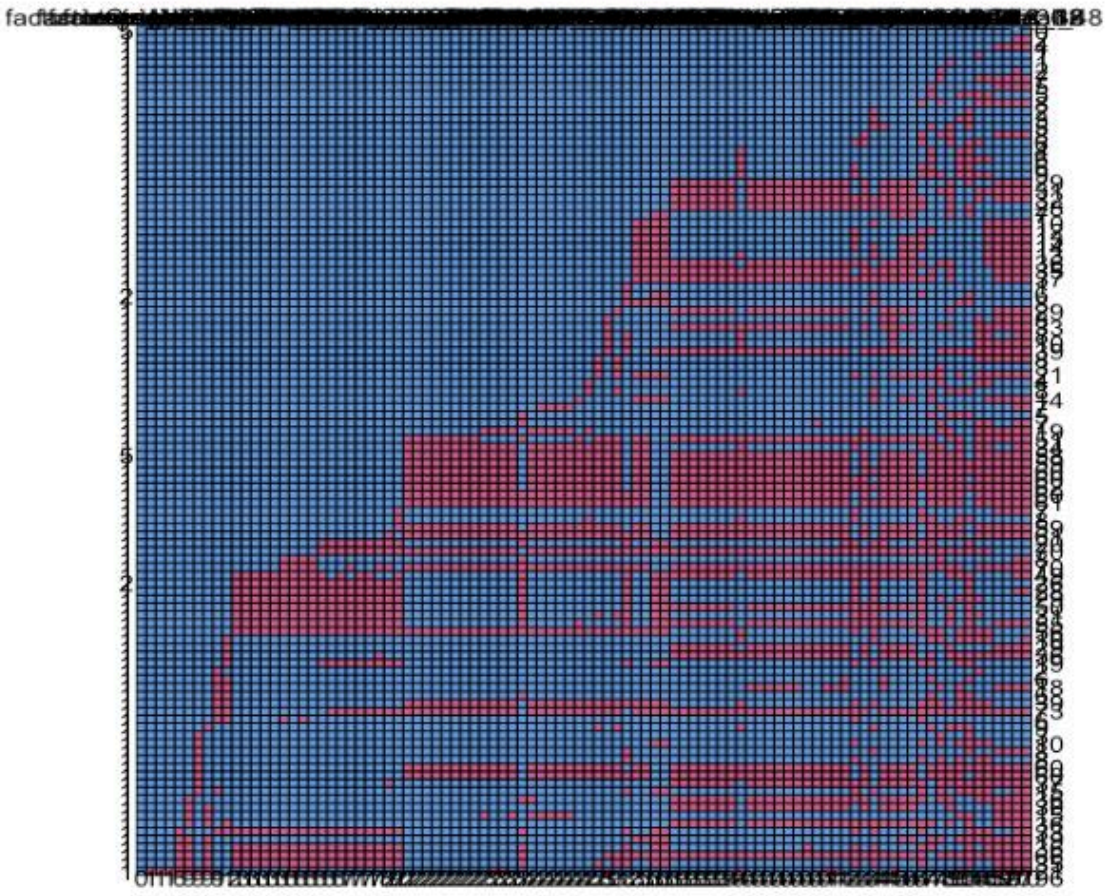

Figura 23: Patrón gráfico de datos perdidos

Del gráfico se infiere que la pérdida de datos es multivariable, ya que hay datos faltantes en varias variables del dataset. Y también es no monótona, ya que las celdas rojas que marcan los datos individuales perdidos no están conectadas desde su aparición hasta el final del ensayo.

Sobre esta misma base de datos se calculan varias columnas que mediante variables dicotómicas codifican la existencia o no de fracaso virológico en las semanas 24,36 o 48 del estudio. Combinando esas tres columnas se obtiene la variable que codifica si se ha observado el evento del estudio (fracaso virológico = valor 1) o no se ha observado el evento del estudio (censura = valor 0).

Se codifica como fracaso virológico:

- Semana 24: Que no se haya alcanzado la indetectabilidad en esa semana (carga viral menor o igual a 50), o que se ha alcanzado la indetectabilidad en la semana 12, pero en la 24 la carga viral a subido a más de 200

- Semana 36: Que habiendo tenido indetectabilidad en la semana 24, en la semana 36 la carga viral sea mayor de 200.
- Semana 48: Que habiendo tenido indetectabilidad en la semana 24, en la semana 36 la carga viral sea mayor de 200. O bien que el contaje de linfocitos CD4A en esta semana sea inferior a 300.

Acto seguido se crea otra nueva columna que codifica el valor del tiempo (en semanas) de la última medición obtenida del paciente. Ésta variable indica el tiempo hasta el abandono, sea este por fracaso virológico(semanas 24,36,48), por finalización del ensayo (semana 48), o por cualquier otra causa (semanas 4,12,24,36,48). Los tiempos de abandono se obtuvieron manualmente viendo columna por columna cuál es la última en la que había datos.

La última columna en ser creada es la que mediante una variable dicotómica indica la existencia de abandono (hay datos faltantes) en el paciente en cuestión. No se diferencia entre causas de abandono.

Ahora se realizará un procedimiento específico para datos que van a ser imputados. Calcular el influx y el outflux de todas las columnas.

Ambos estadísticos toman valores contínuos que van del 0 al 1. El influx de una variable cuantifica cómo de bien conectan sus datos faltantes a los datos observados en otras variables, por lo que una variable con un elevado influx será más fácil de imputar. Mientras que el outflux de una variable cuantifica cómo de bien conectan sus datos observados a los datos faltantes en otras variables, por lo que las variables con buen outflux serán buenas variables predictoras.

	pobs <dbl>	influx <dbl>	outflux <dbl>	ainb <dbl>	aout <dbl>	fico <dbl>
CD4A_48	0.3965517	0.5248954	0.1948870	0.6336237	0.1394486	0.8913043
CargaViral_48	0.3965517	0.5259056	0.1974805	0.6348432	0.1413043	0.8913043
CargaViral_36	0.4913793	0.4206956	0.2656539	0.6025217	0.1534018	0.9122807
CD4A_36	0.4913793	0.4206956	0.2656539	0.6025217	0.1534018	0.9122807
Sodio_24	0.6034483	0.2675711	0.2723231	0.4915164	0.1280488	0.9285714
Potasio_24	0.6034483	0.2675711	0.2723231	0.4915164	0.1280488	0.9285714
LDL_mg_24	0.5775862	0.3040843	0.2738051	0.5243902	0.1345104	0.9253731
ProteinasTotales_24	0.6120690	0.2599221	0.2834383	0.4880759	0.1313981	0.9295775
HDL_mg_24	0.6293103	0.2368307	0.2856614	0.4653999	0.1288005	0.9315068
CargaViral_24	0.6206897	0.2499639	0.2886254	0.4800443	0.1319444	0.9305556

