Análisis de supervivencia de fallas del corazón

Alfie González, Santiago Battezzati & Elena Villalobos

Maestría en Ciencia de datos
Instituto Tecnológico Autónomo de México

27 de Mayo, 2021.

1. Introducción

La idea del trabajo es que muestren que saben usar los modelos que vimos en clase de manera apropiada para resolver un problema práctico. Una vez teniendo la base de datos se van a formular una serie de objetivos a resolver y en el trabajo van a plasmar que fue lo que hicieron para resolver esos objetivos.

Descripción del probelam, contexto y objetivos a resolver

Qué son fallas del corazón, estadísticas mundiales y en México.

Las fallas del corazón son de las muertes más comunes

1.0.1. Objetivo

El objetivo del presente proyecto es evaluar lo

2. Base de datos

Este conjunto de datos contiene los registros médicos de 299 pacientes que tuvieron una falla en el corazón, que el término médico adecuado es disfunción sistólica ventricular. Estos datos se colectaron durante el periodo de seguimiento, y cada paciente tiene las siguientes 13 características clínicas[1].

- `age`: Edad del paciente (años).
- `sex`: Mujer u hombre (binaria).
- `anaemia`: Disminución de glóbulos rojos o hemoglobina (booleana).
- `diabetes`: Si el paciente tiene diabetes (booleana).

[1]Base de datos obtenidas del Machine Learning Repository: https://archive.ics.uci.edu/

- `smoking`: Tabaquismo, si el paciente fuma o no (booleana).
- `high-blood-pressure`: Si el paciente tiene hipertensión (booleana).
- `ejection-fraction`: Fracción de ejección, porcentaje de sangre que sale del corazón en cada contracción (porcentaje).
- `creatinine-phosphokinase`: Nivel de la enzima CPK en sangre (mcg/L).
- `platelets`: Plaquetas en la sangre (kiloplaquetas/ml).
- `serum-creatinine`: Nivel de creatinina sérica en sangre (mg/dl).
- `serum-sodium`: Nivel de sodio sérico en sanre (mEq/L).
- `time`: Periodo de seguimiento (días).
- `DEATH-EVENT`: Si el paciente falleció durante el período de seguimiento (booleana).

A continuación, se presenta un análisis exploratorio para observar el comportamiento general de las variables.

2.1. Análisis exploratorio

En la Figura 1, se observa de lado izquierdo un gráfico de barras para la variable de sexo, que nos muestra que tenemos casi el doble de hombres que mujeres. En el histograma de lado derecho, podemos las edades de todos los participantes, el rango de edad va de 40 a casi 100; observamos que tenemos más conteos de edades en los 50, también se observa una concentración de conteos en los 40, 45, 50, 60, 65 y 70.

La Figura 2, presenta también gráficos de barras de todas las variables booleanas que tenemos en el estudio que son anemia, diabetes, hipertensión y fumar. En todas, el 1 significa presencia y el cero ausencia. En la mayoría observamos más presencia de personas que se podrían considerar sanas en estas características pues no tiene ni anemia, ni diabetes, ni hipertensión. Así mismo, la variable que más diferencia tiene es la de tabaquismo, pues casi el doble de los sujetos no fuman, en comparación con los que sí fuman.

La Figura 3 es un gráfico que contiene los scatterplots de las variables continuas que tenemos, así mismo su contraparte muestra las correlaciones. Existen dos colores porque el rosa hace referencia a los hombres y el azul a las mujeres. Además, las variables de salida sangre y plaquetas están en logaritmo para poder apreciar mejor la relación con otras variables. El objetivo de mostrar este gráfico es para apreciar que no existen correlaciones claras entre las variables continuas, que se confirma con su contraparte con el estadístico de correlación, esto sucede para ambos sexos.

