Regresión logística multitarea para análisis de supervivencia: de los modelos tradicionales a los de aprendizaje automático.

Diego Vallarino Navarro
Máster en Bioinformática y Bioestadística
Área 2, subárea 2: Análisis de Datos

Directora: **Nuria Pérez Álvarez**
PRA: **Carles Ventura Royo**
Junio de 2022

Título del trabajo:	*Regresión logística multitarea para análisis de supervivencia: de los modelos tradicionales a los de aprendizaje automático.*
Nombre del autor:	*Diego Vallarino Navarro*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Carles Ventura Arroyo*
Fecha de entrega (mm/aaaa):	06/2022
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 2 – Subárea 2: Análisis de datos*
Idioma del trabajo:	Español
Número de créditos:	15
Palabras clave	*Modelos, Análisis de supervivencia, Aprendizaje Automático (ML)*

Resumen del Trabajo (máximo 250 palabras)

En el presente trabajo hemos utilizado una base de datos real, del paquete *survival*, para poder testear si existía una mejora en la performance en la utilización de diferentes modelos de supervivencia.
Luego de hacer una discusión conceptual sobre cuatro modelos, un modelo paramétrico, uno semi paramétrico, otro no paramétrico, y otro dentro de la categoría de *machine learning*, hemos evidenciado que los modelos tienen performances diferentes. Posiblemente la respuesta a esta mejora en la performance radique en la utilización de los datos censurados en forma diferente dentro del desarrollo de cada modelo, según se evidencia en la teoría analizada en el presente trabajo.
La hipótesis anterior la fundamentamos en el hecho que el modelo que mejor performance tiene, medido por el C-index, es el modelo de regresión logística multitarea (MTLR) el cual es esencialmente una colección de modelos de regresión logística construidos en diferentes intervalos de tiempo para determinar la probabilidad de que el evento de interés ocurriera durante cada intervalo. Los resultados proporcionados por el MTLR son similares al modelo CoxPH sin basarse en la suposición de CoxPH de que la función de peligro para los dos sujetos es constante en el tiempo. La mejora de performance del MTLR respecto al modelo Cox, el más cercano en performance, fue de aproximadamente un 6%.
Github: github.com/DiegoVallarino/MTLR-for-Suvirval-

Abstract (250 words or less):

In the present work we have used a real database, from the *survival* package, to be able to test if there was an improvement in performance in the use of different survival models.

After making a conceptual discussion about four models, a parametric model, a semi-parametric model, a non-parametric model, and another within the category of machine learning, we have shown that the models have different performances. Possibly the answer to this improvement in performance lies in the use of censored data differently within the development of each model, as evidenced in the theory analysed in this paper. We base the previous hypothesis on the fact that the model that has the best performance, measured by the C-index, is the multitask logistic regression (MTLR) model, which is essentially a collection of logistic regression models built at different time intervals. to determine the probability that the event of interest would occur during each interval. The results provided by the MTLR are similar to the CoxPH model without relying on the CoxPH assumption that the hazard function for the two subjects is constant over time. The performance improvement of the MTLR over the Cox model, the closest in performance, was approximately 6%.
Github: github.com/DiegoVallarino/MTLR-for-Suvirval-

Contenido

1. Resumen 6
2. Introducción 6
2.1 Contexto y justificación del Trabajo 6
2.1 Objetivos del Trabajo 8
2.1.1 Objetivo General 8
2.1.2 Objetivo Específicos 8
2.1.3 Enfoque y método seguido 9
2.1.4 Planificación del Trabajo 10
2.1.5 Breve sumario de productos obtenidos 13
2.1.6 Breve descripción de los otros capítulos de la memoria 13
3. Estado del arte 14
3.1 Marco Teórico 14
3.2 Modelos de Predicción de Supervivencia. 17
3.2.1 No paramétrico: Kaplan-Meier 19
3.2.2 Semi-paramétrico: Cox Regression 20
3.2.3 Paramétrico: Regresión Lineal, Modelo de Tobit 21
3.2.4 Machine Learning: Multi-Task Learning 22
4. Metodología 24
4.1 Métricas de evaluación: C-Index 26
5. Resultados 26
5.1. Análisis Descriptivo de los Datos 26
5.2 Existencia de Valores censurados por Derecha. 27
5.3 Análisis de data.train y data.test 27
5.4 Primeros análisis de supervivencia 27
5.5 Relevancia de las variables a utilizar usando XGBoost 27
5.6 Análisis de los diferentes modelos a probar 28
5.6.1 Modelo de Tobit – modelo paramétrico de supervivencia 28
5.6.2 Modelo Kaplan-Meier – modelo no paramétrico de supervivencia 29
5.6.3 Modelo Cox – modelo semi paramétrico de supervivencia 29
5.6.4 Modelo MTLR – Modelo de Machine Learning 29
5.6.5 Comparativa entre la performance de los modelos 30
6. Conclusiones y Discusión 32
7. Trabajo futuro 33
Glosario 35
Bibliografía 36
Anexo 39

1. Resumen

En el presente trabajo hemos utilizado una base de datos real, del paquete *survival*, para poder testear si existía una mejora en la performance en la utilización de diferentes modelos de supervivencia.

Luego de hacer una discusión conceptual sobre cuatro modelos, un modelo paramétrico, uno semi paramétrico, otro no paramétrico, y otro dentro de la categoría de *machine learning*, hemos evidenciado que los modelos tienen performances diferentes. Posiblemente la respuesta a esta mejora en la performance radique en la utilización de los datos censurados en forma diferente dentro del desarrollo de cada modelo, según se evidencia en la teoría analizada en el presente trabajo.

La hipótesis anterior la fundamentamos en el hecho que el modelo que mejor performance tiene, medido por el C-index, es el modelo de regresión logística multitarea (MTLR) el cual es esencialmente una colección de modelos de regresión logística construidos en diferentes intervalos de tiempo para determinar la probabilidad de que el evento de interés ocurriera durante cada intervalo. Los resultados proporcionados por el MTLR son similares al modelo CoxPH sin basarse en la suposición de CoxPH de que la función de peligro para los dos sujetos es constante en el tiempo. La mejora de performance del MTLR respecto al modelo Cox, el más cercano en performance, fue de aproximadamente un 6%.

2. Introducción

2.1 Contexto y justificación del Trabajo

Uno de los problemas determinantes en la toma de decisiones radica en predecir el tiempo antes de que ocurra un evento en particular. Por ejemplo, podemos estar interesados en la supervivencia de un paciente con cáncer, el tiempo antes de que se deba reemplazar una pieza de maquinaria o el tiempo antes de que un cliente interrumpa el comportamiento de otro cliente. El único aspecto del problema, que justifica el uso de métodos especializados, es la censura. Cierto tipo de problema de escasez de datos.

Supongamos que tenemos un conjunto de datos de pacientes con cáncer y estamos interesados en predecir el momento de la muerte relacionada con el cáncer. Para los pacientes que fallecieron en el momento del análisis, la solución sería simple: podíamos tomar su fecha de muerte (suponiendo un registro completo y preciso) y tratarla como un

problema de aprendizaje. Se trataría de un ejercicio supervisado, o más comúnmente conocido como "dicotómico, 0,1".

Una ventaja de este enfoque es la simplicidad: funciona casi instantáneamente con cualquier implementación de regresión logística, sin necesidad de implementar algoritmos personalizados. Sin embargo, un gran problema es que mirar la dicotomía puede ocultar información importante. Por ejemplo, considere dos pacientes con cáncer: uno que vive 5 años y el otro vive 10 años después del diagnóstico.

Obviamente, estos son casos completamente diferentes, pero desde el punto de vista del modelo binario, ambos pertenecen a la misma clase. Esto dificulta el aprendizaje, ya que se recopilan diferentes muestras en clase y las predicciones en sí son menos informativas. Además, para muchos problemas, es posible que no esté claro qué hora específica elegir. Lo ideal es no elegir nada y utilizar toda la información disponible. ¿Qué pasa con los pacientes cuya muerte aún no hemos notado? (muerte controlada en el lenguaje del análisis de supervivencia o valores NA, por ejemplo, porque todavía están vivos o se han mudado a otro país). La exclusión de estos casos puede reducir significativamente el tamaño de los datos de entrenamiento y puede sesgar el análisis.