1–10 of 40 rows Previous 1 2 3 4 Next

Figura 24: Variables con menor outflux

Los valores de outflux de la mayoría de las variables de la base de datos utilizada se sitúan entre 0,5 y 0,8. Estos valores se consideran bajos, siendo los problemas de datos faltantes my severos en ellos, pero potencialmente este grupo puede contener variables importantes.

	pobs <dbl>	influx <dbl>	outflux <dbl>	ainb <dbl>	aout <dbl>	fico <dbl>
CD4P_0	0.8620690	0.10737480	0.7836236	0.5670732	0.2579268	0.9500000
CD8P_0	0.8620690	0.10737480	0.7836236	0.5670732	0.2579268	0.9500000
CD4A_0	0.8620690	0.10867369	0.7869581	0.5739329	0.2590244	0.9500000
CD8A_0	0.8620690	0.10867369	0.7869581	0.5739329	0.2590244	0.9500000
Hematocrito_0	0.8879310	0.08428345	0.8165987	0.5478424	0.2609519	0.9514563
Hemoglobina_0	0.8879310	0.08428345	0.8165987	0.5478424	0.2609519	0.9514563
Plaquetas_0	0.8879310	0.08428345	0.8165987	0.5478424	0.2609519	0.9514563
Leucocitos_0	0.8879310	0.08428345	0.8165987	0.5478424	0.2609519	0.9514563
LinfosTotales_0	0.8879310	0.08428345	0.8165987	0.5478424	0.2609519	0.9514563
VHC_0	0.9224138	0.08110839	0.9314561	0.7615176	0.2865284	0.9532710

31–40 of 40 rows Previous 1 2 3 4 Next

Figura 25: Variables con mayor outflux

Se eliminan las columnas que tienen un outflux y un influx menor de 0,4; salvo las de carga viral y CD4A.En este caso son:

“Calcio_12”, “pH_12”,“Bicarbonato_12”, “AcidoLactico_12”, CD4P_24“,”CD8A_24“,”CD8P_24“,”Hematocrito_24“,”Hemoglobina_24“,”Plaquetas_24“, ”Leucocitos_24“,”LinfosTotales_24“,”Glucosa_mg_24“,”Urea_mg_24“,”Creatinina_mumol_24“,”Sodio_24“,”Potasio_24“,”Bilirrubina_mumol_24“,”GPT_24“,”GOT_24“,”GGT_24“,” ProteinasTotales_24“, Colesterol_mg_24”, “LDL_mg_24”, “HDL_mg_24”, "Trigliceridos_mg_24

Tras la eliminación de estas variables el porcentaje de datos faltantes baja considerablemente hasta el 20,43%, lo que también mejorará el procedimiento de imputación.

Otro de los pasos dados antes de la imputación es el de buscar outliers. En general se ha buscado la menor manipulación de la variabilidad existente en los datos. Pero hay valores que están muy fuera de lo que se puede tomar incluso como valores muy poco comunes.

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
111	-657.49781	0.0000e+00	0.0000e+00
60	-16.00247	1.2281e-57	3.3159e-56

2 rows

Figura 26: Resultado del test de valores atípicos

Claramente la observación de CD4A_12 de la fila 60 es errónea, ya que en una persona sana el contaje de linfocitos CD4A se encuentra en el rango de 530 - 1570 unidades por microlitro (10). Así que se procede a sustituir su valor por NA, y se rellenará su valor mediante imputación.

En el caso de la fila 111 no se observa ninguna medición tan anómala como para ser eliminada.

Con este procedimiento finaliza la gestión de los datos.

5.2 Imputación múltiple con MICE

También se puede realizar imputación múltiple mediante las librerías *Hmisc*, *Amelia*, *missForest*, y *mi*. Incluso el paquete utilizado para la realización del bosque aleatorio de supervivencia *randomForestSRC* en caso de datos faltantes, realiza una imputación previa a la obtención del modelo de aprendizaje automático.

El primer paso propiamente dicho de la imputación es el de establecer los métodos que se van a utilizar durante el proceso. Para ello se realiza una imputación con 0 iteraciones, que servirá de base para poblarla de los métodos que se quieran aplicar a la imputación definitiva.

Concretamente se utilizaron los siguientes métodos:

- polr - Para las variables categóricas ordenadas. Es un método específico para ellas.
- logreg.boot - Para las variables categóricas dicotómicas. Es un método específico para ellas.
- cart - Para variables que dan problemas cuando se utiliza el método pmm, que es el que por definición utiliza MICE. Se trata de un método robusto frente a los outliers, y que puede manejar variables con elevada multicolinearidad y distribuciones sesgadas.

Posteriormente mediante la función *quickpred* del paquete MICE se definen las variables que van a intervenir en la predicción. Serán principalmente las variables con elevado outflux. Al contrario, también se define que las variables con bajo outflux, y las categóricas, no van a intervenir en la predicción.

Utilizando estos métodos y estos predictores, se realiza una imputación con 25 iteraciones. No se obtienen errores por parte de la funciòn *mice*, lo que indica que no existen problemas de colinearidad en los datos.

Una vez finalizada la imputación de la base de datos, mediante un generador de números aleatorios se selecciona uno de las 5 bases de datos con los datos completos mediante imputación. Inicialmente, la idea fue utilizar conjuntamente las 5 bases de datos generadas mediante imputación. Esto da una base de datos de 580 filas. Pero esa idea se retiró para evitar el sobreentrenamiento.

La solución de utilizar sólo 116 filas como en la base de datos original es más fiel a la situación real de número de pacientes del ensayo. Además, evita el sobreentrenamiento ya que no expone a los algoritmos a 580 filas que prácticamente

las mismas 116 filas repetidas 5 veces. El inconveniente es que la base de datos es pequeña, y eso genera unos *datasets* de entrenamiento y de test de tamaño muy pequeño, que consecuentemente no surtirán de suficientes ejemplos a los algoritmos; resultando en peor capacidad predictiva.

