Redes bayesianas y análisis de supervivencia

Modelos gráficos

Ricardo Cruz Martínez, Vazquez Arriaga Jorge & Martínez Mejía Gerardo

1. Introducción.

Las *redes bayesianas* se han ido popularizando a lo largo del tiempo ya que se tienen varios campos de aplicación como la medicina, genética y la toma de decisiones, por mencionar algunos ejemplos; sin embargo, la mayoría de aplicaciones recae en Inteligencia Artificial. Esto tiene sentido pues en el enfoque bayesiano se busca "aprender" de los datos observados.
En este proyecto se buscará introducir el tema de *redes bayesianas* así como conceptos básicos del *análisis de supervivencia* con el fin de juntar ambas ramas de la estadística e implementar modelos que nos ayuden a estimar probabilidades de supervivencia dado un tiempo $\mathbf{T}$.
La base de datos con la que trabajaremos se encuentra en la paquetería `survival` llamada `nafld`, la cual consta de 3 datasets. Para los fines de este proyecto usaremos el dataset llamado `nafld1`, que tiene 17549 observaciones y 9 variables, la cual consta de un estudio poblacional de la enfermedad del hígado graso (no alcohólico) donde cada observación pertenece a alguna persona y se tiene registro sobre género, edad, peso, estatura, etc. Los detalles los veremos en la sección 4.

2. Redes Bayesianas.

Las *redes bayesianas* se pueden pensar como una mezcla de probabilidad con la teoría de gráficas. Dicho en otras palabras, son modelos gráficos que representan las relaciones probabilísticas entre un número de variables. Para poder explicar de manera más clara este tema debemos introducir algunos conceptos de teoría de gráficas.

2.1. Conceptos Básicos de Teoría de Gráficas.

Para fines de este proyecto no revisaremos a profundidad esta área de las matemáticas pues sólo necesitaremos las siguientes definiciones:

1. *Gráfica.* Decimos que G es una gráfica si está compuesta por dos conjuntos: el conjunto de vértices $V(G)$ y el conjunto de aristas $A(G)$. Donde $V(G)$ es finito no vacío y $A(G)$ es un conjunto finito cuyos elementos están formados por elementos de $V(G)$. Lo denotamos por $G = (V, A)$.
2. *Gráfica Dirigida.* Decimos que $G = (V, E)$ es una gráfica dirigida donde E lo definimos como:
$$E \subseteq \{(a, b) \in V \times V : a \neq b\}.$$
De manera más intuitiva, dado un elemento $(a, b) \in E$, lo podemos pensar como la arista que va del vértice a al vértice b en ese sentido. De ahí el nombre de gráfica dirigida.
3. *Gráfico Dirigido Acíclico.* Decimos que $G = (V, E)$ es una gráfica dirigida acíclica (GAD) si no hay un camino dirigido $A_1 \to A_2 \to \cdots \to A_n$ tal que $A_1 = A_n$ con, $A_1, \cdots, A_n \in V$.
En teoría de gráficas se entiende por *camino* a la secuencia de vértices distintos $v_1, \ldots, v_n$ tal que v_i se conecta con v_{i+1} para cada $i = 1, \ldots, n-1$. En este caso decimos que es un camino de tamaño n.

2.2. Conceptos Básicos de Redes Bayesianas.

Vistas las definiciones en la sección anterior (2.1) tenemos los elementos necesarios para definir lo que es una red bayesiana.

1. *Red Bayesiana.* Una red bayesiana (RB) es una gráfica acíclica dirigida en el que cada vértice representa una variable aleatoria que tienen asociada una función de probabilidad condicional, las cuales pueden ser interpretadas como relaciones causa-efecto. Las aristas que van de un vértice v_i a otro v_{i+1} nos indican la influencia directa que hay de v_i sobre v_{i+1}, en este caso, diremos que v_i es un vértice "padre" de v_{i+1}, a este último vértice lo podemos pensar como un vértice "hijo". Por otra parte, para cada nodo hay una tabla de probabilidad condicional que sirve para cuantificar los efectos de los "padres"sobre el nodo.

La estructura de la red bayesiana provee información sobre las relaciones de dependencia e independencia condicional existentes entre variables. Estas relaciones simplifican la representación de la función de probabilidad conjunta como el producto de las funciones de probabilidad condicional de cada variable. En términos matemáticos, se puede representar de la siguiente manera:

$$P(\mathbf{X}) = P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)),$$

Donde $\mathbf{X}$ representa las variables involucradas y la notación $Pa(X_i)$ los vértices padres de la variable X_i.

2.3. Supuestos distribucionales.

En principio hay muchas posibilidades para elegir las distribuciones locales y la global, sin embargo, es común encontrar solamente dos casos:

- *Variables Multinomiales.* Principalmente utilizado para variables categóricas. Tanto la distribución global como las distribuciones locales siguen una distribución multinomial; esta última representada como tablas de probabilidad condicional. A este tipo de redes bayesianas se les conoce como discretas.
- *Variables Normales Multivariadas.* Caso contrario a la definición anterior. Esta representación es utilizada principalmente para variables continuas. Se hace la suposición de que la distribución global tiene distribución normal multivariada, mientras que las distribuciones locales tienen distribución normal univariada limitadas por restricciones lineales. Las distribuciones locales son modelos lineales en los que los padres juegan el papel de variables explicativas. Este tipo de redes bayesianas se le conocen como Gaussianas.

2.4. Discretización.

Como ya mencionamos antes, existen las *redes bayesianas discretas* y las *redes bayesianas Gaussianas*, sin embargo, cuando se tienen variables continuas y discretas, la manera sencilla de resolver este inconveniente es discretizando las variables que sean continuas. Dado que nuestro dataset contiene variables explicativas continuas y discretas (ver sección 4), será conveniente utilizar un modelo de *red bayesiana discreta*.
A continuación presentamos un ejemplo de modelo de red bayesiana para nuestros datos.

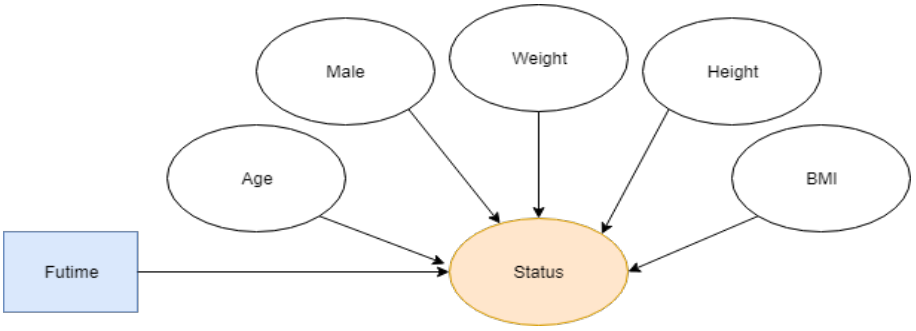


Figura 1: Ejemplo de Red Bayesiana.