Además, las observaciones censuradas aún contienen una información: sabemos con certeza que alguien que todavía estaba vivo después de un estudio de 5 años no murió antes de este punto.

La censura es un tema central en el análisis de la existencia, y muchos métodos y algoritmos están impulsados por esfuerzos para manejarla de manera objetiva e imparcial. Con respecto al análisis de supervivencia, en general, se debe considerar la cuestión de la moderación, es decir, se desconoce el momento real del evento para algunos pacientes.

La idea principal que se utiliza en muchos algoritmos de análisis de supervivencia es que los pacientes de control aún están parcialmente informados, es decir, no experimentan el evento hasta que conocemos su estado por última vez. En muchos casos, las NA representan un estado desconocido: podría ser 1 o 0, porque no sabemos el tiempo muerte real, simplemente no sucedió antes del momento en que hicimos la observación.

Bajo el supuesto de moderación independiente (no es más probable que los pacientes de control experimenten un evento que otros con covariables), todas estas secuencias son igualmente probables.

Considerando estas restricciones, es importante para el presente trabajo intentar analizar por medio de modelos de *tradicionales* el comportamiento de análisis de supervivencia

en base a variables 0 y 1, para luego levantar esta restricción y analizar críticamente si existe otra herramienta de *machine learning* que permita mejorar los modelos de predicción de supervivencia incorporando más información al modelo, manteniendo los mismos datos. Y entender la magnitud de esa mejora.

2.1 Objetivos del Trabajo

2.1.1 Objetivo General

El objetivo general del presente trabajo es el de comparar técnicas paramétricas, semi paramétricas y no paramétricas, con un modelo de *machine learning* para análisis de supervivencia. Específicamente el objetivo generar es poder tener una aproximación a la pregunta. Específicamente el objetivo generar es poder tener una aproximación a la pregunta

¿Se puede mejorar el entendimiento de la supervivencia en los pacientes de Cáncer de pulmón de la Administración de Veteranos contenidos en la base "veteran" utilizando técnicas de machine learning?

2.1.2 Objetivo Específicos

En relación con los objetivos específicos, los mismos los podemos enumerar de la siguiente forma:

1. Realizar un relevamiento sobre las principales modelos paramétricas, semi paramétricos y no paramétricos, con otro modelo de *machine learning (ML)* que pueden ser utilizados para un análisis de supervivencia, y seleccionar los que utilizarán en el presente trabajo.

   ¿qué modelos de ML se pueden ser utilizados para el análisis de supervivencia para el cáncer en el caso del Veterans' Administration Lung Cancer study?
2. Comprender conceptual y numéricamente las diferentes técnicas de *machine learning* que se van a comparar.

   ¿Cuáles son las diferencias entre las técnicas de ML que se relevaron en el punto anterior?
3. Realizar un análisis crítico respecto de las ventajas e inconvenientes de cada una de las técnicas de *machine learning* que se utilizarán.

   ¿Cuáles son las principales limitantes y beneficios que tienen las diferentes técnicas a utilizar?
4. Comprender claramente su funcionamiento práctico. Es decir, saber usar R o Python para obtener los resultados e interpretarlos.

¿Cuál es el código en R que permite el análisis de los modelos y la comparativa entre ellos sobre el análisis del cáncer en el caso de Veterans' Administration Lung Cancer study?

5. Realizar un estudio conceptual y práctico que permita compararlas de manera crítica.
6. Realizar una presentación sucinta de otros métodos que permiten realizar lo mismo, y hacer una pequeña discusión sobre la aplicabilidad de estos métodos al *dataset* de *Veterans' Administration Lung Cancer study*. Este análisis será el punto de partida para futuras investigaciones.
7. Realizar una interpretación de los resultados obtenidos en el área de aplicación clínica, o en este caso particular en el estudio del Cáncer.

   ¿Qué implicaciones tiene una mejora en el desempeño del análisis de supervivencia para la toma de decisiones clínicas?

2.1.3 Enfoque y método seguido

La metodología que se aplicará en el presente trabajo seguirá el siguiente orden:

1. **Análisis descriptivo**: como primer paso en cualquier análisis de datos, es importante conocer la información con la que se va a trabajar. En este trabajo de fin de maestría se trabajará con una base de datos real, que contiene información sobre pacientes de Cáncer de pulmón de la Administración de Veteranos. En tal sentido, se realizará un análisis descriptivo con sus respectivas evidencias.
2. **Análisis de supervivencia**: se aplicarán los métodos más tradicionales para estudiar el tiempo hasta el fracaso al tratamiento para los diferentes tipos de células: 1=squamous, 2=smallcell, 3=adeno, 4=large. Se usarán *Kaplan-Meier, Cox Regression y Linear Regression.*
3. **Análisis de Machine Learning**: se aplicará *Multi-Task Learning* para el análisis de supervivencia que permitirá contrastar los resultados del punto anterior, con un modelo de mejora en la utilización de la información que se puede extraer de los datos.
4. **Conclusiones y discusión**: una vez se haya llevado a cabo el trabajo de análisis de supervivencia y la comparación entre las diferentes técnicas de análisis, se procederá a extraer las correspondientes conclusiones, y la discusión de estas.
5. **Siguiente línea de investigación**: los modelos predictivos en los cuales se utiliza *machine learning* evolucionan en forma permanente, por lo que en esta sección se plantearan nuevas áreas de avance en el sentido del tema de este TFM, particularmente en el estudio de nuevos modelos que pueda mejorar la

performance, así como potenciales indicadores capaces de afinar la forma en que se compara la performance de los diferentes modelos, haciéndola lo más homogénea posible.

Dentro de los riesgos propios del TFM de estas características están los siguientes:

A. No haber estimado correctamente el Alcance del trabajo con el Tiempo disponible que se tiene para desarrollar el mismo.
B. Encontrar metodología relevante para el uso de herramientas *machine learning.* Si bien se puede desarrollar los modelos que se mencionan en este trabajo, posiblemente los resultados esperados no se concreten por un mal desarrollo del código o que los supuestos de cada uno de los modelos no se den en la prueba empírica con la base de datos a utilizar. Esto tendría un impacto directo en el tiempo necesario para desarrollar del presente trabajo. Para mitigar este riesgo se ha hecho una investigación secundaria en diferentes foros técnicos para poder entender la viabilidad *ex ante* de las herramientas y las limitantes de las mismas.
C. Algún tipo de pérdida o rotura de las herramientas con las cuales trabajares durante los próximos meses. En tal sentido, todo material desarrollado es "subido" a la nube en archivos compartidos para poder tener respaldo del trabajo realizado.

2.1.4 Planificación del Trabajo

Con respecto al plan de trabajo, el cual se adjunta en la siguiente figura, podemos identificar lo siguiente:

1. El relevamiento de los principales modelos de *machine learning* que pueden ser utilizados para el análisis de supervivencia implicarán las siguientes tareas:
    a. Relevamiento de los modelos tradicionales y de *machine learning* que son mayormente utilizados
    b. Relevar cuales son
    c. Identificar sus principales características
    d. Seleccionar algunos de estos modelos para avanzar en las siguientes etapas de este trabajo
2. Una vez se hayan seleccionado las diferentes herramientas de *machine learning* se realizará un conceptual y numérico las diferentes técnicas. Para cumplir con este objetivo, se realizarán las siguientes actividades:
    a. Entender cuáles son las principales características conceptuales
    b. Entender cuál es el funcionamiento que tienen estas técnicas desde la perspectiva del análisis de los datos numéricos.