Tras sufrir varios errores generados durante la ejecución de los diferentes algoritmos de aprendizaje automático, se decidió convertir las columnas tiempo al fracaso y fracasos a la clase numérica. Y una vez hecho esto, se procede a separar los *datasets* de entrenamiento y test mediante muestreo aleatorio. Se escogió un tamaño del dataset de entrenamiento de un 75%, para compensar en lo posible el pequeño tamaño del *dataset* original.

5.3 Aplicación de los algoritmos de aprendizaje automático

5.3.1 Bosque de supervivencia aleatorio

Se comienza aplicando la función *tune* del paquete *randomForestSRC* para calcular los valores óptimos de dos parámetros del árbol. Concretamente *mtry* y *nodesize*. Una vez obtenidos estos parámetros, definimos el modelo de supervivencia siendo la columna de Tiempo_al_fracaso la columna que marca el tiempo, y la que marca el estatus de censura o no es la variable fracasos. Como valores de *mtry* y *nodesize* ponemos los obtenidos como óptimos según la función *tune*.

Luego se utiliza la función *predict.rfsrc* para obtener la tasa de error cuando el modelo entrenado con los datos de entrenamiento se enfrenta a los datos de test.

5.3.2 Máquina de soporte vectorial de supervivencia

Para este algoritmo, se va a utilizar la función *survivalsvm* del paquete *survivalsvm*. El modelo de supervivencia es el mismo que en el caso anterior. La variable *Tiempo_al_fracaso* es la que indica el tiempo, y la variable *fracasos* nos indica el estatus de censura o no.

Luego se utiliza la función *predict* del paquete *stats*, en este caso para obtener los valores predichos.

5.3.3 Boosting

Para este algoritmo, se va a utilizar la función *glmboost* del paquete *mboost*. El modelo de supervivencia sigue siendo el mismo que en los dos algoritmos anteriores.

Al igual que en el algoritmo de máquina de soporte vectorial, se utiliza la función *predict* del paquete *stats*, y también es para obtener los valores predichos.

5.4 Evaluación del rendimiento de los algoritmos aplicados

5.4.1 Índice C de Harrell

El índice C de Harrell es una medida que puede ser calculada mediante diferentes fórmulas en múltiples librerías disponibles en R:

- *cindex* de la librería *pec*.
- *cindex* de la librería *dynpred*.
- *estC* del paquete *compareC*.
- *concordance* del paquete *survival*.
- *concordance.index* del paquete *survcomp*.
- *UnoC* del paquete *SurvAUC*

De entre todas estas opciones se probaron infructuosamente las funciones *UnoC*, *concordance.index*, y *cindex*. Finalmente se dio con la metodología correcta para lo que se estaba realizando en este caso.

Se comienza evaluando el índice C de Harrell del algoritmo de Boosting. Para ello se utiliza la función *rcorr.cens* del paquete *Hmisc*. Se le indica que utilice los datos de las predicciones calculadas previamente, y que las aplique al modelo de supervivencia previamente comentado.

Posteriormente se procede de la misma manera para calcular el mismo índice para el modelo de máquina de soporte vectorial de supervivencia.

Para calcular el índice C de Harrell del modelo de bosque de supervivencia aleatorio se calcula el contrario del error de predicción del modelo. Para ello se utiliza la función *get.cindex* del paquete *randomForestSRC*. De manera previa se deben calcular las predicciones *Out Of Bag* del modelo.

Los índices C de Harrell obtenidos por los algoritmos de aprendizaje automático son los siguientes:

Índice	Bosque supervivencia	Máquina soporte vectorial	Boosting
C de Harrell	0.7960526	0.7333333	0.8

Figura 27: Índices C de Harrell

5.5 Análisis de sensibilidad

El análisis de sensibilidad se realiza con el fin de estudiar la influencia que tienen las violaciones del modelo de datos perdidos (en este caso MAR) sobre la inferencia obtenida utilizando datos imputados.

En este caso se van a alterar los valores de la variable CD4A_36 con el fin de observar el efecto que se produce sobre ésta y otras variables. Los valores que se van a utilizar (sumar o restar) para alterar el valor de esa variable son: (-80,-60,-40,-20,0,20,40,60,80).

A continuación se realiza un gráfico de la supervivencia para datos censurados. Para ello se utiliza la función *survfit* del paquete *survival*. En el gráfico se puede ver que los pacientes con datos faltantes tienen diferente probabilidad de fracaso virológico con respecto a la población de pacientes sin datos faltantes.

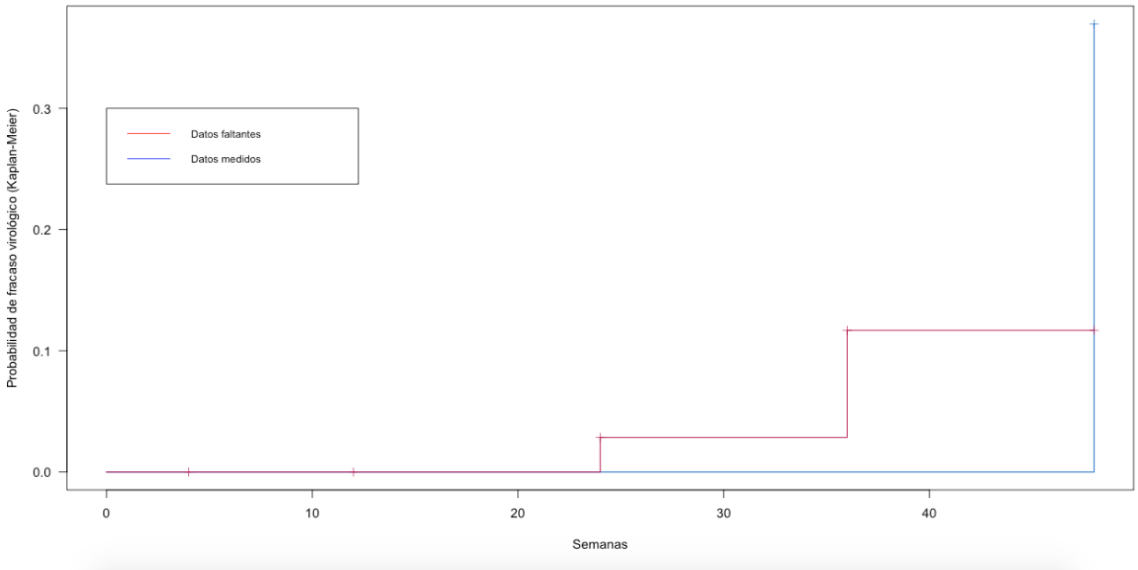

Figura 28: Probabilidad de fracaso virológico vs tiempo

Posteriormente mediante el empleo de los valores acumulados en el vector delta, simulamos imputaciones en las que la variable CD4A_36 presenta un tipo de pérdida de datos que es MNAR.