De la figura 1 se tiene que todas nuestras variables predictoras de nuestro dataset son nodos "padres" de la variable `Status` por lo que, en principio, solo tendríamos una tabla de probabilidad la cual consistiría en calcular las probabilidades condicionales $P($`Status` $\mid$ `Age, Male, Weight, Height, BMI`$)$. La variable `futime` será nuestra representación del tiempo de manera explícita, es decir, no se considerará un nodo padre de nuestra variable `Status`. Daremos más detalles de este supuesto en la sección 5, sin embargo, igualmente veremos que sucede si metemos la variable del tiempo directamente en nuestro modelo.
Es claro que este modelo puede cambiar, por ejemplo, se sabe que el índice de masa corporal (variable *bmi*) depende del peso (variable *weight*) por lo que este último podría pasar a ser un nodo padre del IMC. De igual manera la estatura con el género podría ser una opción, no obstante, se tendría que hacer un análisis descriptivo de los datos para poder crear diferentes estructuras que podrían ser de ayuda (ver sección 4). Una vez que se defina la estructura de nuestra red, dado que no conocemos la tabla de probabilidades, el siguiente paso sería "entrenar" nuestra red. El entrenamiento es el proceso de estimación de esta(s) tabla(s). Para esta tarea existen varios algoritmos, sin embargo, en este caso usamos el método bayesiano que es usando el valor de la esperanza de la distribución *a posteriori* tomando una distribución inicial poco informativa.
Una gran ventaja de usar redes bayesianas es que comúnmente nos encontramos con datos faltantes y el modelo de riesgos proporcionales de Cox (véase la siguiente sección) no puede realizar inferencia sobre la probabilidad de

supervivencia, sin embargo, las redes bayesianas pueden realizar esta estimación aún si se tienen datos faltantes, no obstante, se necesitaría cierto tipo de arquitectura en nuestra red para que esto suceda. Por simplicidad, consideraremos que nuestros datos no contienen valores nulos.
En la literatura existen dos perspectivas para el ajuste de una red bayesiana: las hay *estáticas* y *dinámicas*[5]. La principal diferencia que hay entre una y otra es que la red bayesiana *dinámica* modela asociaciones que surgen de la dinámica temporal entre entidades de interés, mientras que la red bayesiana *estática* hace una captura instantánea del sistema en un tiempo determinado, por lo que necesitaríamos representar el tiempo explícitamente agregando una variable de indexación T para el tiempo, capturando cada punto discreto en el tiempo que sea de interés. Situación que no ocurre con el modelo de riesgos proporcionales de Cox (ver sección 3.4).

3. Conceptos Básicos de Análisis de Supervivencia.

El *análisis de supervivencia* es una de las ramas más antiguas de la estadística cuyo principal objetivo es, como su nombre lo dice, estimar la probabilidad de "falla" o "muerte" de un aparato electrónico o una persona con ciertas características en cierto tiempo. Esto se lleva a cabo estimando la *función de supervivencia* la cual definiremos más adelante.
Por otra parte, hay un concepto a tener en cuenta y son los *datos de supervivencia*. Estos datos se generan a partir del interés de estudiar el tiempo que transcurre entre un evento inicial y un evento final que define el término del estudio para cada individuo (recordemos que no necesariamente esto se aplica a personas, sin embargo, estos estudios son comunes en el área biomédica). Al tiempo transcurrido entre estos dos eventos se le denomina *tiempo de falla, tiempo de supervivencia o tiempo de muerte*.
Además, los estudios de supervivencia pertenecen a los *estudios longitudinales* en los que los individuos se siguen a través del tiempo, en este caso, el seguimiento es desde que comienza el experimento hasta el tiempo de falla.

3.1. Características de los datos de supervivencia.

Principalmente se tienen dos características:

- *Tiempo de supervivencia.* El tiempo de supervivencia se toma como una variable aleatoria positiva (generalmente tienen cola derecha larga, es decir, sesgo positivo).
- *Observaciones censuradas.* En la vida real ocurre que durante algún experimento algunos individuos lo abandonen, por lo que el concepto de censura proviene de la necesidad de tomar esto en cuenta al momento de realizar estimaciones de la supervivencia de algún individuo.

  Los hay de 3 tipos:

  1. *Censura tipo I.* Se da cuando se fija un tiempo máximo de observación/seguimiento de los individuos para que presenten la falla. Las observaciones que no hayan presentado fallas al momento máximo de observación son llamadas *observaciones censuradas*.
  2. *Censura tipo II.* Se decide prolongar el tiempo de observación de los individuos en el estudio hasta que ocurran k fallas de n posibles. Los individuos que no presentaron la falla al completarse estas primeras k representan observaciones censuradas.
  3. *Censura aleatoria.* Como su nombre lo indica este tipo de censura ocurre sin ningún control de investigador, es decir, se puede dar por abandono del individuo, muerte del mismo por causas ajenas al experimento o por alguna otra causa externa al experimento.

  A continuación mencionaremos otros tipos de clasificación de censuras:

  1. *Censura por la derecha.* Para las 3 definiciones anteriores sabemos que de haber ocurrido la falla ocurre después del tiempo de censura observado, es decir, a la derecha.
  2. *Censura por la izquierda.* En este caso, la falla ocurre antes de que el individuo ingrese al estudio por lo que se sabe que su tiempo de falla no observado es menor que el tiempo de censura observado.
  3. *Censura por intervalo.* La observación de los individuos no se realiza de manera continua y los periodos de observación pueden ser muy largos entre dos observaciones consecutivas. En este caso la falla se presenta en el intervalo determinado por estos dos periodos de revisión.

Para este proyecto únicamente consideraremos la *censura tipo I* por la derecha.
Algo importante por mencionar es la notación utilizada para diferenciar tiempos censurados y no censurados. Para esto definimos la siguiente función:

$$\delta = \begin{cases} 1 & si \quad \mathbf{T} \leq C \\ \\ 0 & si \quad \mathbf{T} > C \end{cases}$$

Donde C es el tiempo de censura. Esta notación es importante pues la paquetería `survival` la utiliza para realizar la diferencia antes mencionada.

3.2. Funciones importantes en el análisis de supervivencia.

Ahora revisaremos las 2 principales funciones del análisis de supervivencia que nos serán de ayuda y que serán las que trataremos de estimar por diversos métodos. Para esto definiremos la variable aleatoria $\mathbf{T}$ la cual será no negativa y representará el tiempo de falla de los individuos en el estudio.