3. Luego de comprender profundamente el funcionamiento de las diferentes herramientas, se realizará un análisis crítico respecto de las ventajas e inconvenientes de cada una de las técnicas de *machine learning* que se utilizarán.
    a. Se analizarán los beneficios y debilidades de las diferentes técnicas
    b. Se analizarán desde la perspectiva conceptual
    c. Se plantearán algunas hipótesis a ser analizadas en las siguientes etapas del presente trabajo, las cuales podrán ser demostradas a través del análisis de datos.
4. Una vez que haya realizado las etapas anteriores, se trabajará en comprender claramente su funcionamiento práctico. Es decir, saber usar R o Python para obtener los resultados e interpretar estos.
    a. Se hará una investigación de las mejores prácticas que existen
    b. Se analizarán modelos en R o en Python que ya fueron utilizados y se hará un análisis crítico.
    c. Con el conocimiento relevado de los pasos anteriores, se realizará un análisis especifico a nuestro *dataset*
5. Después de terminar con la etapa anterior, se realizará un estudio conceptual y práctico que permita compararlas de manera crítica.
    a. Se correrán los diferentes modelos en nuestro *dataset veteran*.
    b. Se seleccionarán los principales indicadores de desempeño de los modelos
    c. Se compararán esos indicadores de performance y se concluirá al respecto
6. Como actividad final se realizará una presentación sucinta de otros métodos que permiten realizar lo mismo, y hacer una pequeña discusión sobre la aplicabilidad de estos métodos al *dataset* con el que trabajaremos. Este análisis será el punto de partida para futuras investigaciones.

	PEC 1 - Plan de Trabajo		PEC 2 - Desarrollo Trabajo I				PEC 3 - Desarrollo Trabajo II				Memoria y Presentacion		Crear Presentacion	Defensa Publica
	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10	Semana 11	Semana 12	Semana 13	Semana 14
1. El relevamiento de los principales modelos de *machine learning* que pueden ser utilizados para el análisis de supervivencia implicarán las siguientes tareas:														
a. Relevamiento de los modelos tradicionales y de *machine learning* que son mayormente utilizados														
b. Relevar cuales son														
c. Identificar sus principales características														
d. Seleccionar algunos de estos modelos para avanzar en las siguientes etapas de este trabajo														
2. Una vez se hayan seleccionado las diferentes herramientas de *machine learning* se realizará un conceptual y numérico las diferentes técnicas. Para cumplir con este objetivo, se realizarán las siguientes actividades:														
a. Entender cuales son las principales características conceptuales														
b. Entender cual es el funcionamiento que tienen estas técnicas desde la perspectiva del análisis de los datos numéricos.														
3. Luego de comprender profundamente el funcionamiento de las diferentes herramientas, se realizará un análisis crítico respecto de las ventajas e inconvenientes de cada una de las técnicas de *machine learning* que se utilizarán.														
a. Se analizarán los beneficios y debilidades de las diferentes técnicas														
b. Se analizarán desde la perspectiva conceptual														
c. Se plantearán algunas hipótesis a ser analizadas en las siguientes etapas del presente trabajo, las cuales podrán ser demostradas a través del análisis de datos.														
4. Una vez que haya realizado las etapas anteriores, se trabajará en comprender claramente su funcionamiento práctico. Es decir, saber usar R o Python para obtener los resultados e interpretar estos.														
a. Se hará una investigación de las mejores prácticas que existen														
b. Se analizarán modelos en R o en Python que ya fueron utilizados y se hará un análisis crítico.														
c. Con el conocimiento relevado de los pasos anteriores, se realizará un análisis específico a nuestro *dataset*														
5. Después de terminar con la etapa anterior, se realizará un estudio conceptual y practico que permita compararlas de manera crítica.														
a. Se correrán los diferentes modelos en nuestro *dataset*														
b. Se seleccionarán los principales indicadores de desempeño de los modelos														
c. Se comparan esos idendicadores de performance y se concluirá al respecto														
6, Como actividad final se realizará una presentación sucinta de otros métodos que permiten realizar lo mismo, y hacer una pequeña discusión sobre la aplicabilidad de estos métodos al *dataset*														

2.1.5 Breve sumario de productos obtenidos

Siguiendo con los requerimientos planteados para la realización de un Trabajo Final de Maestría, al finalizar este trabajo se entregarán los siguientes documentos:

1. Memoria en formato PDF.
2. Fichero ".RData", que contiene los datos del estudio realizados.
3. Fichero ".Rmd" que contendrá el informe dinámico RMarkdown y el código utilizado en el software RStudio para la realización del estudio de los datos.
4. Presentación MS PowerPoint.
5. Video con la defensa de la memoria.

2.1.6 Breve descripción de los otros capítulos de la memoria

El presente Trabajo de Final de Maestría (TFM) presentará en el siguiente capitulo un pequeño relevamiento del estado del arte sobre la utilización de técnicas paramétricas, semi paramétricas, no paramétricas y de *machine learning* para el análisis de supervivencia. Específicamente se hará énfasis en el análisis de las herramientas con mayor frecuencia de uso, solo por mencionar algunas:

1. No paramétrico: *Kaplan-Meier*
2. Semi-paramétrico: *Cox Regression*
3. Paramétrico: *Linear Regression – Tobit model*

Para luego profundizar sobre la utilización del modelo de *machine learning*: *Multi-Task Learning*. En este análisis se tratará de mostrar la evidencia teórica y empírica que se ha desarrollado al respecto en los últimos 5 años.

Luego se hará una exposición de la metodología que se utilizará en el presente trabajo, en donde se definirán los distintos pasos metodológicos para alcanzar los objetivos previamente planteados en este trabajo.

Posteriormente se realizará el trabajo empírico en donde se analizarán los diferentes modelos sobre la misma base de datos *veteran*, y compararemos la *performance* de cada modelo en función de los indicadores utilizados dentro de las mejores prácticas para dichos análisis.

Por último, se realizarán ciertas conclusiones y discusiones en función de la evidencia encontrada.

3. Estado del arte

El análisis de supervivencia (análisis de tiempo hasta el evento) se usa ampliamente en economía y finanzas, ingeniería, medicina y muchas otras áreas. Un problema fundamental es comprender la relación entre las covariables y la (distribución de) los tiempos de supervivencia (tiempos hasta el evento).

Gran parte del trabajo realizado hasta la fecha ha abordado el problema considerando el tiempo de supervivencia como el primer tiempo de respuesta de un proceso estocástico, asumiendo una forma específica para el proceso estocástico subyacente, utilizando los datos disponibles para conocer la relación entre las covariables y los parámetros del modelo, y luego deducir la relación entre las covariables y la distribución de los primeros tiempos de acierto (el riesgo). Sin embargo, los modelos anteriores se basan en suposiciones paramétricas sólidas que a menudo se violan.

Con el objetivo de poder levantar estas limitantes, en los últimos años se ha comenzado a estudiar la eficiencia y eficacia de modelos de *machine learning* en el análisis de supervivencia.

3.1 Marco Teórico

A continuación, se presenta una sucinta revisión bibliográfica que, si bien no pretende ser exhaustiva en su cobertura, si se considera útil para presentar la situación conceptual en el marco del presente trabajo final de maestría. Por ende, en los próximos párrafos se expondrán diferentes trabajos sobre el análisis de supervivencia, y la utilización de diferentes modelos de *machine learning.*

En el trabajo de Zhang, Y (et al. 2022) se propone evaluar la *AppLife*, bajo el concepto de la predicción de supervivencia de aplicaciones de servicios móviles. Para ello desarrollan un marco que fusiona múltiples fuentes de factores de influencia y se utiliza el aprendizaje multitarea (MTL) para combinar la información de estado de las aplicaciones móviles (Apps) para la predicción de supervivencia.

Para ello los autores analizaron cómo la supervivencia de la aplicación se ve afectada por factores de múltiples fuentes, incluido el historial de descargas, las calificaciones y las reseñas. En segundo lugar, para superar acumulación de errores en la predicción a largo plazo, se propuso un nuevo enfoque basado en MTL.

El enfoque estima si una aplicación sobrevive en cada intervalo de tiempo durante el ciclo de vida de las aplicaciones y aprovecha la relación entre las tareas para mejorar la eficacia de la predicción. Por último, se recopila un conjunto de datos a gran escala con más de

35.000 aplicaciones, en función de las cuales evalúan el marco propuesto. Los resultados muestran que supera a los siete métodos de última generación con los cuales se comparó el desempeño del MTL.

En Zhang X. (et al. 2021) se propone una novedosa *multi-task network* (MTN) con características de varios niveles para la predicción del riesgo de pacientes con cáncer gástrico mediante la predicción simultánea de etapas clínicas. Los resultados indicaron que la red multitarea captura características multinivel compartiendo información de pronóstico de tareas correlacionadas de predicción de etapas clínicas, lo que permite a la MTN generalizar mejorar predicción de supervivencia y mejorar el rendimiento del modelo. La evidencia mostrada por Zhang X. (et al. 2021) permite evidencia la posibilidad de que la MTN sea una herramienta superior para la predicción del riesgo de cáncer gástrico en comparación con los métodos convencionales de aplicación clínica dado este modelo permite mejorar la predicción de supervivencia.