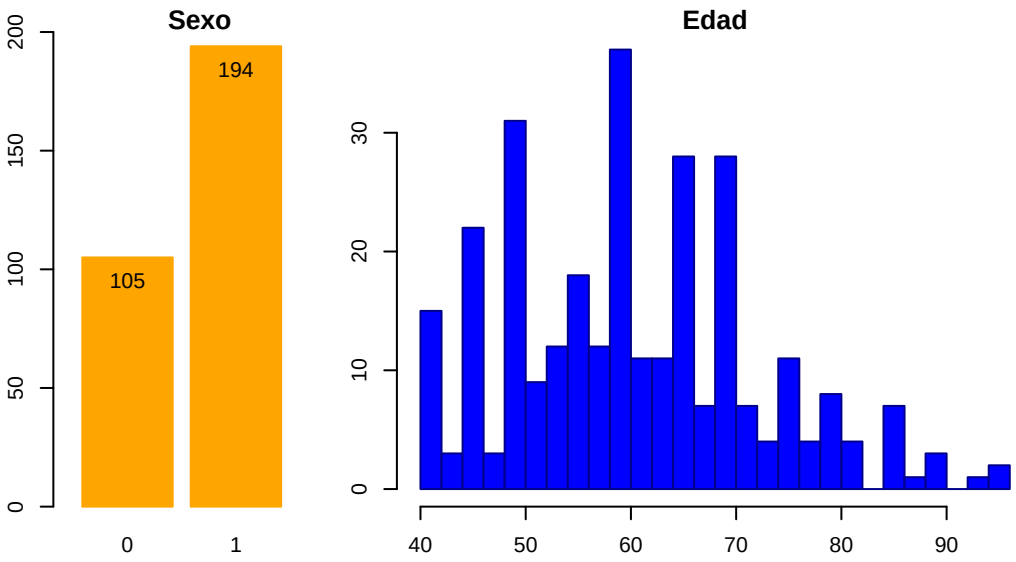

Figura 1: Gráficos de sexo y edad: De lado izquierdo es un gráfico de barras para sexo, done 1 significa hombre y 0 mujer. De lado derecho, un histograma con las edades de los participantes.

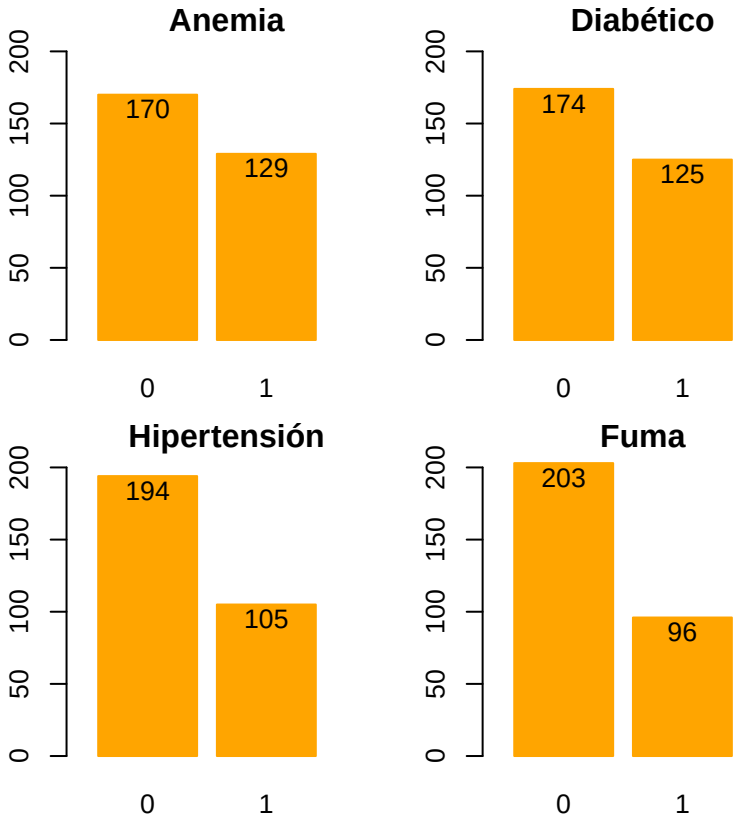

Figura 2: Graficos de barras para las variables de anemia, diabetes, hipertensión y tabaquismo.