- *Función de Supervivencia.* Esta función representa la probabilidad de que un individuo sobreviva a un tiempo determinado t, la denotaremos por $S(t)$ y será representada de la siguiente manera:

$$S(t) = P(\mathbf{T} > t) = 1 - F(t)$$

  Además cumple lo siguiente:

$$S(0) = 1, \quad S(t) = 0 \text{ si } t \to \infty \quad S(t_1) \geq S(t_2) \text{ si } t_1 < t_2$$

- *Función de Riesgo.* Esta función determina la tasa instantánea de falla al tiempo $\mathbf{T} = t$. Dado que el individuo ha sobrevivido un instante antes de t, la denotamos por $h(t)$ y está dada por:

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

  Para el caso discreto se tiene lo siguiente:

$$h(t_j) = P(\mathbf{T} = t_j | \mathbf{T} \geq t_j), \ j = 1, 2, ...$$

- *Función de Riesgo Acumulado.* Para el caso donde $\mathbf{T}$ es continua, esta función se define como:

$$H(t) = \int_0^t h(u)du$$

  Para el caso discreto se define como:

$$H(t) = \sum_{t_j \leq y} h(t_j)$$

3.2.1. Relaciones entre las funciones para el análisis de supervivencia.

Las siguientes relaciones serán de gran ayuda al momento que queramos pasar de una función a otra. En particular, nos fijaremos en las funciones de riesgo y supervivencia. Las demostraciones de estas propiedades quedan fuera de los propósitos de este proyecto, pero se pueden encontrar en [2].

- $h(t) = -\frac{d \log S(t)}{dt}$
- $S(t) = \exp\{-\int_0^t h(u)du\}$

3.3. El estimador Kaplan-Meier (K-M).

El estimador *Kaplan-Meier* es un estimador no paramétrico de la *función de supervivencia* que toma en cuenta las censuras. Se puede probar que este estimador es el estimador máximo verosímil de la *función de supervivencia*, (la prueba queda fuera de los límites de este proyecto, para ver una demostración consulte [1]).
Definimos el estimador K-M de la siguiente manera:

$$\widehat{S}(t) = \prod_{i|t_i<t} \left(1 - \frac{d_i}{n_i}\right),$$

Donde:

- d_i es el número de muertes en el momento t_i.
- n_i es el número de sujetos en *riesgo* justo antes de t_i. De no haber censura, n_i es el número de supervivientes inmediatamente antes del momento t_i. Con censura, es el número de supervivientes menos el número de casos censurados: sólo se observan los sujetos vivos que no se han caído del estudio en el momento en que ocurre una muerte.

3.4. Modelo de riesgos proporcionales de Cox.

Dado lo anterior, se tuvo el supuesto de que $\{T_1, T_2, \ldots, T_n\}$ son variables aleatorias independientes e idénticamente distribuidas, lo cual, es un supuesto muy fuerte pues en la realidad es muy poco probable encontrar poblaciones idénticas de personas (recordemos que trabajaremos con datos de personas) por lo que este modelo va a resultar de gran ayuda.
La idea de este modelo es considerar un vector de covariables Z_i para cada individuo que van a afectar el tiempo de falla. El vector $\underline{Z}_{p\times 1}$ puede denotar el tratamiento o características del individuo, puede ser constante o depender del tiempo. Si tenemos dos individuos con el mismo vector de covariables, definiremos $\underline{Z}$ tal que $\underline{Z} = \underline{0}$ sea nuestra población de referencia, posteriormente, se describe la forma de cómo cambia **T** al pasar de un individuo con covariables $\underline{Z} = \underline{0}$ a otro con covariables $\underline{Z} \neq \underline{0}$.
Una vez definida nuestra población de referencia $\underline{Z} = \underline{0}$, con función de riesgo $h_0(t)$, supondremos que existe $\rho(\underline{Z}) > 0$, tal que:

$$h(t \mid \underline{Z}) = \rho(\underline{Z})h_0(t),$$

Donde

$$\rho(\underline{Z}) = \exp\{\underline{\beta}'\underline{Z}\}$$

Con $\underline{\beta}' = (\beta_1, \ldots, \beta_p)$ (notemos que cuando $\underline{Z} = \underline{0}$ tenemos que $\rho(\underline{Z}) = 1$). Haciendo un sencillo despeje tenemos lo siguiente:

$$\frac{h(t \mid \underline{Z})}{h_0(t)} = \exp\{\underline{\beta}'\underline{Z}\} \Leftrightarrow \log\left(\frac{h(t \mid \underline{Z})}{h_0(t)}\right) = \sum_{i=1}^{n} \beta_i z_i.$$

Una observación importante es que la variable $\rho(\underline{Z})$ no depende del tiempo. En caso de que esto ocurra, es decir, que alguna variable $\underline{Z}_i$ dependa del tiempo, podemos recurrir al modelo extendido de Cox , consulte [1]. Este modelo queda fuera de los propósitos de este proyecto.
Otra propiedad importante a destacar es la siguiente:

$$S(t \mid \underline{Z}) = (S_0(t))^{\exp\{\underline{\beta}'\underline{Z}\}} = (S_0(t))^{\rho(\underline{Z})} \tag{1}$$

La igualdad anterior será de gran importancia, pues es con la que realizaremos nuestras aproximaciones a la función de supervivencia.

4. Análisis Exploratorio del Dataset.

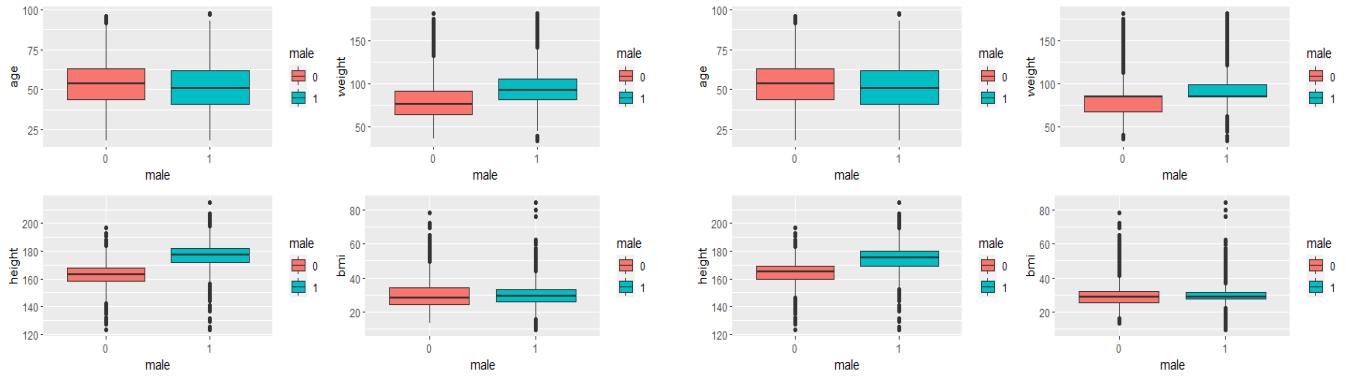
Las variables de nuestro dataset son las siguientes:

1. id	2. age	3. male
4. weight	5. height	6. bmi
7. case.id	8. futime	9. status

Las variables *id, case.id* para nuestro proyecto las descartaremos pues son llaves para poder conectarlas con los demás datasets. Por otra parte, *age* es la edad de los individuos al momento de entrar al estudio, *male* es una variable categórica que es igual a cero si el individuo es femenino y 1 si es masculino. Las variables *weight* y *height* son el peso en kg y la estatura en centímetros de los individuos, *bmi* es el índice de masa corporal, *futime* es el tiempo de muerte o del último seguimiento en días y por último *status* es una variable categórica donde 0 indica que el paciente estaba vivo hasta su último seguimiento y 1 indica fallecimiento del paciente. Dicho lo anterior podemos pensar a la variable *status* como la censura de los datos de la variable *futime*, no obstante, esta variable en nuestra red sería el nodo "hijo" de las variables predictoras, es decir, la podemos interpretar como la variable que nos indicará que una persona sobreviva o no en cierto tiempo (explicamos esto con más a detalle en la sección 5).
Por otra parte, las variables *heigth*, *weigth* y *bmi* contienen valores nulos. Hay 4961 de 17549 tuplas que contienen valores nulos que representa el 28 % de las observaciones por lo que quitar estas tuplas no sería buena idea. Para solucionar esto procederemos a realizar imputaciones en los datos usando la mediana (pues es menos sensible a valores atípicos) de las poblaciones de hombres y mujeres para las respectivas variables. La tabla 1 nos presenta un breve resumen de las variables que no son categóricas una vez realizada la imputación.

	age	weight	height	bmi	futime
Min.	18	33.40	123	9.207	7
1° Qu.	42	75.40	164	26.54	1132
Mediana	53	83.90	169.0	28.88	2148
Media	52.66	85.68	169.4	29.74	2411
3° Qu.	63	92.70	175	31.58	3353
Máximo	98	181.70	215	84.40	7268

Tabla 1: Estadísticas relevantes de los datos no categóricos.



(a) Boxplots por género sin imputación.

(b) Boxplots por género con imputación.

Figura 2: Boxplots de variables continuas por género antes y después de la imputación.

De la tabla 1 notamos que la variable *age* tiene mucha variabilidad, tenemos individuos que tienen desde 18 a 98 años. De igual manera con el peso (variable *weight*), pues el valor máximo es casi 6 veces más grande que el valor mínimo. La variable *height* (altura) presenta valores atípicos, registrando alturas desde 1.23 metros hasta de 2.15 metros, sin embargo, estos casos se pueden dar por lo que no sería conveniente modificarlos u omitirlos. Por otra parte, el índice de masa corporal (variable *bmi*) a ojo parece que los valores son coherentes acorde con los datos que ya tenemos sobre el peso y la altura (pues el IMC está en función de la altura y el peso). Por último, la variable del tiempo de supervivencia (*futime*) presenta tanto valores pequeños como muy grandes, esto aunado a la variable *status* será clave al momento de ajustar algún modelo de supervivencia.
De las figuras 2a y 2b notamos que la imputación de datos hizo que tuviéramos más outliers, sin embargo, las cajas se mantuvieron prácticamente igual y lo mencionado anteriormente queda plasmado de manera gráfica con estas figuras; no obstante notamos que el género influye en algunas variables como el peso y la estatura, la figura 3 nos comprueba esto. Además, es notoria la alta correlación que existe entre las variable *bmi* y *weight* (esto es un indicio muy fuerte para que la variable *bmi* sea vértice padre de la variable *weight* en nuestro modelo pero lo veremos más adelante para el ajuste del modelo). Tenemos que la variable *height* y *weight* tienen una correlación medianamente alta y por último las variables *age* y *height* con una correlación algo baja (igual a -0.161) por lo que el diseño de nuestra red puede ser con base a estas correlaciones. Igualmente viendo que el género tiene un gran peso para algunas variables.

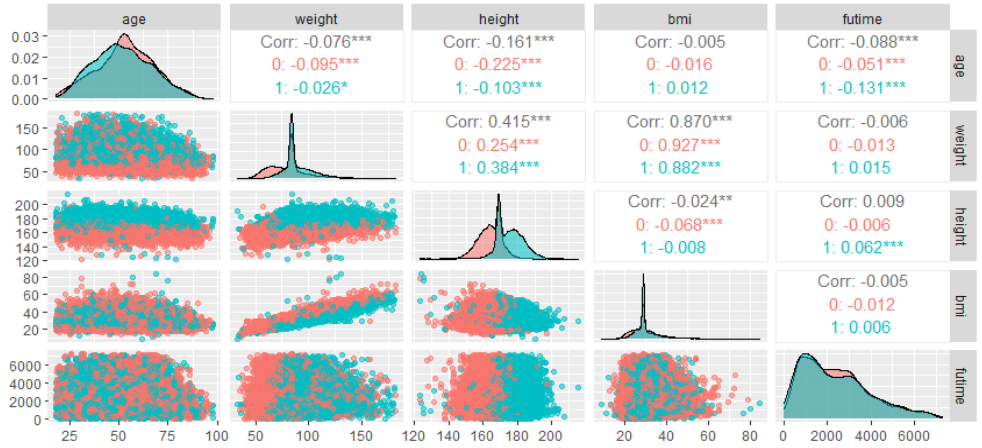


Figura 3: Gráfico de dispersión por pares y correlaciones entre variables.

Finalmente, de las variables categóricas, tenemos lo siguiente:

	0	1
male	9348	8201
status	16185	1364

Tabla 2: Conteos por clase de variables categóricas

De la tabla 2 tenemos que la población de hombres y mujeres está más o menos balanceada, pues tenemos 1147 registros más en el grupo de mujeres que el de hombres. Por otra parte, en la variable *status* tenemos 16185 registros con censura y solamente 1364 fallas. Esto nos va a ocasionar peores estimaciones de la función de supervivencia.
Finalmente, tal como se vio en la sección 2.3, necesitamos que todas nuestras variables explicativas sean continuas o discretas y tal como vimos en esta sección, tenemos una variable binaria `male` por lo que tocaría discretizar las demás variables. Para esta tarea se decidió tomar intervalos de cada variable de tal manera que tuvieran la misma cantidad de elementos, pues por la cantidad de datos que tenemos, si lo hacíamos de otra forma muy posiblemente no tendríamos muestras representativas.
Las variables una vez discretizadas quedaron de la siguiente manera:

Edades	Observaciones
[18,46)	5529
[46,59)	6002
[59,98]	6018

(a) Variable Age.