En Ping Jin (et al. 2021) propone en su trabajo de tesis un marco novedoso de aprendizaje de distribuciones del "precio de reserva" (RP), en el cual desarrolla un modelo de formulación de la relación entre los RP de los consumidores y sus decisiones de compra y un método de recopilación de datos.

Dentro de este marco, se muestra una forma de estimar la distribución de RP específica del consumidor utilizando técnicas de la predicción de supervivencia, viendo las opciones de compra de los consumidores como observaciones censuradas. Para validar el nuevo marco de RP, se realiza experimentos con datos realistas, con cuatro métodos de supervivencia.

Todos los modelos se desempeñaron muy bien (además del planteado por el autor se consideraron los modelos de *Kaplan-Meier Estimator, Cox Proportional Hazards Model, Accelerated Failure Time Model (Tobit Model)*) en la tarea de estimar las distribuciones de RP específicas del consumidor. Pero la conclusión apunta a que el modelo propuesto por Ping Jin (et al 2021) en relación con el RP evidencia una mayor eficiencia con respecto al resto de los modelos analizados.

En el estudio de Bingzhong J. (et al 2020), se propone utilizar el aprendizaje profundo para integrar la resonancia magnética multiparamétrica con etapas clínicas para predecir la ORS (overall risk score) y la estratificación del riesgo en pacientes con NPC (nasopharyngeal carcinoma). Los autores han utilizado el aprendizaje profundo para modelar el pronóstico de NPC en el conjunto de datos de resonancia magnética más

grande. Los resultados se probaron entre 429 pacientes que se incluyeron consecutivamente en un estudio de cohorte retrospectivo entre abril y noviembre de 2015 en la misma institución, lo que garantiza que es probable que el estudio sea reproducible en un entorno clínico real. Cuando se utilizaron resonancias magnéticas multiparamétricas con o sin etapas clínicas para establecer una red de supervivencia profunda multimodal, la *multi-modality deep survival network* (MDSN) propuesta logró resultados más precisos que los otros métodos de vanguardia. Los autores utilizaron el aprendizaje profundo para extraer una representación profunda de resonancias magnéticas multiparamétricas, que complementa el sistema de estadificación en la estratificación de riesgo para pacientes con NPC. Cuando en el trabajo se utilizaron la agrupación por etapas combinada con la etapa T, la etapa N y características profundas para establecer un modelo de pronóstico para NPC, el índice C de la agrupación por etapas aumentó de 0,610 a 0,672, lo que es una mejora significativa y demuestra que la MDSN tiene el potencial para ayudar a los médicos a tomar decisiones de tratamiento.

En su trabajo Yan Li (et al 2018) recopila conjuntos de datos de múltiples fuentes y exhiben un patrón de datos perdidos, es decir, cada paciente toma diferentes tipos de pruebas y recibe varios tratamientos, y cada prueba/tratamiento se asocia con un conjunto correspondiente de características.

Sin embargo, todos los métodos de análisis de supervivencia existentes están diseñados para conjuntos de datos totalmente observados y es posible que no se apliquen directamente cuando se presenten dichos datos faltantes por bloques.

El trabajo propuesto por los autores tiene como objetivo abordar los desafíos de investigación antes mencionados. Específicamente, han empleado un método de partición que descompone el bloque de fuentes múltiples datos faltantes en la submatriz múltiple completada; por lo tanto, transforma el problema original en una serie de problemas de análisis de supervivencia multifuente relacionados.

Para hacer frente a estos problemas, se propone un modelo de aprendizaje multitarea de dos capas que logra tanto el análisis a nivel de característica como a nivel de fuente, y además el modelo propuesto es capaz de aprovechar la información de la estructura en el patrón faltante por bloques. Los resultados experimentales utilizado en el documento sobre el Alzheimer permiten demostrar que el método propuesto supera a los otros métodos de última generación.

Como se ha demostrado en este sucinto relevamiento conceptual, en la literatura actual existe casi unanimidad en que los trabajos sobre estimación de la supervivencia mejoran en su desempeño utilizando modelos de *machine learning* algo más complejos que los modelos tradicionales. Por ende, es importante entender cuáles son las características de estos modelos, así como sus fortalezas y debilidades.

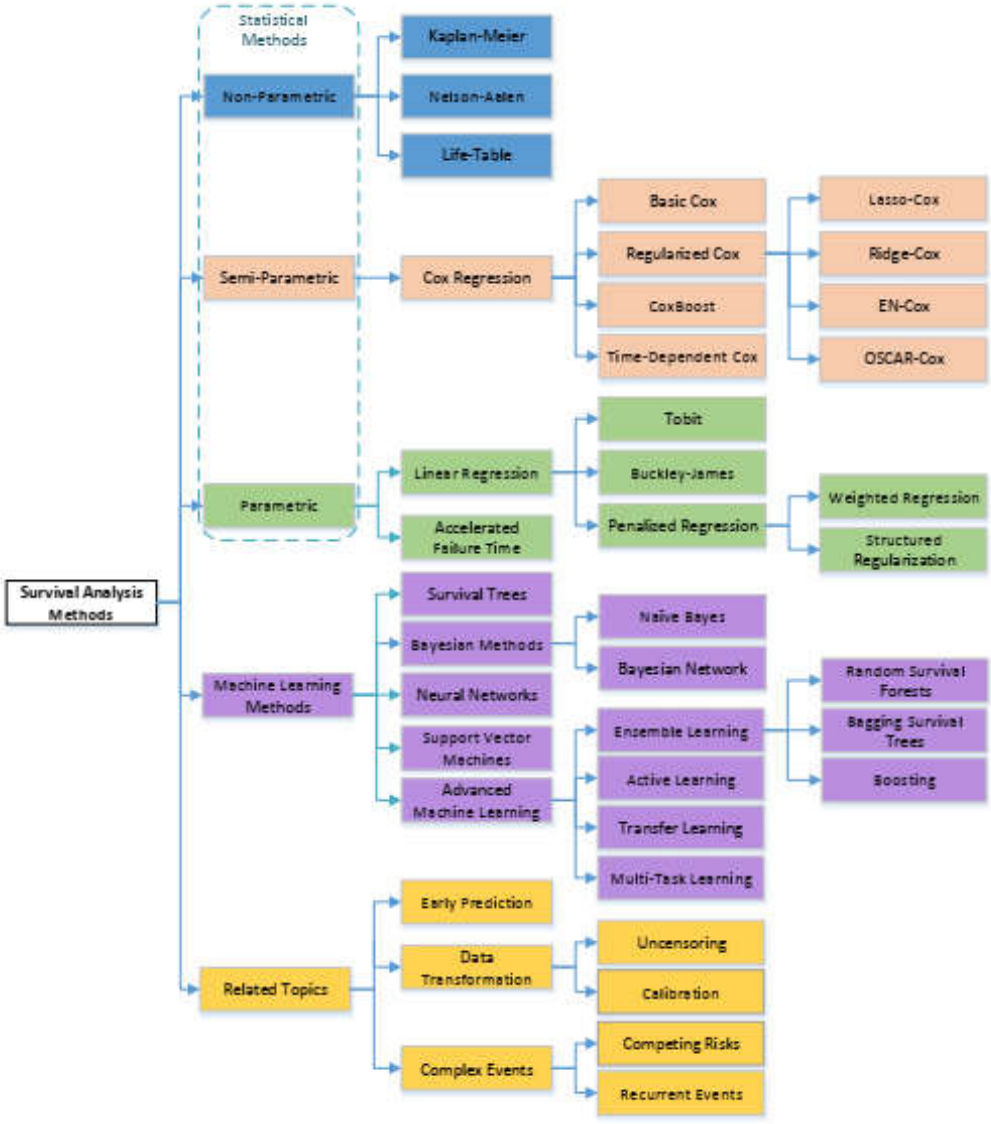

Figura 1: Ping Wang (et al. 2017) pág. 1:7

3.2 Modelos de Predicción de Supervivencia.

Los modelos de predicción de la supervivencia son diversos y con diferentes características. Ping Wang (et al 2017) se hace un análisis exhaustivo de los distintos modelos para el análisis de supervivencia.