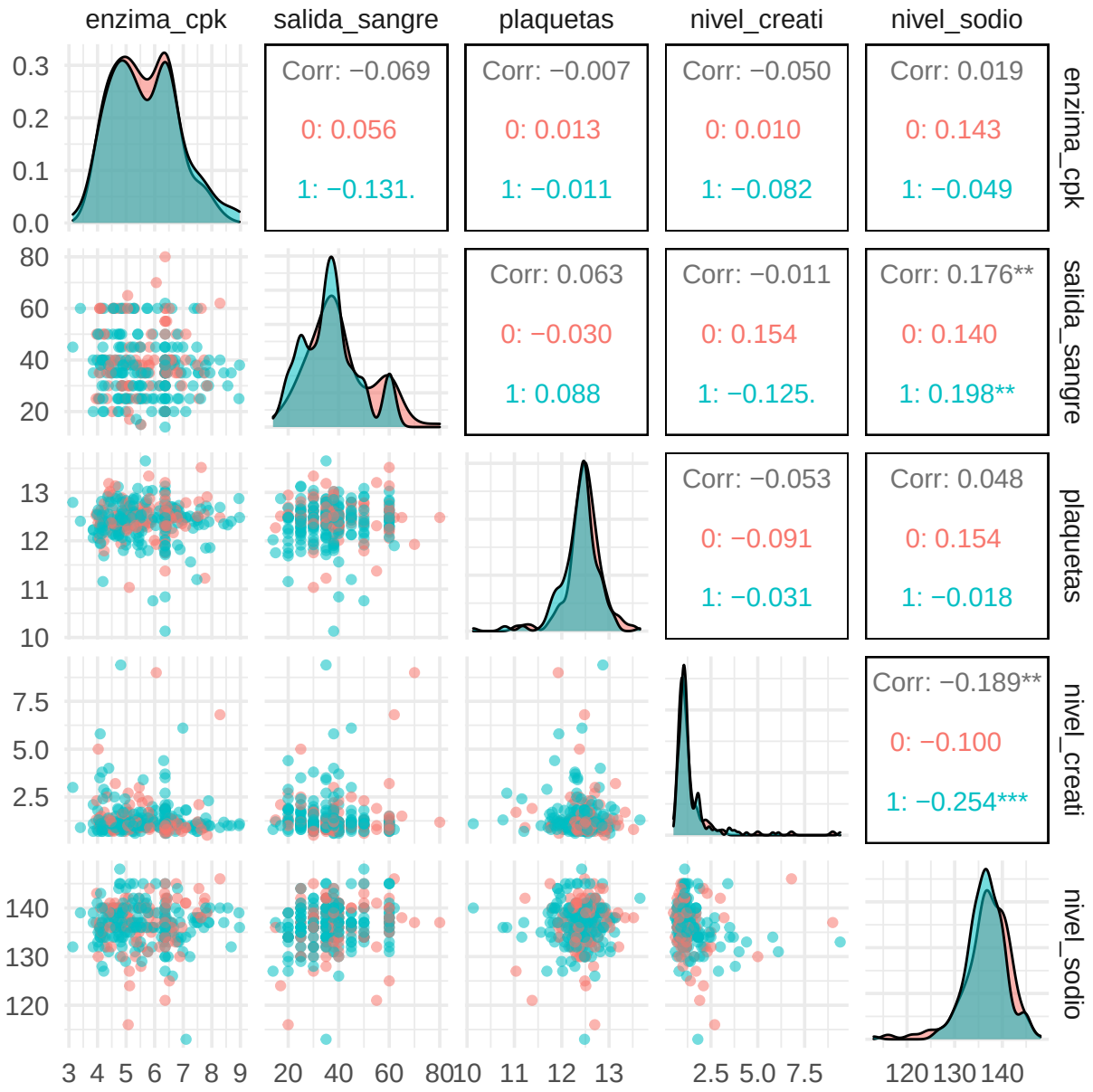

Figura 3: hola

2.2. Datos censurados

La variable evento muerte nos indica si el paciente falleció o no durante el periodo de seguimiento. En la presente base de datos, el 32 % de los pacientes fallecieron durante el periodo de seguimiento del estudio. En la Figura 4 podemos observar una línea que indica el periodo de seguimiento para cada paciente y si es un dato censurado o no, indicado por color. Lo importante de este gráfico es observar la mayoría de los pacientes que fallecieron durante el estudio, lo hicieron en el primer tercio del periodo, que equivale a 90 días aproximadamente. También, existe un conjunto de varios decesos presentados en el periodo de casi 180 días. Por último, parece ser que los pacientes que

tuvieron un periodo más largo de seguimiento fueron los que no presentaron el deceso.