Pesos	Observaciones
[33.4,81.7)	5847
[81.7,86.2)	5831
[86.2,182]	5871

(b) Variable Weight.

Alturas	Observaciones
[123,166)	5422
[166,172)	6080
[172,215]	6047

(c) Variable Height.

Tabla 3: Frecuencia de variables una vez discretizadas.

5. Ajuste del Modelo.

Para ajustar nuestro modelo primero supondremos que el supuesto de proporcionalidad se cumple, es decir, que $\rho(\underline{Z})$ no dependa del tiempo. Además, supondremos que nuestras variables tampoco dependen del tiempo. Por lo que vimos en la figura 3 las variables `bmi` y `weight` están altamente correlacionadas, por lo que se optó por descartar la variable `bmi`.
Por otra parte, se ha decidido ajustar un modelo de red bayesiana estática la cual se vió en la sección 2, además, tomaremos la variable correspondiente al tiempo (`futime`) y la modificaremos de tal manera que podamos hacer nuestros ajustes de manera anual.
Ajustando un modelo clásico de riesgos proporcionales de Cox considerando todas las variables menos `bmi` sin interacciones tendríamos lo siguiente:

	coef	exp. coef	pval
age[46,59)	1.19	3.30	$<$ 2e-16
age[59,98]	2.73	15.32	$<$ 2e-16
male1	0.61	1.85	$<$ 2e-16
weight[81.7,86.2)	-0.22	0.80	0.002
weight[86.2,182]	-0.18	0.84	0.012
height[166,172)	-0.27	0.77	0.0007
height[172,215]	-0.51	0.60	3.73e-08

Tabla 4: Resumen del modelo de riesgos proporcionales de Cox.

De la tabla 4 tenemos que todas las variables involucradas son estadísticamente significativas, además, notamos que al aplicarle la función exponencial a los coeficientes de las variables las variables de la edad y el género son mayores a 1. Esto nos indica un aumento significativo en la probabilidad de muerte para estos grupos respecto a la población de referencia (la cual consta de las categorías que no aparecen en esta tabla). Caso contrario con las variables que corresponden al peso y la altura, por lo que las respectivas variables de la población de referencia tendrían más probabilidad de muerte.
Dicho todo lo anterior, la arquitectura de nuestra red se vería igual a la figura 1 salvo que no estaremos utilizando la variable `bmi` tal como se ve en la figura 4. Con esta arquitectura de red estaríamos estimando la probabilidad de sobrevivir al tiempo $\mathbf{T} = t$ dadas nuestras variables explicativas. Es una manera diferente de interpretar la ecuación

1, es decir, las probabilidades de nuestra tabla de probabilidades obtenidas de nuestra red nos ayudarán a estimar lo siguiente:

$$P(\texttt{Status} \mid \texttt{male}, \texttt{age}, \texttt{weight}, \texttt{height}, \texttt{T = t}) = S_0(t)^{\rho(Z)}$$

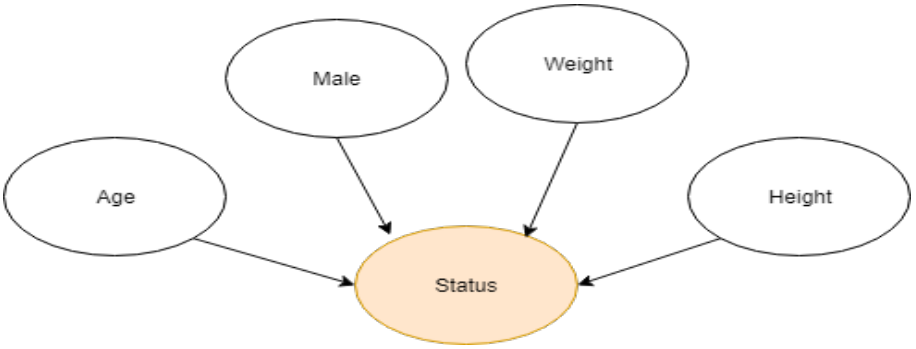


Figura 4: Arquitectura de la primera red bayesiana.

Ahora, vimos en la tabla 1 que la variable `futime` toma valores que van de 7 hasta 7268 y como mencionamos que haremos nuestro análisis de manera anual, se optó por separar nuestros datos de la siguiente manera: en un grupo los datos tales que la variable `futime` fuera menor o igual a 365, luego tomar los datos que fueran mayores a 365 pero menores a $365 \times 2 = 730$ y así sucesivamente hasta finalmente tomar las observaciones que fueran mayores a $365 \times 19 = 6935$.
Hecho lo anterior, para cada subconjunto de variables creadas según el valor asociado a la variable `futime` (dado que se generaron 19 subconjuntos de datos) para cada uno de estos subconjuntos se procedió a ajustar el modelo mencionado anteriormente, es decir, se ajustaron 19 redes bayesianas. Por otra parte, como solamente estamos trabajando con variables discretas tendremos un total de 54 funciones de supervivencia (pues es una por cada combinación de variables de la ecuación anterior) las cuales mostramos en el anexo (figuras 6, 7 y 8). Las gráficas nos dan una idea de la estimación de probabilidades, sin embargo, concluir a partir de gráficas es complicado pues se tienen muchas probabilidades estimadas.
A continuación daremos los modelos que presentan probabilidades menores a 0.5 en algún valor del tiempo:

No. Modelo	Modelo	Tiempo(s)
3	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	1
8	$P(s \mid \texttt{male} = 0, \texttt{weight} = [81.7, 86.2), \texttt{height} = [123, 166), \texttt{age} = [46, 59))$	19
11	$P(s \mid \texttt{male} = 1, \texttt{weight} = [81.7, 86.2), \texttt{height} = [123, 166), \texttt{age} = [46, 59))$	5
12	$P(s \mid \texttt{male} = 1, \texttt{weight} = [81.7, 86.2), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	3,6,13
15	$P(s \mid \texttt{male} = 0, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	17
17	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [46, 59))$	5
18	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	1,13
21	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
24	$P(s \mid \texttt{male} = 1, \texttt{weight} = [33.4, 81.7), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
36	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
39	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [172, 215), \texttt{age} = [59, 98))$	1
42	$P(s \mid \texttt{male} = 1, \texttt{weight} = [33.4, 81.7), \texttt{height} = [172, 215), \texttt{age} = [59, 98))$	1,15

Tabla 5: Tiempos cuyas probabilidades condicionales de sobrevivir son menor a 0.5 del la primera red.