A modo de resumen, en su trabajo se identifican diferentes métodos y análisis útiles para el análisis de la supervivencia. Como se mostró se muestra una tabla resumen (ver pag. 1:7)

En esta sección, describimos los conceptos fundamentales del análisis de supervivencia. Presentamos las notaciones formales y las métricas antes de introducir nuestra contribución, que modela el evento de supervivencia como una función continua del tiempo.

El análisis de supervivencia, a veces denominado análisis de tiempo hasta el evento, generalmente se define como un conjunto de métodos para analizar datos donde la variable de resultado es el tiempo hasta un evento de interés. En el análisis de supervivencia, los sujetos se observan típicamente durante un período determinado, y el énfasis está en el momento en que ocurre el evento de interés.

Un conjunto de datos típico utilizado para el análisis de supervivencia consta de tres elementos clave: los datos del paciente (atributos), la duración hasta el tiempo y un indicador de evento, es decir, si se ha producido un evento como se ilustra en Katzman et al. 2018. Si se detecta un evento (por ejemplo, la muerte), el período de tiempo se refiere al tiempo transcurrido entre el momento en que se obtuvieron los datos del paciente y el momento en que ocurrió el evento. Si no se detecta un evento, el intervalo de tiempo es el tiempo transcurrido entre la recopilación de datos basales y la última interacción del paciente (por ejemplo, el final del estudio). Para este último caso, los pacientes se conocen como $TE = 1\ o\ TE = 0$ *censurados por derecha*.

Por lo general, para usar modelos de regresión en dichos datos, se deben eliminar los datos censurados a la derecha, lo que imputa un sesgo en el modelo; por lo tanto, modelar datos censurados por la derecha requiere un modelo de supervivencia.

Suponiendo que es la función de densidad de probabilidad y es la función de probabilidad acumulativa, la probabilidad de un evento que ha ocurrido por el tiempo $f(t)F(t)t$ se define en la Ecuación 1.

$$F(t) = \int_0^t f(x)dx \qquad \textit{Eq. 1}$$

Como alternativa a la función, comúnmente estudiamos la $F(t)$ *función de supervivencia,* $S(t)$, que se define como la probabilidad de que el tiempo de un evento sea más

significativo que un tiempo especificado, *T*. La función de supervivencia proporciona la idoneidad de un evento a lo largo del tiempo y se puede definir como la Ecuación 2.

$$S(t) = F(T \geq t) = \int_t^{\infty} f(x)dx = 1 - F(t) \qquad \textit{Eq. 2}$$

Por otro lado, la *función de peligro,* generalmente denotada por $h(t)$ *o* $\lambda(t)$*,* es la posibilidad instantánea por unidad de tiempo para el evento, siempre que la persona haya sobrevivido hasta el tiempo. Es la probabilidad de fracaso en un período de tiempo infinitesimalmente corto entre y desde que el sujeto ha vivido hasta. La $ttt + \Delta tt$*función de peligro* $h(t)$ se define en la ecuación 3.

$$h(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt \mid T \geq t]}{dt} \; or \; h(t) = \frac{f(t)}{S(t)} \qquad \textit{Eq. 3}$$

La *función de peligro acumulativo* es el peligro acumulado a lo largo del tiempo y se puede expresar como la Ecuación 4.

$$\Lambda(t) = \int_0^t h(x)dx \qquad \textit{Eq. 4}$$

Existe una relación entre la supervivencia, el peligro y la función de peligro acumulativo descrita en la Ecuación 5.

$$S(t) = \exp\big(-\Lambda(t)\big) = \exp\left(-\int_0^t h(x)\,dx\right) \qquad \textit{Eq. 5}$$

Ahora, veamos algunos de los modelos clásicos de supervivencia y discutamos sus inconvenientes.

3.2.1 No paramétrico: Kaplan-Meier

Entre todas las funciones, la función de supervivencia o su presentación gráfica es la más utilizada. En 1958, Kaplan y Meier [Kaplan y Meier 1958] desarrollaron la curva de Kaplan-Meier (KM) o el estimador de límite de producto (PL) para estimar la función de supervivencia utilizando la duración real del tiempo observado. Este método es el más utilizado para estimar la función de supervivencia. Sea $T_1 < T_2 < \cdots < T_k$ un conjunto de tiempos de eventos distintos ordenados observados para instancias $N(K \leq N)$. Además de estos tiempos de eventos, también hay tiempos de censura para casos cuyos tiempos de eventos no se observan. Fuera tiempo de evento específico $T_j(j = 1,2, \dots, K)$, el número de eventos observados es *dj* ≥ 1, y las instancias de *rj* se considerarán "en riesgo" ya que su tiempo de evento o tiempo censurado es mayor o igual que *Tj*. Cabe

señalar que no podemos simplemente considerar *rj* como la diferencia entre $r_{(j-1)}$ y $d_{(j-1)}$ debido a la censura. La forma correcta de obtener r_j es $r_j = r_{(j-1)} - d_{(j-1)} - c_{(j-1)}$, donde $c_{(j-1)}$ es el número de instancias censuradas durante el período de tiempo entre $T_{(j-1)}$ y T_j. Entonces la probabilidad condicional de sobrevivir más allá del tiempo T_j se puede definir como:

$$p(T_j) = \frac{r_j - d_j}{r_j} \qquad \textit{Eq. 6}$$

Ventajas	**Desventajas**
Más eficiente cuando se conocen distribuciones teóricas inadecuadas.	Difícil de interpretar; arroja estimaciones inexactas.

3.2.2 Semi-paramétrico: Cox Regression

El Modelo de Riesgo Proporcional de Cox (CoxPH) es un modelo de regresión comúnmente utilizado para investigar la asociación entre el tiempo de supervivencia de un paciente y una o más variables predictoras. Es un enfoque semi paramétrico que se centra en el modelado de la función de peligro, que depende de la suma lineal de las características y el modelo de peligro de referencia. La ecuación de Cox muestra el modelo de regresión, que generalmente no se especifica, y es la función de riesgo expresada a través de una representación lineal tal que, donde es $\hat{h}_\beta(x)\vec{x}_i\, h_o(t)\boldsymbol{\eta}(\vec{x}_i)\boldsymbol{\eta}(\vec{x}_i) = \exp\left(\sum_{j=1}^{p} x_j^i \omega_j\right) x_j^{i\text{ith}}$ característica, son los coeficientes a determinar, y $\omega_j p$ es el conjunto de covariables.

$$h(t, \vec{x}_i) = h_o(t)\boldsymbol{\eta}(\vec{x}_i) \qquad \textit{Eq. 7}$$

La ecuación 7 muestra la probabilidad parcial de Cox, que está parametrizada por los valores $h(t)$ y $h_0\ (t)$ son el tiempo de evento respectivo, el indicador de evento $\omega\,.T_i E_i x_i i$ y los datos de referencia para la observación. El producto se define sobre el conjunto de pacientes con $E_i = 1$, es decir, evento que ocurrió antes de que finalice el período de estudio. El conjunto de riesgos es el conjunto de pacientes que aún están en riesgo de fracaso en el momento $\mathfrak{R}(t) = \{i: T_i \geq t\}t$. La probabilidad parcial logarítmica de la función debe minimizarse para obtener los resultados deseados $L_c(\omega)$

$$L_c(\omega) = \prod_{i:E_i=1} \frac{\exp(x^i \omega)}{\sum_{j \in \Re(T_i)} \exp(x^j \omega)} \qquad Eq.\ 8$$

El modelo no lineal de riesgos proporcionales de Cox fue introducido por Katzman et al. (2018) utilizando redes neuronales. La función de peligro para tales modelos se expresa en la ecuación 8, donde, = exp(y es una función de riesgo no lineal.$\boldsymbol{\Psi}(\vec{x}_i)\ \Psi(\vec{x}_i)), \Psi$

$$h(t, \vec{x}_i) = h_o(t)\boldsymbol{\Psi}(\vec{x}_i) \qquad Eq.\ 9$$

La combinación lineal de los parámetros de CoxPH lineal se reemplaza por, la salida de la red neuronal. La función de pérdida se expresa en la ecuación 9. $\Psi(\vec{x}_i)$

$$L_c = \prod_{i:E_i=1} \frac{\exp(\Psi(\vec{x}_i))}{\sum_{j \in \Re(T_i)} \exp(\Psi(\vec{x}_i))} \qquad Eq.\ 10$$

El modelo de red neuronal calcula los atributos no lineales y su combinación lineal se calcula para estimar la función de riesgo.