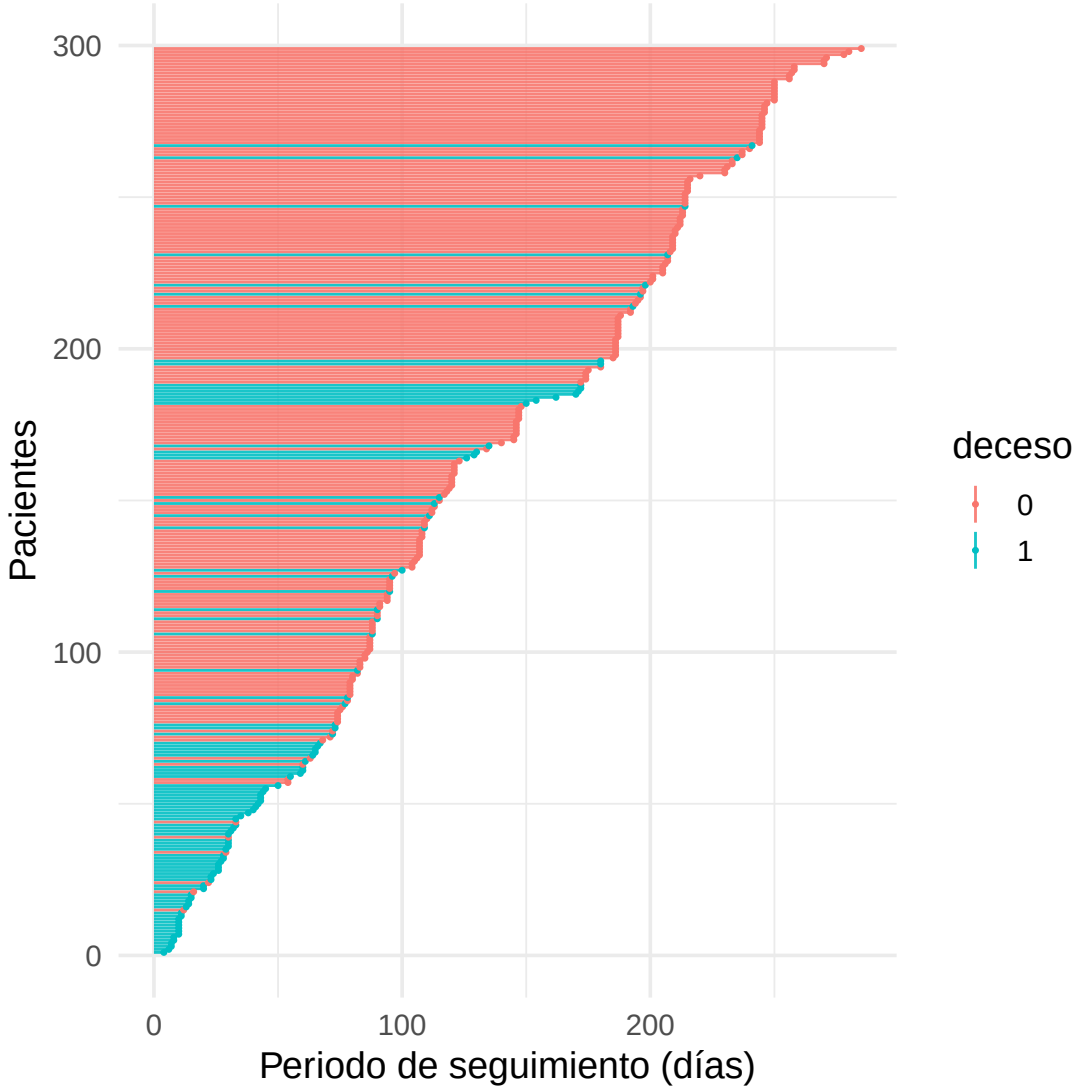

Figura 4: En el eje horizontal tenemos el tiempo de duración del estudio y en el eje vertical cada uno de los pacientes. Los pacientes están ordenados de acuerdo a los días del periodo de seguimiento. El color azul indica si fue un deceso y el rosa lo contrario.

3. Modelado e implementación

3.1. Verificación de supuestos

describan con detalle el modelo, con todas sus especificaciones, que usarán para resolver sus objetivos. Corran su modelo en R, Python o el paquete que mejor les acomode.