De la tabla 5 tenemos que es un poco más común que los hombres tengan menos probabilidad de sobrevivir comparado con las mujeres. Por otra parte, el peso de los individuos parece estar más equilibrado aunque los individuos con peso entre $[33.4, 81.7)$ son los que más predominan; además, para las estaturas en el intervalo $[123, 166)$, igualmente tienen menos probabilidad de sobrevivir. Por último, la edad parece solamente afectar a las personas mayores a 46 años. La combinación de estos factores nos indican que tenemos más probabilidad de sobrevivir al tiempo indicado a causa del hígado graso no alcohólico en ciertos periodos de tiempo, donde el tiempo que más se repite es el que corresponde al primer año. Esta información podría ser crucial para poder tomar decisiones a tiempo antes de complicaciones severas.
Por último, ajustaremos otra red donde incluiremos el tiempo de manera explícita en la red, es decir, tendremos un modelo con la siguiente arquitectura:

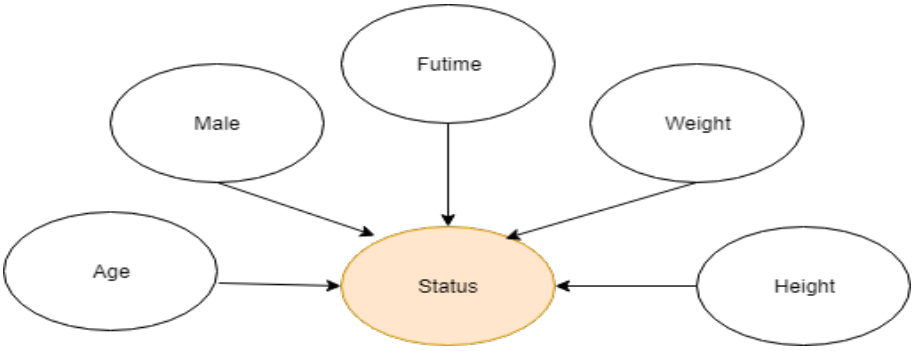


Figura 5: Arquitectura de la segunda red bayesiana.

Las gráficas de la aproximación de las probabilidades las presentamos en el anexo (figuras 9, 10 y 11), aunque nuevamente es complicado concluir con estas gráficas.
Por último, haremos un análisis análogo al que presentamos en la tabla 5 pero con este nuevo modelo. Los resultados se presentan a continuación:

No. Modelo	Modelo	Tiempo(s)
3	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	1
11	$P(s \mid \texttt{male} = 1, \texttt{weight} = [81.7, 86.2), \texttt{height} = [123, 166), \texttt{age} = [46, 59))$	5
12	$P(s \mid \texttt{male} = 1, \texttt{weight} = [81.7, 86.2), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	3,6,13
15	$P(s \mid \texttt{male} = 0, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	17
17	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [46, 59))$	5
18	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [123, 166), \texttt{age} = [59, 98))$	1,13
21	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
24	$P(s \mid \texttt{male} = 1, \texttt{weight} = [33.4, 81.7), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
36	$P(s \mid \texttt{male} = 1, \texttt{weight} = [86.2, 182), \texttt{height} = [166, 172), \texttt{age} = [59, 98))$	1
39	$P(s \mid \texttt{male} = 0, \texttt{weight} = [33.4, 81.7), \texttt{height} = [172, 215), \texttt{age} = [59, 98))$	1
42	$P(s \mid \texttt{male} = 1, \texttt{weight} = [33.4, 81.7), \texttt{height} = [172, 215), \texttt{age} = [59, 98))$	1,15

Tabla 6: Tiempos cuyas probabilidades condicionales de sobrevivir son menor a 0.5 de la segunda red.

Notamos que la tablas 6 y 5 son práctimanente iguales pues solo difieren por la estimación número 8, sin embargo, las probabilidades no son las mismas en todos los casos. Esto lo podemos ver si comparamos las gráficas que se encuentran en el anexo 7.

6. Conclusión

Como conclusión tenemos que el hecho de que nuestras redes realicen estimaciones durante periodos de tiempo es algo que resulta intuitivo de pensar. Si tomamos como ejemplo alguna enfermedad, tal que los primeros días son los menos probables para sobrevivir pero que conforme avance el tiempo esta probabilidad aumenta considerablemente, es algo que las redes bayesianas plasman de manera clara. Contrario a lo que ocurre con el modelo de riesgos proporcionales de Cox.

7. Anexo.

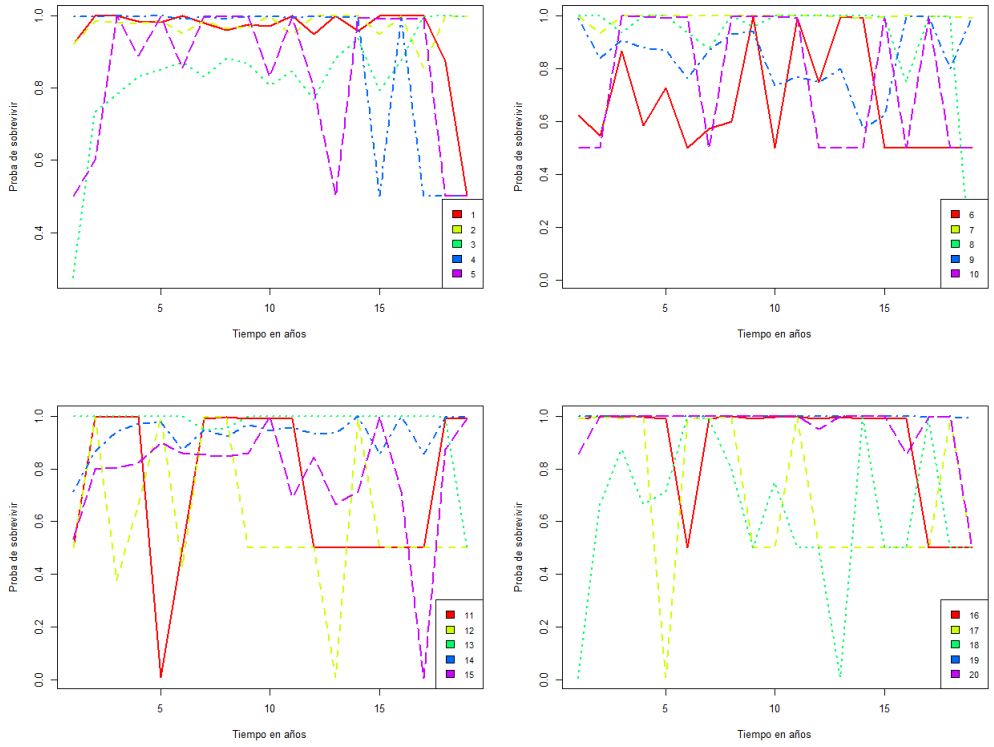


Figura 6: Primeras 20 estimaciones de las probabilidades de supervivencia del primer modelo.