Aunque el modelo CoxPH determina las funciones de supervivencia y peligro en función del vector de características, posee las siguientes limitaciones:

- El registro de la razón de riesgo es impulsado por la combinación lineal de características de un individuo.
- Se basa en la suposición de peligro proporcional, que establece que la función de peligro de dos individuos debe ser constante en el tiempo.
- El modelo exacto no es computacionalmente eficiente y a menudo se aproxima utilizando Efron (1977) o Breslow (1974). La oscuridad del componente de tiempo de la función de peligro hace que el modelo CoxPH no sea adecuado para las predicciones de supervivencia.

Para superar los inconvenientes mencionados anteriormente, Yu (et al 2021) introdujo el modelo de Regresión Logística Multitarea (MTLR) para aproximar la función de supervivencia que analizaremos más adelante en este trabajo.

3.2.3 Paramétrico: Regresión Lineal, Modelo de Tobit

No podemos aplicarlo directamente para resolver problemas de análisis de supervivencia porque faltan los tiempos reales de los eventos para los individuos controlados. Varios modelos lineales [Miller et al. Halpern1982; Kuul et al. XIX] Buckley y James 1979; Wang et al. 2008; Li et al, 2016e], incluidas la regresión de Tobit y la regresión de Buckley-James, se han propuesto para tratar casos controlados en un análisis de supervivencia. Estrictamente hablando, la regresión lineal es una regresión controlada específica de los parámetros, sin embargo, este método es fundamental en el análisis de

datos y, por lo tanto, discutimos el método de regresión por separado. La linealidad de los datos controlados está aquí.

Regresión de Tobit: El modelo de Tobit [Tobit 1958] fue uno de los primeros intentos de extender la regresión lineal usando una distribución gaussiana para analizar datos usando observaciones controladas. En este modelo, se introduce una variable latente y∗ y la suposición hecha aquí es que depende linealmente de X a través del parámetro β como $y = X_\beta + \epsilon; \epsilon \sim N(0; \sigma_2)$ donde ε es un término de error normalmente distribuido. Entonces, para la i^{th} instancia, la variable observable y_i será $y_i *$ si $y_i * > 0$, de lo contrario será 0. Esto significa que, si la variable latente es mayor que 0, entonces la variable observada es igual a la variable latente y viceversa. Dependiendo de la variable latente, los parámetros del modelo pueden ser estimados por el método de estimación de probabilidad. Máxima probabilidad (MLE) con complejidad temporal $O(NP^2)$.

3.2.4 Machine Learning: Multi-Task Learning

El modelo de regresión logística multitarea (MTLR) es esencialmente una colección de modelos de regresión logística construidos en diferentes intervalos de tiempo para determinar la probabilidad de que el evento de interés ocurriera durante cada intervalo. Los resultados proporcionados por el MTLR son similares al modelo CoxPH sin basarse en la suposición de CoxPH de que la función de peligro para los dos sujetos es constante en el tiempo.

Los modelos de regresión censurados paramétricos sufren de debilidades aún más críticas. El rendimiento de predicción de la regresión censurada paramétrica depende en gran medida de la elección de la distribución. Sin embargo, en las aplicaciones del mundo real hay demasiadas interacciones y escenarios complejos que pueden afectar el evento de interés de varias maneras; por lo tanto, en la práctica, elegir una distribución teórica adecuada para aproximar los datos de supervivencia es muy difícil, si no imposible.

Para superar estas debilidades de ambos tipos de métodos, en este documento, proponemos el modelo MTLR. Formulamos el problema original de predicción del tiempo de supervivencia en un problema de aprendizaje de tareas múltiples. La principal motivación para usar el aprendizaje multitarea es su capacidad para aprender una representación compartida en tareas relacionadas y reducir el error de predicción de cada tarea. Por lo tanto, el modelo puede proporcionar una estimación más precisa de si ocurre o no un evento al comienzo de cada intervalo de tiempo, lo que proporcionará una

estimación precisa del tiempo de supervivencia para cada caso. Otra ventaja de usar el aprendizaje multitarea para la estimación del tiempo de supervivencia es que traduce el problema de regresión en una serie de problemas de clasificación binaria relacionados, y en cada intervalo de tiempo el clasificador correspondiente solo se enfoca en modelar el problema local y, por lo tanto, proporciona una estimación más precisa que los modelos de regresión que pretenden modelar todo el problema a la vez. Este modelo se construye sin ninguna hipótesis adicional excepto la hipótesis lineal, es decir, la característica y el objetivo exhiben una relación lineal, a diferencia del modelo de riesgos proporcionales de Cox y los modelos de regresión paramétrica censurada.

Como se desconoce el estado de supervivencia de una instancia censurada después del tiempo censurado correspondiente, la matriz de etiquetado de objetivos no está completa; por lo tanto, los métodos estándar de aprendizaje multitarea no pueden manejar las instancias censuradas. Para superar este problema, proponemos utilizar una matriz indicadora adicional que permite que el modelo aprenda simultáneamente de instancias censuradas y no censuradas y, por lo tanto, el modelo propuesto puede aprovechar simultáneamente tanto las instancias censuradas como las no censuradas. Dado que se alienta a múltiples predictores a compartir patrones de escasez similares, no solo seleccionará características importantes y aliviará el sobreajuste en espacios de características de alta dimensión, sino que también aprenderá una representación compartida en todas las tareas en diferentes momentos o intervalos de tiempo.

En el análisis de supervivencia, para cada instancia de datos, observamos un tiempo de supervivencia (Oi) o un tiempo censurado (Ci), pero no ambos. El conjunto de datos está censurado por la derecha si y solo si se puede observar $Ti = min\,(Oi, Ci)$ durante el estudio. Una instancia en los datos de supervivencia generalmente se representa mediante un triplete $(Xi, Ti, \delta i)$, donde Xi es un vector de características $1\ q;\ \delta i$ es el indicador de censura, es decir, $\delta i = 1$ para una instancia sin censura, y $\delta i = 0$ para una instancia censurada; y Ti denota el tiempo observado y es igual al tiempo de supervivencia Oi para instancias no censuradas y Ci de lo contrario, es decir,

$$T_i = \begin{cases} O_i & \text{if} \quad \delta_i = 1 \\ C_i & \text{if} \quad \delta_i = 0 \end{cases}$$

Para instancias censuradas, Oi es un valor latente, y el objetivo del análisis de supervivencia es modelar la relación entre Xi y Oi usando los tripletes $(Xi, Ti, \delta i)$ para instancias censuradas y no censuradas.

En la práctica, el tiempo se considera como intervalos de tiempo contables en lugar de un número real (un número con una fracción). Traducimos la etiqueta original a una matriz objetivo de k columnas Y, donde $k = max(Ti), i = 1, 2, , n$, es el tiempo máximo de seguimiento de todas las instancias. Cada elemento en la matriz de destino indica si el evento ocurrió ("0") o no y el problema original de predicción de supervivencia puede transformarse así en un problema de aprendizaje de tareas múltiples. La principal motivación para transformar el análisis de supervivencia en un problema de aprendizaje de tareas múltiples es que la dependencia entre los resultados en varios puntos de tiempo se captura con precisión a través de una representación compartida entre tareas relacionadas en esta transformación multitarea que reducirá el error de predicción en cada tarea.

4. Metodología

Para poder cumplir con los objetivos del presente trabajo, se propone el análisis de una base de (*veteran*). El conjunto de datos de cáncer de pulmón de la Administración de Veteranos muestra el tiempo de supervivencia de los pacientes masculinos con cáncer de pulmón inoperable avanzado. El estudio fue realizado por la Administración de Veteranos de EE. UU. Una vez más, este conjunto de datos no parece contener dependencias no lineales obvias.