Tras realizar la imputación anidada en la que se han utilizado los valores de delta. Se cambia el formato en el que se encuentra almacenada este *dataset*, con el fin de poder realizar los siguientes procedimientos. Para ello se utiliza la función *nested.datlist* del paquete *miceadds*. En el siguiente paso se estraen las imputaciones realizadas con 3 valores de delta (-80,0,80), para ello se recurre a la función *subset_datlist* del paquete

miceadds. Casi para finalizar, se realizan gráficos de caja y bigotes para ver gráficamente la influencia que estos valores distorsionados tienen en diferentes variables. Estos gráficos de caja y bigotes se realizan mediante la función *bwplot* de la librería *lattice*.

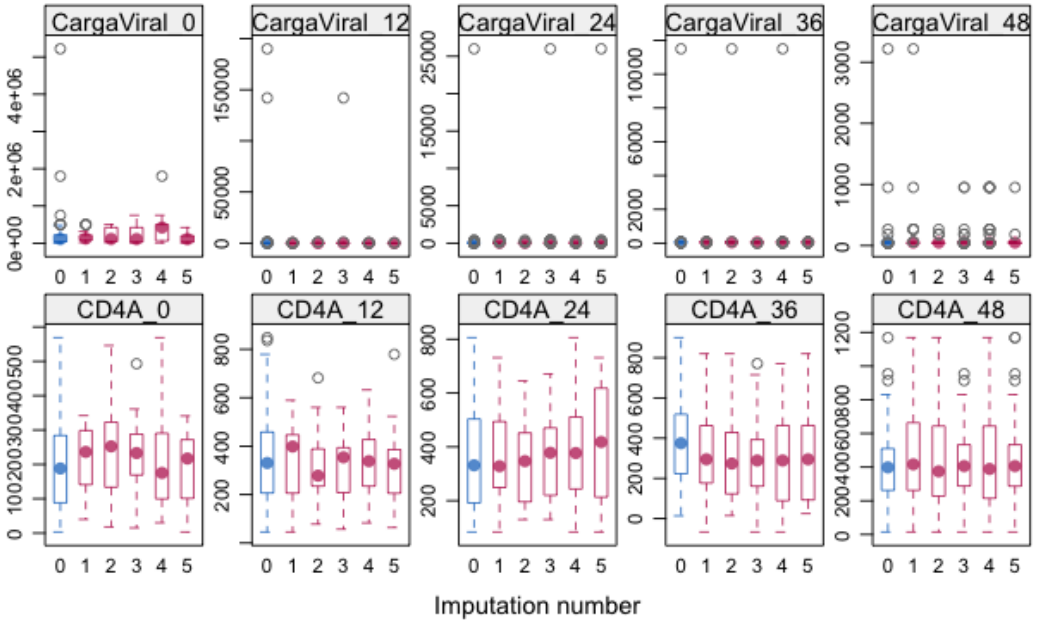

Figura 29: Caja y bigotes con delta valor -80

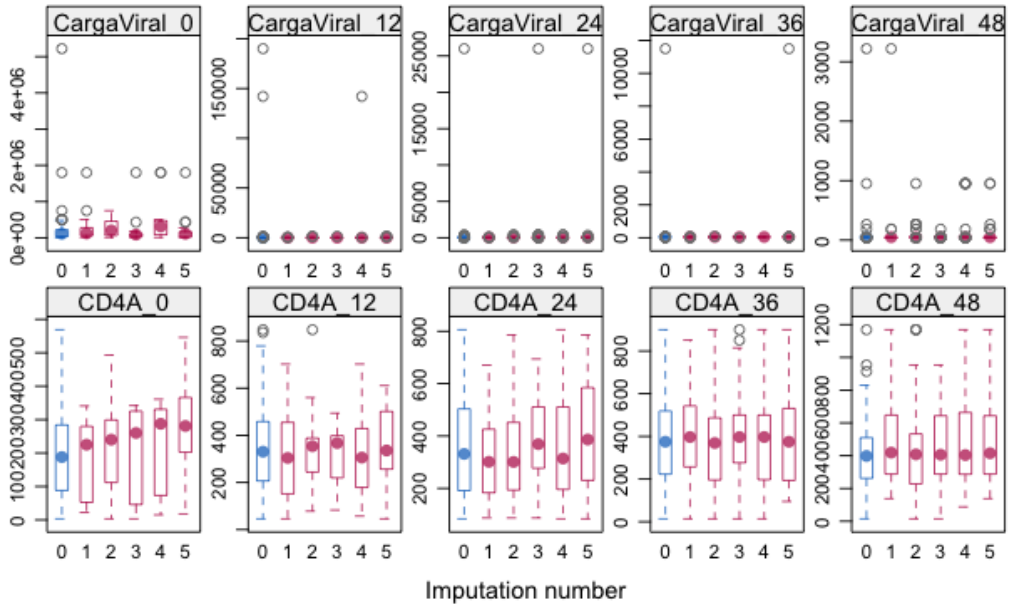

Figura 30: Caja y bigotes con delta valor 0

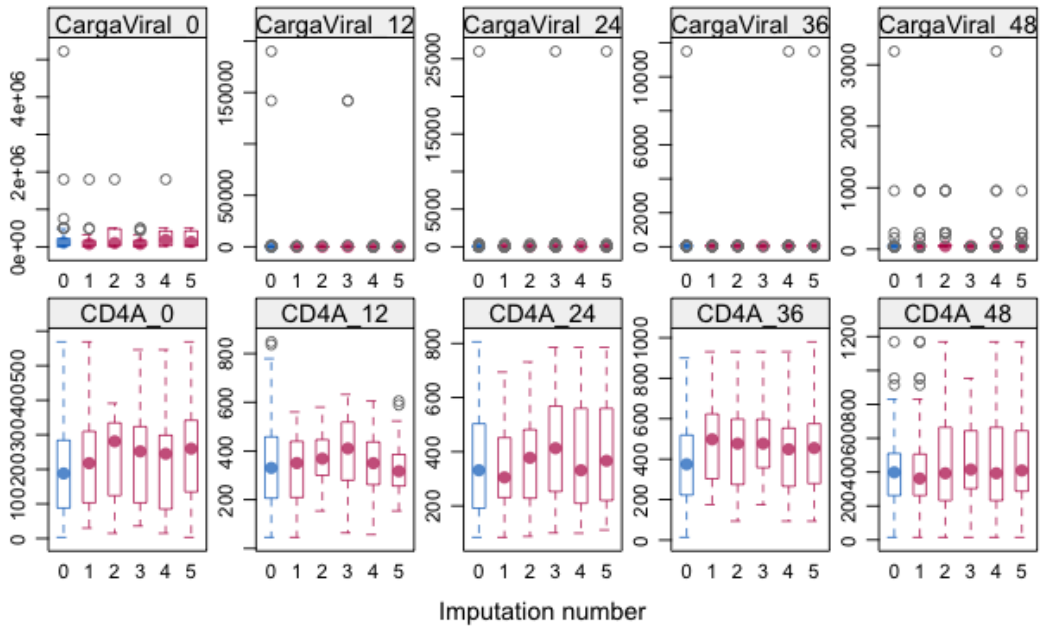

Figura 31: Caja y bigotes con delta valor 80

El último paso en el análisis de sensibilidad es la realización de gráficos de densidad para ofrecer otra visión de la distribución de las variables, y así dirimir si la modificación de los datos altera la distribución de valores de manera significativa. Para ello se utiliza la función *densityplot* del paquete *lattice*