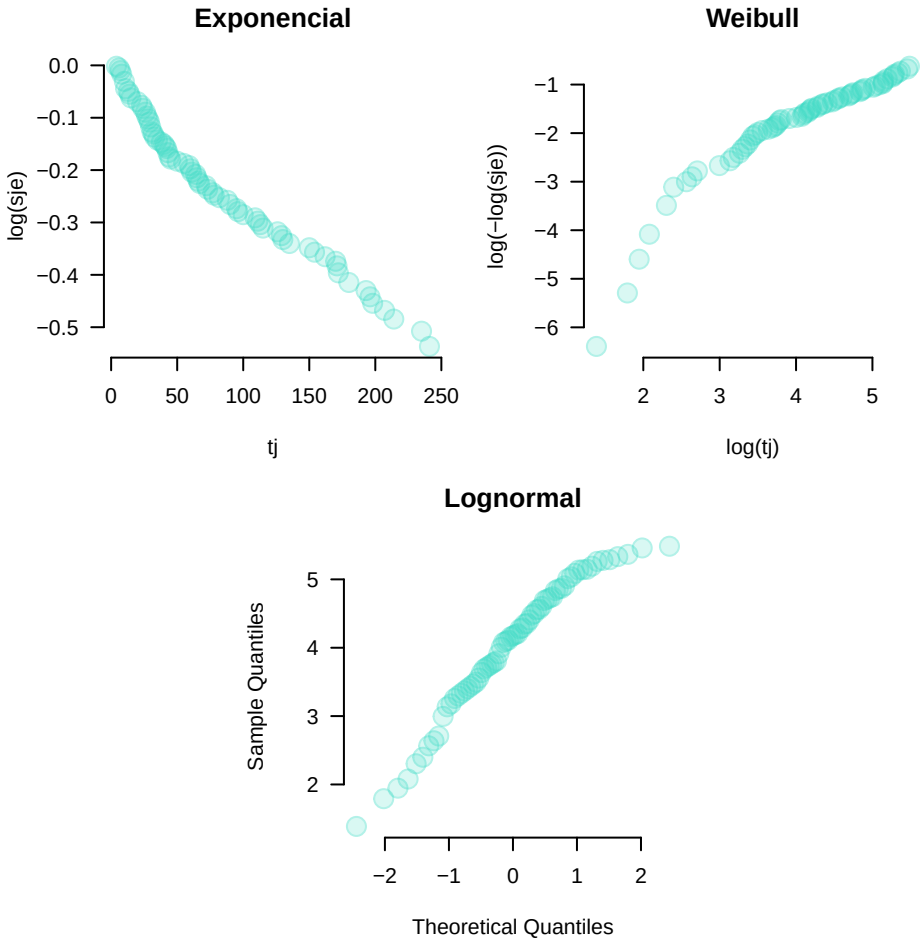

Figura 5: hola

4. Resultados

Interpretación de resultados: presenten un resumen de sus estimadores (puntuales y por intervalo) e interpreten en el contexto del problema. Planteen pruebas de hipótesis y tomen decisiones. Hagan uso de sus resultados para responder a los objetivos planteados e incluyan predicciones.

5. Discusión

Referencias

Incluyan una lista de las fuentes que consultaron para hacer su trabajo, desde páginas de internet, libros, revistas o apuntes de clase.