La labor de predicción se divide en múltiples tareas, donde cada tarea tiene como objetivo estimar el estado de supervivencia de los pacientes. o de generalización de todas las tareas involucradas. Véase Li, Y., Yang, T., Zhou, J., & Ye, J. (2018) para profundizar en la metodología.

Las principales etapas metodologías del presente trabajo se pueden identificar de la siguiente forma:

1. *Entendimiento de los datos y formulación del problema*. La utilización de datos 0,1 para el análisis de supervivencia tiene ciertas limitaciones. La primera etapa de la metodología en poder identificar el conjunto de datos con los que se va a trabajar, entender las limitantes a nivel de disponibilidad de datos (NA, etc), poder comprender

las diferentes variables que están comprendidas en la base, y posibilidad del desarrollo de nuevos atributos que sirvan para poder mejorar el entendimiento y disposición de datos para los posteriores análisis.

2. *Desarrollo del análisis de supervivencia a través de cada modelo.* Una vez que la base de datos es analizada, preparada, y entendida, se procederá a desarrollar el análisis de supervivencia a través de diferentes modelos, los cuales toman en cuenta la información desde diferentes perspectivas. Para ello utilizaremos modelos no paramétricos, semi paramétricos, y paramétricos. Analizaremos el comportamiento/performance de cada uno de los modelos en términos relativos.

3. *Análisis empíricos y comparativa de los Modelos.* Una vez que se dispone de los análisis definidos en la etapa anterior, se procede a utilizar el modelo de *machine learning: Multi-Task Learning* para el análisis de supervivencia en los mismos pacientes con de *veteran* (Cao, H. *et al* 2019). Bajo los supuestos de este trabajo, se considera que la *performance* tendrá diferencias relevantes, se hará una comparación en base a los principales indicadores de desempeño de los modelos estadísticos.

4. *Conclusión y recomendación.* En función de los resultados, se redactarán diferentes conclusiones y recomendaciones para futuros trabajos teóricos y empíricos.

A modo de resumen, se presenta un esquema del proceso metodológico:

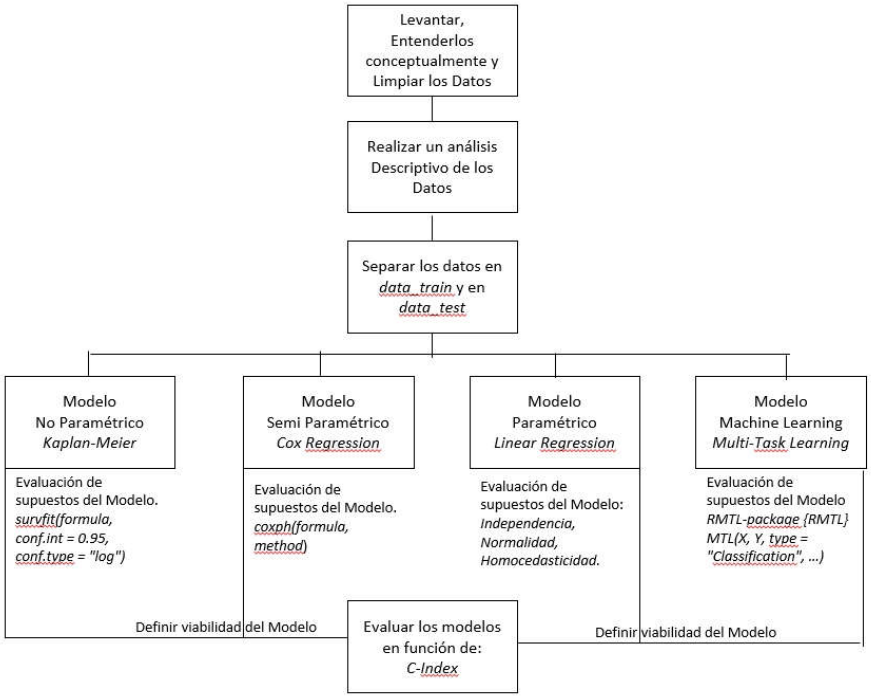

Figura 2: metodología de trabajo en base a elaboración propia

4.1 Métricas de evaluación: C-Index

El índice C de Harrell se expresa en la ecuación 11 donde,

- η_i es la puntuación de riesgo i del paciente
- $\mathbb{I}_{T_i>T_j} = \begin{cases} 1, & if\ T_i > T_j \\ 0, & otherwise \end{cases}$
- $I_{T_i>T_j} = \begin{cases} 1, & if\ T_i > T_j \\ 0, & otherwise \end{cases}$
- $\mathbb{I}_{\eta_i>\eta_j} = \begin{cases} 1, & if\ \eta_i > \eta_j \\ 0, & otherwise \end{cases}$

$$C - index = \frac{\sum_{i,j} \mathbb{I}_{T_i>T_j} \cdot \mathbb{I}_{\eta_i>\eta_j} \cdot \delta_j}{\sum_{i,j} \mathbb{I}_{T_i>T_j} \cdot \delta_j} \qquad \textit{Eq. 11}$$

Al igual que *el área bajo la curva*, el índice C = 1 describe la mejor predicción del modelo, y el índice C = 0.5 es tan igual como una predicción aleatoria.

5. Resultados

5.1. Análisis Descriptivo de los Datos

A continuación, se presenta un análisis descriptivo de la base de datos *suvirval* con la que trabajaremos en este TFM. La base tiene una estructura de 137 observaciones (pacientes) con 8 variables. Esto se muestra en la figura 1 del Anexo, mientras que en la figura 2 del Anexo se muestra la distribución de frecuencia de la variable *karno.*

Desde el punto del componente de las variables, las mismas son las siguientes:

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 10=yes

En la figura 3 del Anexo podemos analizar cómo es la distribución de la población en función del tipo de células. Mientras que en la figura 4 del Anexo se muestra la misma

descripción, pero para entender como es la estadística descriptiva de la población en referencia al tratamiento recibido por los pacientes.

5.2 Existencia de Valores censurados por Derecha.

Como hemos mencionado en el maraco teórico de este trabajo, la no disponibilidad de información completa implica varias limitantes. Una de ellas es no poder contar con toda la información que nos permiten analizar los datos. Otra es la limitante en la *performance* de los modelos, ya que muchos de los modelos, como el KM por ejemplo, tiene una mejor performance cuanto más grande es la muestra con la que trabajaremos.

En este punto es importante remarcar que existen diferentes tipos de datos NA. Uno es el que surge porque la información está mal imputada o porque no aparece en el análisis, algo que es común a cualquier análisis estadístico. Pero, en el caso del análisis de supervivencia el disponer de información NA puede ser sinónimo de tener algún tipo de censura por derecha. Es decir, no poder leer la información que posee el dato porque el modelo de análisis no lo permite realizar, o al menos lo toma en forma parcial.

Por ende, una de las etapas del análisis del trabajo es poder analizar la existencia de valores censurados. Estos se muestran en la siguiente figura 5 del Anexo.

A su vez, podemos confirmar que no existen valores NA de otro tipo dentro de la base de datos.

5.3 Análisis de data.train y data.test

Para poder realizar una contrastación de *performance* entre modelos era crítico que se realizara una división en el *dataset*. Se tomó el 70% de la base para entrenar los modelos y un 30% para usarlo como set de testeo.

5.4 Primeros análisis de supervivencia

Para comenzar con el estudio se realizan algunos primeros análisis de supervivencia a través del análisis *surdiff*. Ésta es una prueba que implica conocer si hay una diferencia entre dos o más curvas de supervivencia utilizando la familia de pruebas G-rho, o para una sola curva frente a una alternativa conocida. Tal cual se evidencia en la figura 6 del Anexo.

5.5 Relevancia de las variables a utilizar usando XGBoost

Para poder analizar qué variables tienen mayor poder de explicación en los modelos, se utilizó el modelo de XGBoost para poder tener mayor comprensión de los datos.