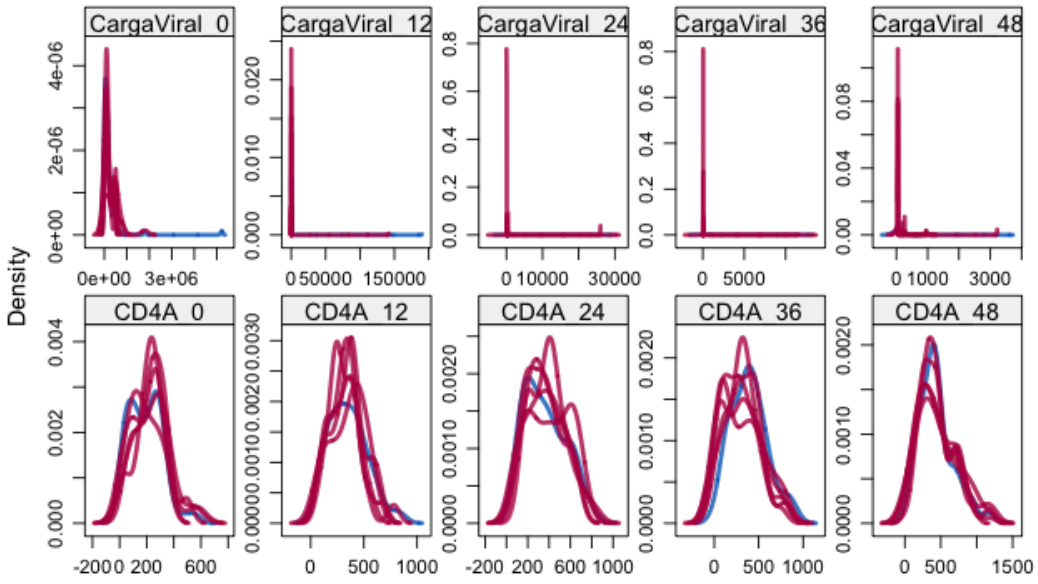

Figura 32: Gráfico de densidad delta valor -80

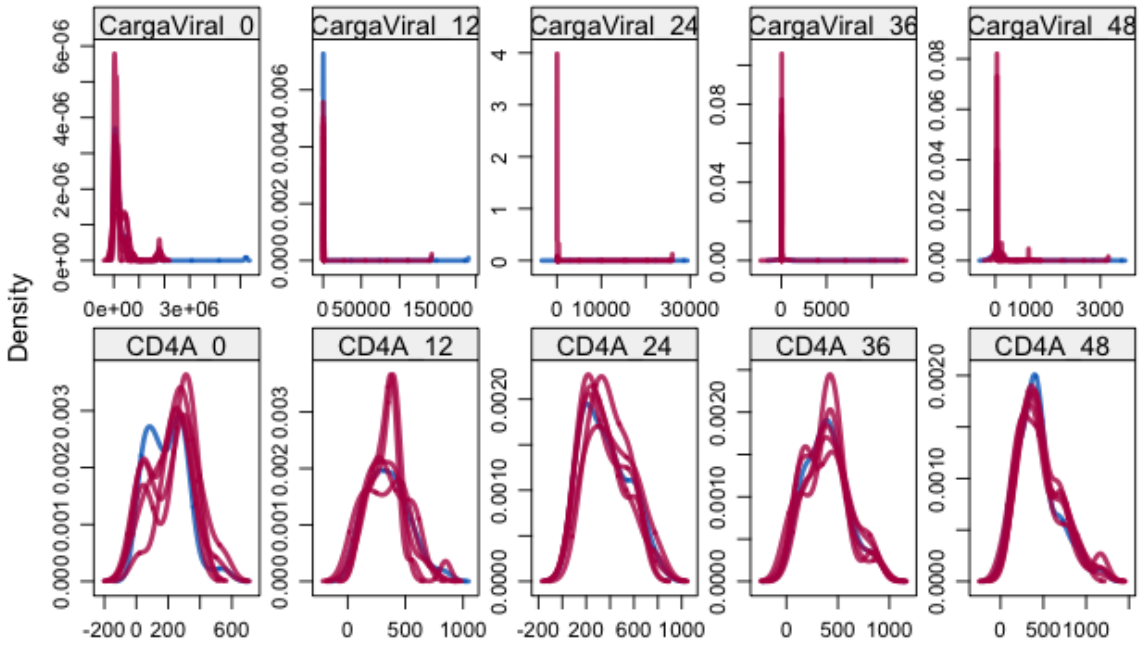

Figura 33: Gráfico de densidad delta valor 0

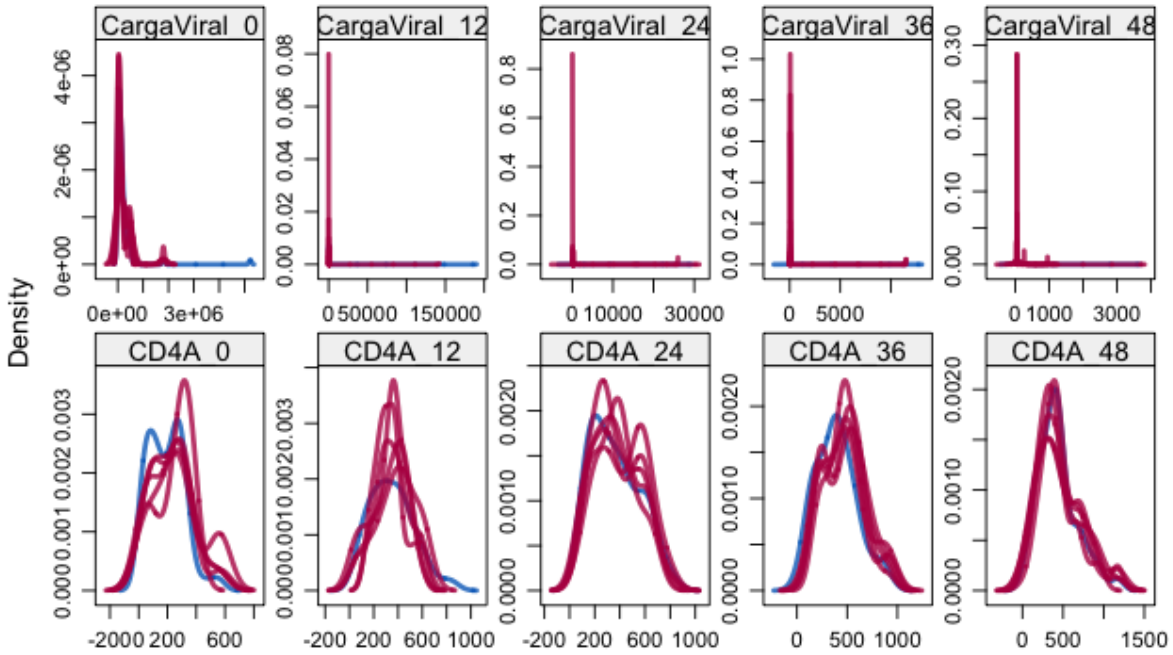

Figura 34: Gráfico de densidad delta valor 80

Se puede observar que las diferencias entre las distribuciones son leves a pesar de los valores extremos que se utilizaron.

5.6 Comparación de los dos tratamientos antirretrovirales

Inicialmente, para realizar esta comparación se pensó en una medición del número de fracasos en cada una de las poblaciones, y posterior comparación entre ellas. Y si ésto no resultaba concluyente, en la que hiciese aumentar en mayor medida el número de linfocitos CD4. Cuando llegó el momento de realizar esta actividad, se decidió que era conveniente cambiar este objetivo para realizarlo de diferente manera. Con un procedimiento más ampliamente utilizado y con mayores visos de ofrecer resultados robustos y concluyentes.

Finalmente el camino seguido fue el de realizar un test T para comparar las medias de los dos tratamientos, que son grupos independientes entre si. Se comenzó utilizando la variable Carga_Viral_48 porque la carga viral es la medida más trascendente para cualquier paciente con VIH. Y como segunda variable para la realización de un test T se escogió el contaje de linfocitos CD4 en la semana 48 (CD4A_48).