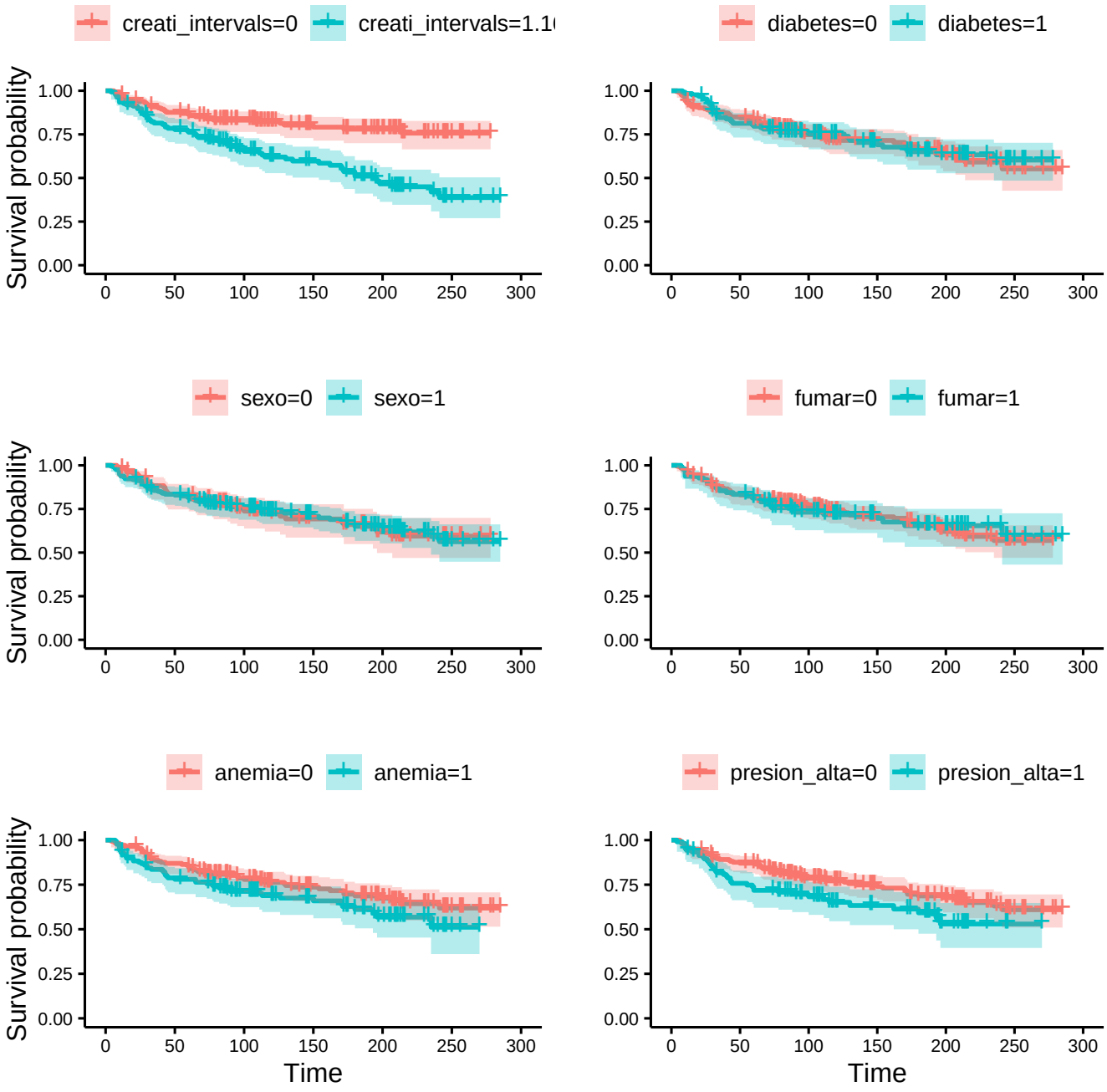

Figura 6: hola

Apéndice

Apéndice. Incluyan si quieren, todo el código utilizado. Por favor no incluyen código dentro de ninguna de las secciones anteriores. NOTA: Las gráfica que consideren útiles las pueden incluir en cualquiera de las secciones de la i-iv con comentarios para que el lector vea lo que ustedes quieren que vean. Las gráficas que no sean indispensables las pueden mandar al apéndice.