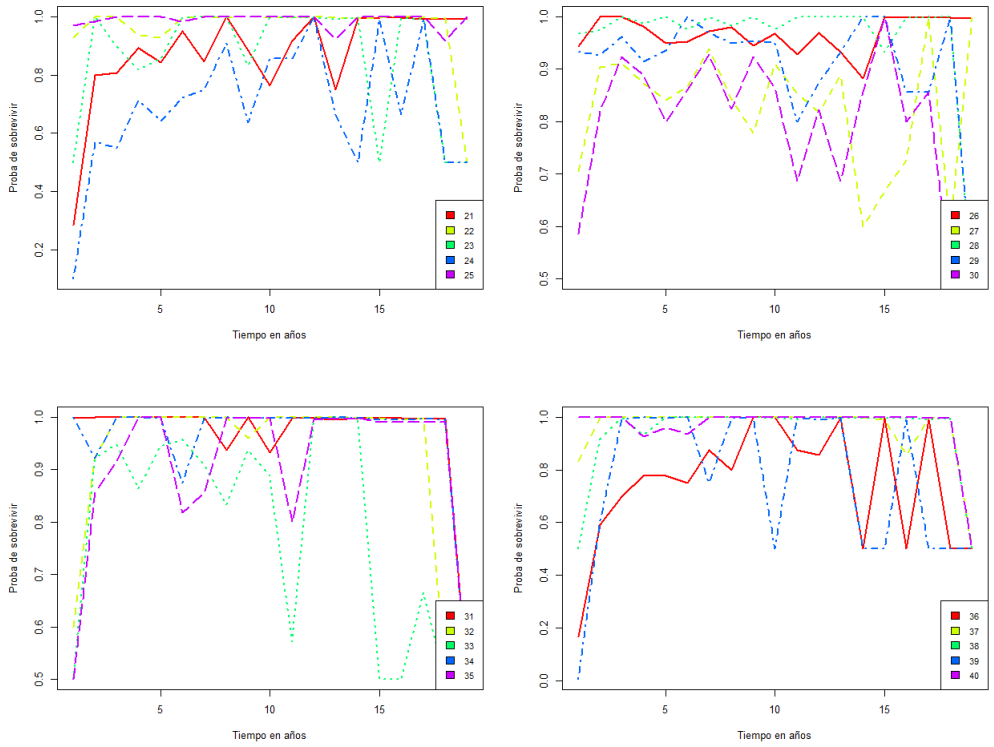


Figura 7: Estimaciones de las probabilidades de supervivencia (de la 21 a la 40) del primer modelo.

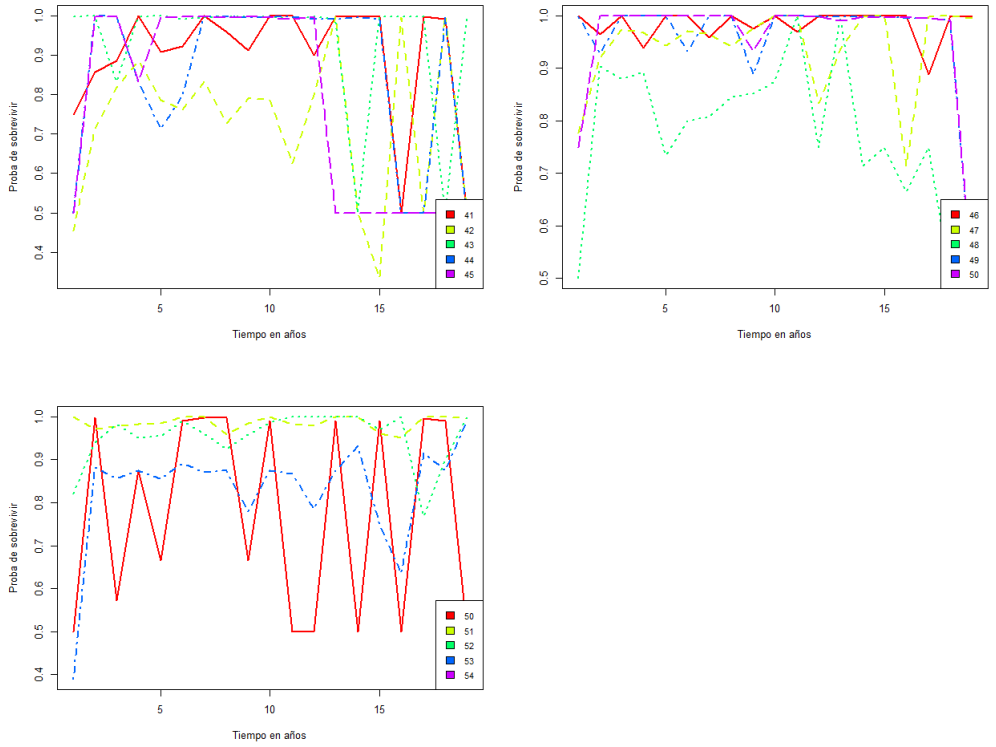


Figura 8: Estimaciones de las probabilidades de supervivencia (de la 41 a la 54) del primer modelo.