Una de las condiciones para la realización de la t de student es la normalidad de los datos. Para ello asegurarse de este hecho, se estudió la normalidad de manera gráfica mediante un gráfico de caja y bigotes, y también se realizó un test de Shapiro-Wilk.

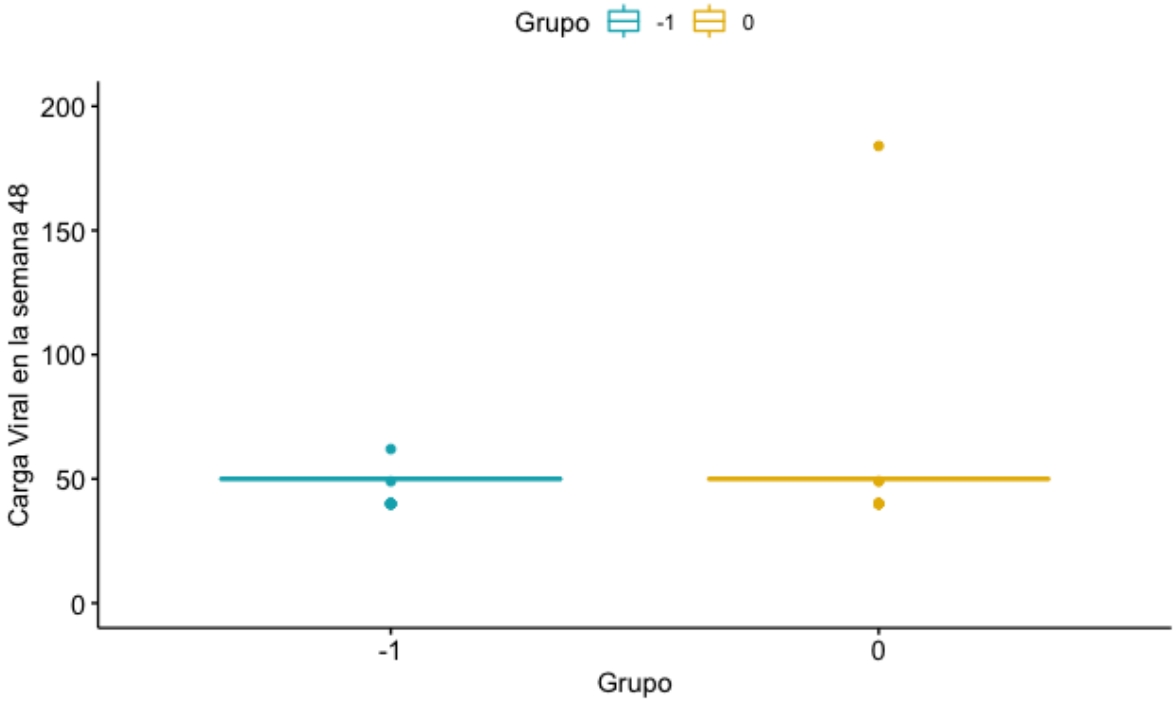

Figura 35: Caja y bigotes de la variable CargaViral_48 para los dos tratamientos

La gráfica de caja y bigotes no parece mostrar indicios que hagan dudar de la normalidad de los datos. Sin embargo al realizar el test de Shapiro-Wilk el resultado muestra que se trata de datos que se distribuyen mediante una distribución no normal. Eso obliga a utilizar el test de Wilcoxon para muestras independientes.

```
        Wilcoxon rank sum test with continuity correction

data:  CargaViral_48 by Grupo
W = 1701, p-value = 0.8976
alternative hypothesis: true location shift is not equal to 0
```

Figura 36: Test de Wilcoxon para la variable CargaViral_48

El resultado del test para la variable CargaViral_48 indica que no podemos rechazar la hipótesis nula de la igualdad de medias entre las dos muestras.