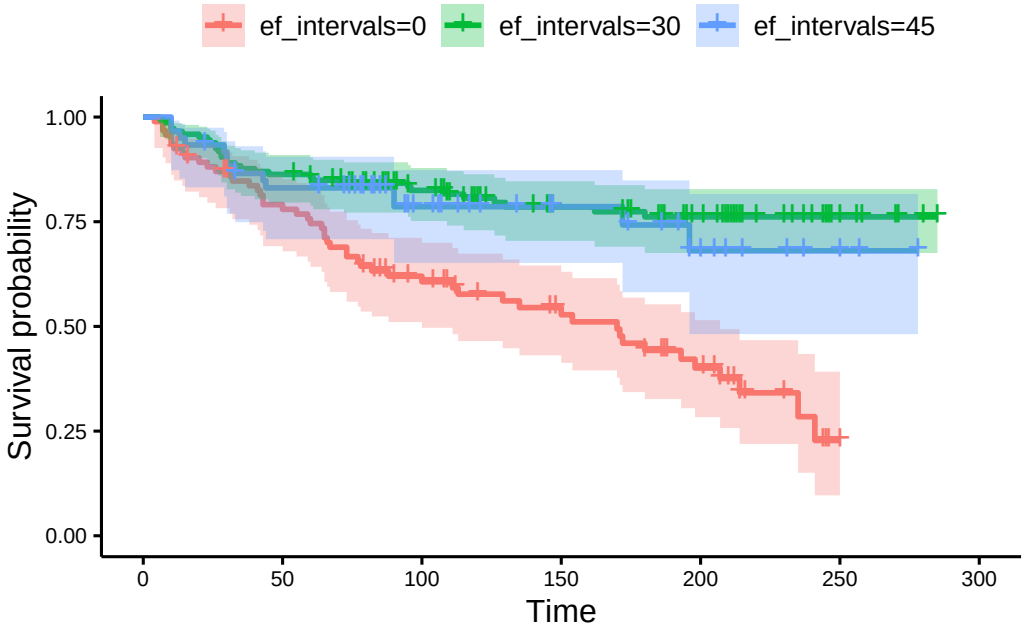

Figura 7: hola

4) Preparen una presentación de 15 minutos más o menos, el formato es libre. Todos los integrantes tienen que hablar y la calificación de la presentación será individual, mientras que la calificación del trabajo será por equipo. Se penalizará a aquellos equipos que se tarden más del tiempo asignado originalmente en su presentación.