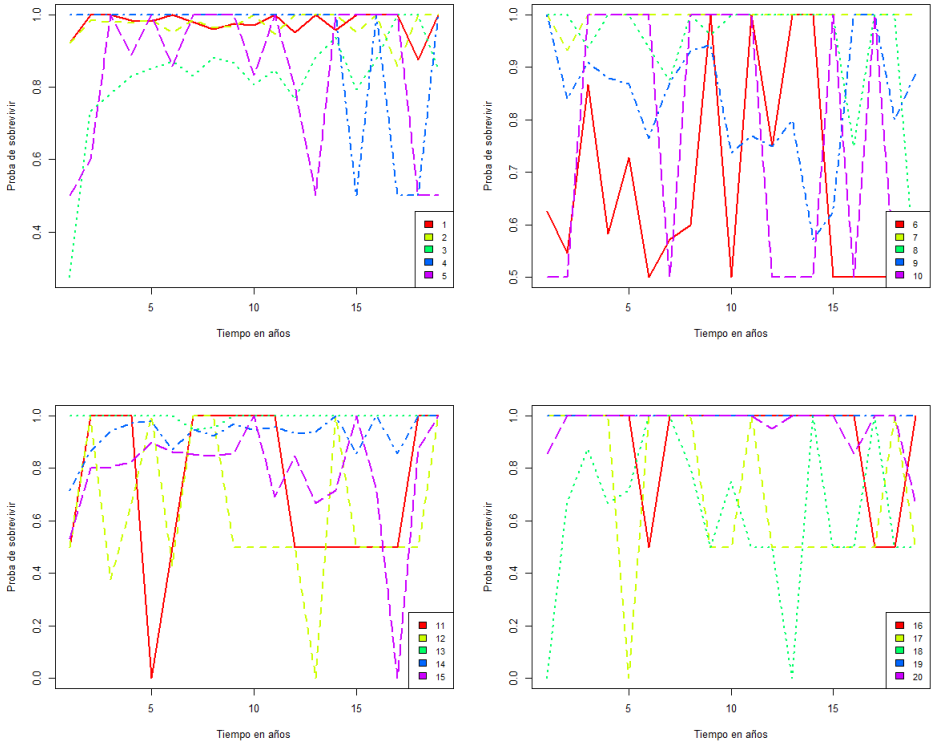


Figura 9: Primeras 20 estimaciones de las probabilidades de supervivencia del segundo modelo.

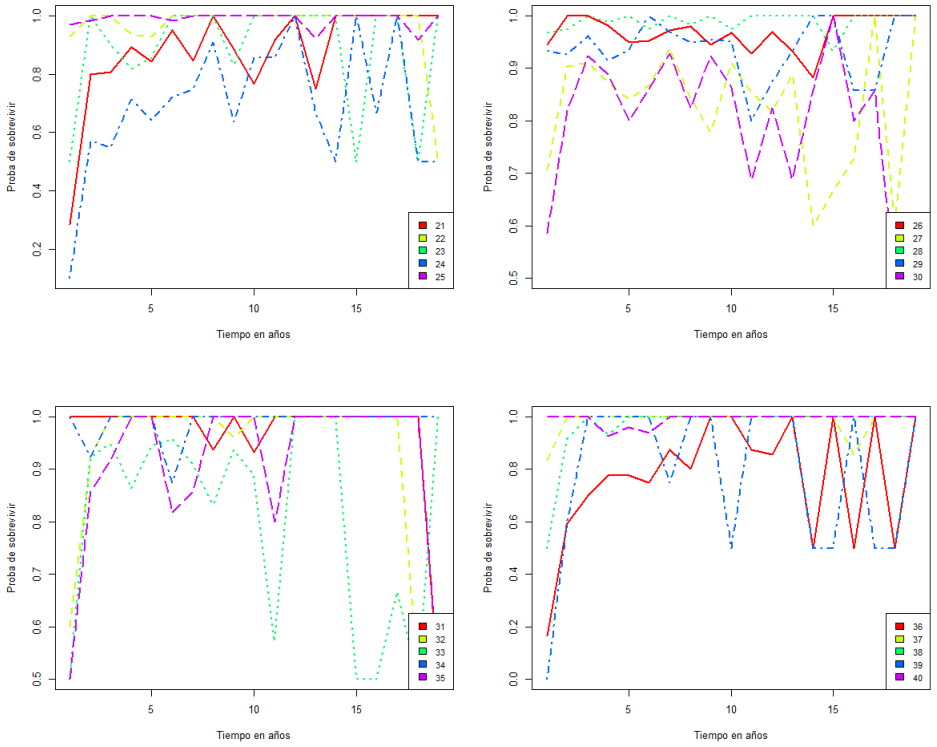


Figura 10: Estimaciones de las probabilidades de supervivencia (de la 21 a la 40) del segundo modelo.

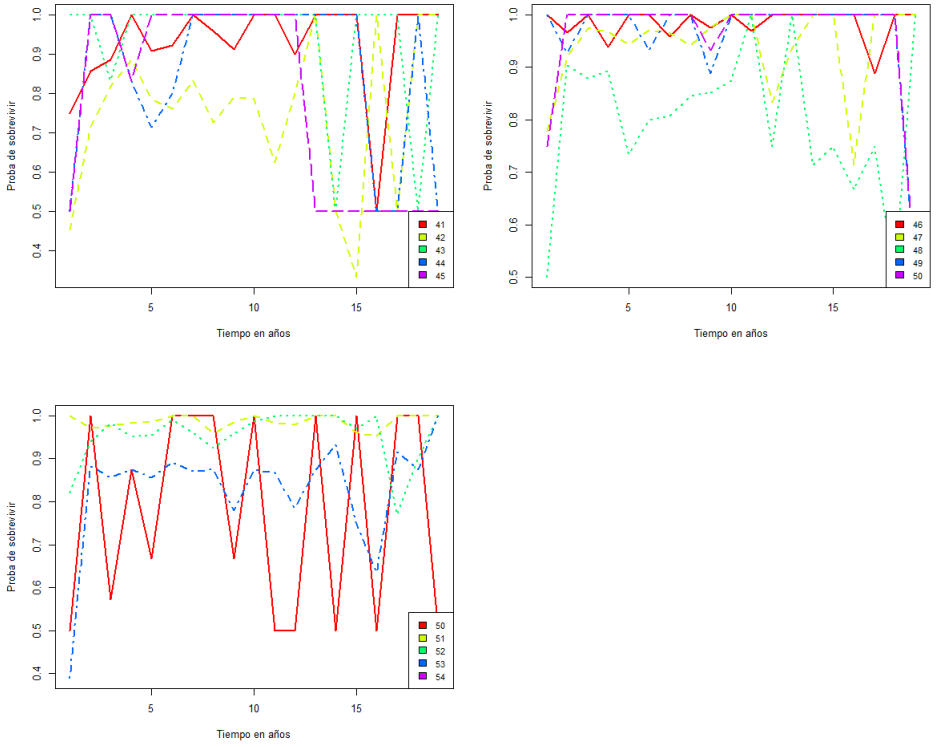


Figura 11: Estimaciones de las probabilidades de supervivencia (de la 41 a la 54) del segundo modelo.

Referencias

[1] Cox, D., & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC.

[2] Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis A Self-Learning Text* (3.a ed.). Springer.

[3] Kraisangka, J., & Druzdel, M. J. (2018, diciembre). A Bayesian Network Interpretation of the Cox's Proportional Hazard Model. ScienceDirect. https://doi.org/10.1016/j.ijar.2018.09.007

[4] Moore, D. F. (2016). *Applied Survival Analysis Using R* (1.a ed.). Springer.

[5] Nagarajan, R., Scutari, M., & Lébre, S. (2013). *Bayesian Networks in R with Applications in Systems Biology* (1.a ed.). Springer.