Como el resultado de la comparación de las cargas virales ha resultado en igualdad, se realiza el mismo procedimiento pero utilizando los datos de la columna CD4A_48.

Se comienza con la gráfica de caja y bigotes, que vuelve a no mostrar indicios que hagan dudar de la normalidad de los datos.

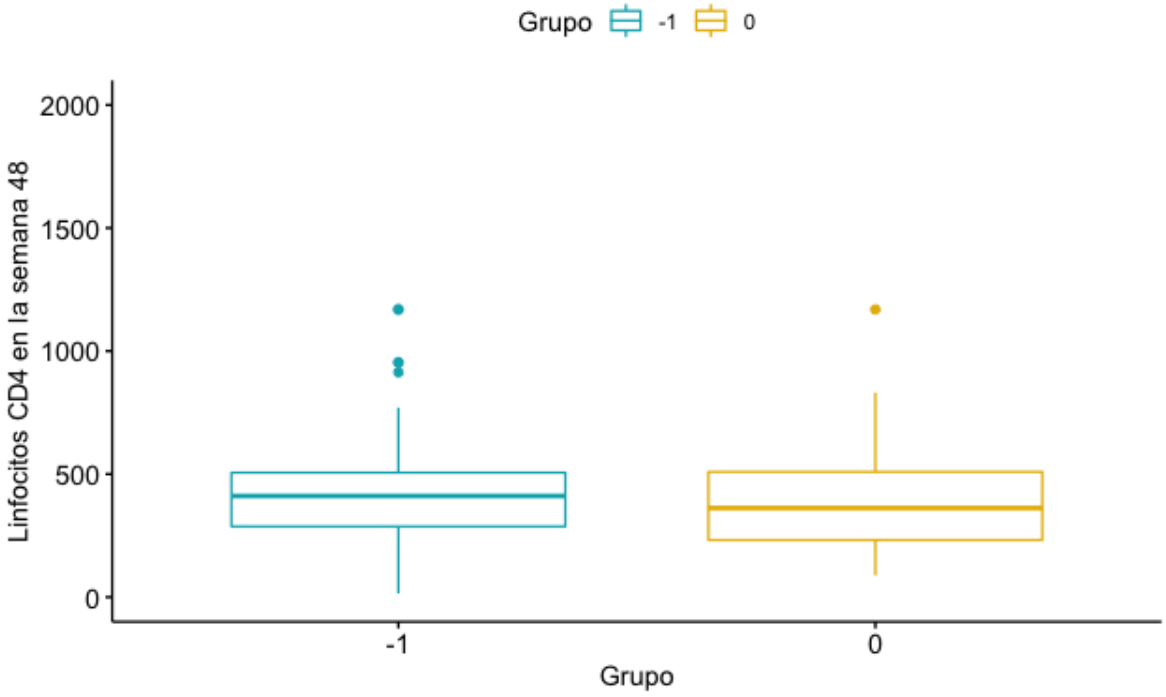

Figura 37: Caja y bigotes de la variable CD4A_48 para los dos tratamientos

De todos modos se realiza un test de Shapiro-Wilk para corroborar este extremo. El resultado del test es que tampoco los datos de esta variable se distribuyan según una normal. Esto obliga a realizar el test de Wilcoxon para muestras independientes

```
        Wilcoxon rank sum test with continuity correction

data:  CD4A_48 by Grupo
W = 1849.5, p-value = 0.3562
alternative hypothesis: true location shift is not equal to 0
```

Figura 38: Test de Wilcoxon para la variable CD4A_48

El resultado del test para la variable CD4A_48 indica que no podemos rechazar la hipótesis nula de la igualdad de medias entre las dos muestras.

6. Conclusiones y autoevaluación

6.1 Conclusiones y discusión

Las conclusiones de este TFM son las siguientes:

No hay diferencias entre grupos en la carga viral en la semana 48 de tratamiento, ni en el contaje de linfocitos CD4.Por lo tanto, los tratamientos antirretrovirales ensayados han tenido un resultado equivalente. Pero teniendo en cuenta que el tratamiento A pretende sustituir al tratamiento B ofreciendo mayor simplicidad en la toma de cápsulas. Debido a esta simplicidad,se elige la combinación de kivexa más efarivenz (tratamiento A) como mejor tratamiento de los testeados.

Tal y como se indica en el punto que trata sobre los antirretrovirales, la adherencia a un tratamiento TARGA es crítica para obtener una supresión viral duradera. Esta supresión viral duradera es la que incrementa la esperanza de vida de los pacientes con VIH. Por lo que cualquier mejora en la adherencia al tratamiento, lleva aparejadas mejoras en la duración de la vida de los pacientes con VIH.

Con respecto a los algoritmos de aprendizaje automático, el algoritmo de boosting es el que ha conseguido mejor desempeño al obtener un índice C de Harrell de 0,8. Los algoritmos de bosque aleatorio de supervivencia y de máquina de soporte vectorial de supervivencia obtienen un índice C de 0,796 y 0,733 respectivamente.

El índice C mide la bondad de ajuste de clasificaciones binarias de una manera equivalente al área bajo la curva ROC. Así que si se tienen en cuenta los valores

obtenidos por los tres algoritmos entrenados con los datos de este TFM, el rendimiento obtenido por los tres modelos es bueno. Aunque todavía les queda un trecho para alcanzar un índice C de 1, que es el de un modelo que predice perfectamente todos los resultados.

6.2 Autoevaluación

En líneas generales, la consecución de los objetivos marcados al inicio de este TFM ha sido prácticamente plena.

Si se miran los objetivos generales, el primero de ellos habla de "la selección de la combinación de datos censurados y algoritmo aprendizaje automático que alcance un mejor desempeño en el índice C de Harrell y en la puntuación de Brier". Es en este objetivo general en el que no se ha conseguido el pleno cumplimiento. La causa ha sido la puntuación de Brier, ya que fue imposible calcular esta puntuación y que diese valores en el rango correcto (de 0 a 1).

De la misma manera que en el objetivo general 1 no hay un cumplimiento total, en el objetivo específico 1.5 tampoco se consigue alcanzar la plena consecución de resultados. Este objetivo consistía en evaluar el desempeño de los algoritmos mediante la puntuación de Brier y el índice C de Harrell. Debido a la imposibilidad de calcular una puntuación de Brier correcta. Ésto hace que la representatividad de la selección del mejor algoritmo de la que habla el objetivo 1.6 sea menor, al basar toda la decisión en un solo indicador (índice C).

El fracaso en la obtención de esta puntuación es fruto del mayor inconveniente del software estadístico R. La multitud de paquetes de los que dispone, generan resultados en los que el objeto generado pertenece a una clase particular. Pero en muchas ocasiones esa clase particular no dispone de métodos con los que calcular determinadas funciones. De todos modos, esta contigencia estaba reflejada en el análisis de riesgos, y se actuó tal y como estaba planificado en este caso.

En el resto de objetivos tanto generales como específicos se logró un 100% de cumplimiento. Para ello tuvieron que ponerse en marcha varias de las medidas de contingencia fijadas al inicio del proyecto, debido a varios imprevistos surgidos en el transcurso de este TFM, y de los cuales se fue informando en las respectivas entregas parciales de este TFM.

Concretamente las medidas de contingencia, y los cambios en los objetivos fueron las siguientes:

- En lugar de utilizar el método de calibración para trabajar con los datos censurados, se vio que dado el número de datos faltantes en la base de datos

utilizada, habría que utilizar una imputación para poder llevar a cabo el TFM. En este caso la medida de contingencia llevada a cabo fue consultar con la tutora para replantear el procedimiento a utilizar, a fin de preparar los datos para su utilización por los algoritmos de aprendizaje automático.

- Se utilizó el algoritmo de máquina de soporte vectorial de supervivencia en lugar del algoritmo de máquina de aprendizaje extremo de supervivencia.
- Se utilizó el algoritmo de boosting en lugar del algoritmo de aprendizaje activo de supervivencia.
- Se calculó únicamente el índice C de Harrell debido a la imposibilidad de obtener una puntuación de Brier correcta.

Otros objetivos se modificaron sin mediar medida de contingencia, porque durante el transcurso del TFM se entendió que era una manera más correcta de realizar esa comparación. Concretamente, se habla de la comparación entre tratamientos.

Las causas de los éxitos conseguidos en el proyecto han sido, principalmente el establecimiento de medidas de contingencia que han funcionado de manera excelente cuando han sido necesarias, y la excelente labor de la tutora ya que hubo que recurrir en múltiples ocasiones a su ayuda para no sufrir un atasco definitivo que diese al traste con este proyecto.

Como aprendizaje principal tras la realización de la parte práctica del TFM, queda la certeza de que antes de plantear qué metodologías aplicar sobre los datos, es mejor mirar cómo están dispuestos estos. Si se hubiese realizado este primer análisis de los datos en un momento más temprano, seguramente se hubiese empezado antes con el procedimiento de imputación. Ésto hubiera dado más tiempo para realizar otras partes del TFM, ya que hubo que destinar tiempo de la escritura de esta memoria definitiva a la parte práctica.

Esa es también la principal modificación que se realizaría sobre la propuesta inicial. Siendo la otra la utilización de otras medidas para medir el desempeño de los modelos de aprendizaje automático. El índice C podría ser así complementado con la exactitud (*accuracy*), con la precisión (*precision*), o con la exhaustividad (*recall*)..

Las línes de trabajo futuro pueden ir por el camino de probar otras librerías para comparar si la imputación mediante estas otras librerías ofrece mayor rendimiento en alguno de los algoritmos utilizados, o en todos ellos en general. Además de manejar más métricas de evaluación del desempeño del modelo.

7. Glosario

Antirretroviral: Medicamento que combate a los retrovirus, como por ejemplo el virus del VIH.

Aprendizaje automático: En inglés *machine learning* es la especialidad, que mediante la computación consigue algoritmos con los cuales los ordenadores pueden realizar tareas como la clasificación, o la predicción.

Ensayo clínico: Investigación llevada a cabo para medir los beneficios de un tratamiento farmacológico. Todos los tratamientos farmacológicos deben superar estos ensayos. Éstos pueden clasificar en fases según la finalidad concreta de la investigación llevada a cabo (fase I la más básica, y fase IV la más avanzada)

CD4: Población de glóbulos blancos que son atacados por el virus VIH, provocando un importante efecto pernicioso sobre el sistema inmunitario del paciente. Pudiendo provocar SIDA y la muerte del enfermo.

VIH: Acrónimo de Virus de la Inmunodeficiencia Humana. Agente infeccioso causante de la enfermedad del SIDA (Síndrome de la Inmunodeficiencia Adquirida)

8. Bibliografía

1. Vinzamuri B, Li Y, Reddy CK. Active learning based survival regression for censored data. CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management. 2014;241–50.

2. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. ACM Computing Surveys [Internet]. 2019;51(6):1–39. Available from: http://arxiv.org/abs/1708.04649

3. Taylor JMG. Random survival forests. Journal of Thoracic Oncology [Internet]. 2011;6(12):1974–5. Available from: http://dx.doi.org/10.1097/JTO.0b013e318233d835

4. Roy J, Lin X. Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. Biometrics. 2005;61(3):837–46.

5. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: From design to analysis. Yale Journal of Biology and Medicine. 2013;86(3):343–58.

6. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2002.

7. Calafati RO. Estrategias para el tratamiento de datos faltantes ("missing data") en estudios con datos longitudinales [Internet] [PhD thesis]. 2017. p. 82. Available from: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/64085/6/romancalafatiTFG0617memoria.pdf

8. Buuren S van. Flexible Imputation of Missing Data, Second Edition [Internet]. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman; Hall/CRC; 2018. Available from: https://www.taylorfrancis.com/books/9780429492259

9. Echeverría P, Negredo E, Carosi G, Gálvez J, Gómez JL, Ocampo A, et al. Similar antiviral efficacy and tolerability between efavirenz and lopinavir/ritonavir, administered with abacavir/lamivudine (Kivexa), in antiretroviral-naïve patients: A 48-week, multicentre, randomized study (Lake Study). Antiviral Research. 2010;85(2):403–8.

10. Medicine AB of I. ABIM Laboratory Test Reference Ranges - January 2020 [Internet]. 2020. pp. 1–12. Available from: https://www.abim.org/{~}/media/ABIM Public/Files/pdf/exam/laboratory-reference-ranges.pdf