XGBoost Extreme Gradient Boosting es un algoritmo predictivo supervisado que utiliza el principio de impulso. La idea del boosting es generar secuencialmente múltiples

modelos de predicción "débiles", y cada uno de éstos toma los resultados del modelo anterior, para generar un modelo "más fuerte", con mejor poder predictivo y mayor estabilidad en sus resultados. Para obtener un modelo más fuerte, se utiliza un algoritmo de optimización, en este caso Gradient Descent. Durante el entrenamiento, los parámetros de cada modelo débil se ajustan iterativamente tratando de encontrar el mínimo de una función objetivo, que puede ser la proporción de error en la clasificación, el área bajo la curva (AUC), el error cuadrático medio (RMSE) o algún otro.

Se puede observar, en la figura 3, que las variables que más aportan información en la clasificación son las siguientes (por orden de importancia): *time, diagtime, age, trt, karmo.*

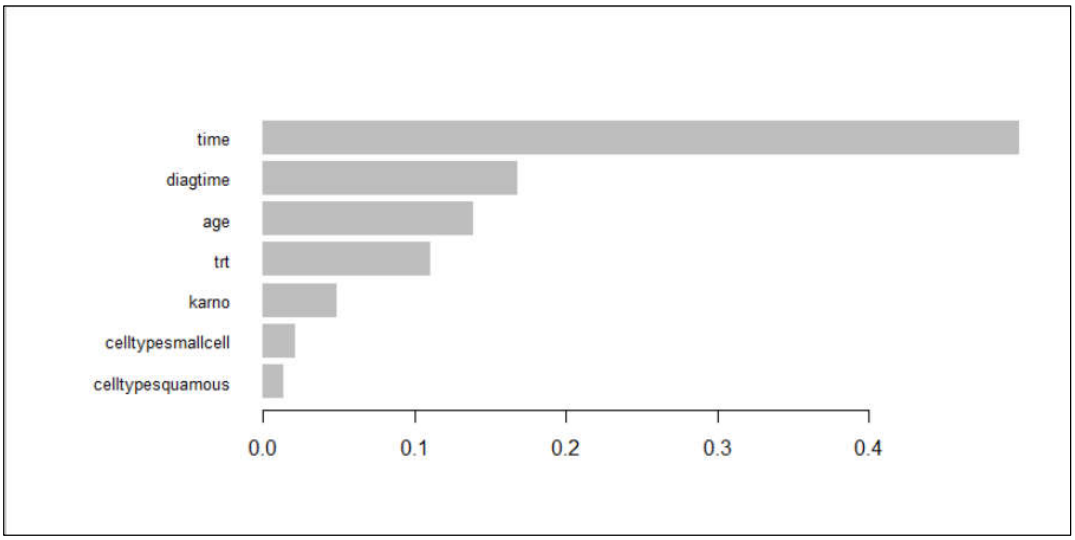

Figura 3: elaboración propia utilizando XGBoosting

5.6 Análisis de los diferentes modelos a probar

A continuación, se mostrarán los principales procedimientos sobre la utilización de los diferentes modelos que se han querido testear dentro del presente trabajo.

5.6.1 Modelo de Tobit – modelo paramétrico de supervivencia

El modelo ha seguido la siguiente estructura:

survreg(formula = Surv(time, status) ~ karno + age + trt, data = data.train)

Como se puede evidenciar te utilizaron las variables relevantes según el estudio del apartado anterior. Se utilizó la función *survreg* del paquete *survival* para entrenar el modelo, y luego poder estimar el desempeño de este.

Con los siguientes resultados a nivel general:

Chisq= 41.4 on 3 degrees of freedom, p= 5.4e-09

5.6.2 Modelo Kaplan-Meier – modelo no paramétrico de supervivencia

Para el caso del modelo de KM se utilizaron todas las variables, no haciendo una selección *ex ante,* de las mismas. En este caso el motivo fue poder analizar todo el conjunto de datos bajo la metodología de KM.

El modelo ha seguido la siguiente estructura:

fit3<-survfit(Surv(time, status) ~ 1, data = data.train)

Los resultados a nivel general se muestran en la figura 7 del Anexo, en donde se muestra como la probabilidad de supervivencia tiene una tendencia a la baja a medida que avanza el tiempo.

5.6.3 Modelo Cox – modelo semi paramétrico de supervivencia

Al igual que el modelo de apartado anterior, y considerando que la evidencia estadística muestra que el modelo KM, como el modelo Cox, son los que tienen mayor frecuencia de uso en el estudio de supervivencia, se tomó la decisión de tomar todas las variables como explicativas.

El modelo ha seguido la siguiente estructura:

fit4 <- coxph(Surv(time, status) ~ ., data=data.train, x = TRUE)

Los resultados del modelo de Cox se presentan en la figura 8, 9 y 10 del Anexo. En los análisis referencias en el Anexo, algunos de los datos relevantes son que la regresión de Cox cumple con la Normalidad de los Residuos, necesaria para poder considerar a la regresión como un modelo viable. A su vez, podemos evidenciar que, al usar todas las variables, no solo las variables con mayor importante como se mencionaron en la sección 5.5, se encuentran algunas variables que tienen algún tipo de significación en el modelo de Cox como por ejemplo *celltype.*

5.6.4 Modelo MTLR – Modelo de Machine Learning

El problema original de predicción de la supervivencia puede transformarse así en un problema de aprendizaje multitarea. La principal motivación para transformar el análisis de supervivencia en un problema de aprendizaje de múltiples tareas es que la dependencia entre los resultados en varios momentos se captura con precisión a través de una representación compartida entre tareas relacionadas en esta transformación de múltiples tareas que reducirá el error de predicción en cada tarea.

El modelo ha seguido la siguiente estructura:

fit7 <- mtlr(Surv(time, status)~., data = data.train, nintervals = 9)

Los resultados a nivel general del modelo MTLR se presentan en la figura 11 del Anexo.

5.6.5 Comparativa entre la performance de los modelos

A continuación, se presenta una gráfica de resumen en donde se muestra como el modelo de *machine learning* tiene una mejor performance cuando mediamos esta por medio del índice de *Concordance*.[1]

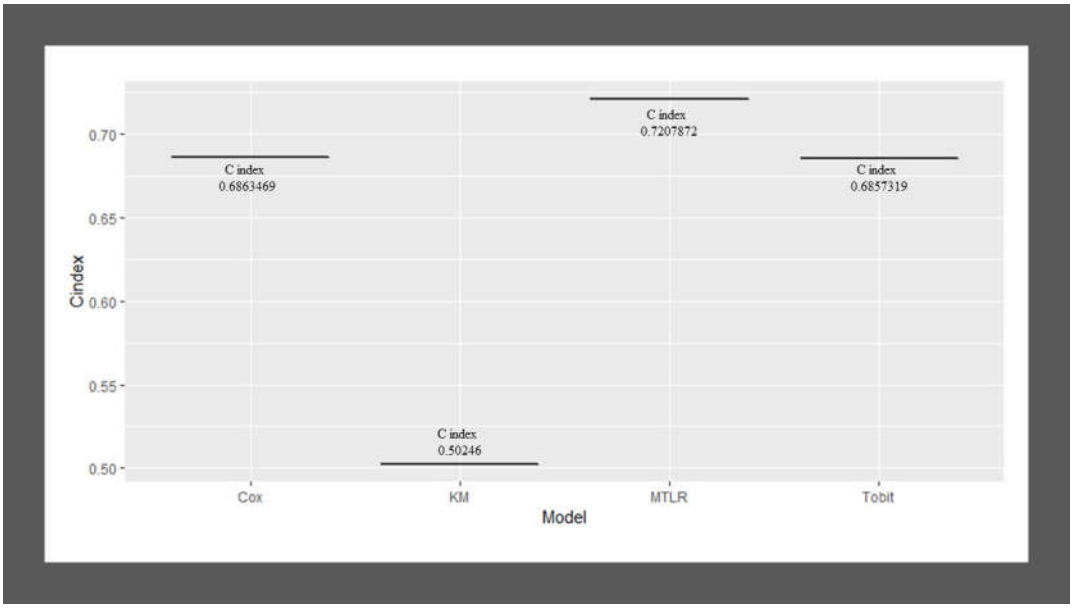

Figura 4: elaboración propia sobre comparativa de modelos (*boxplot)*

Para el proceso de comparación se entrenaron los modelos con al *data.train* y luego se utilizó la función *predict* para poder tener un entendimiento claro de cómo era la performance del modelo.

A continuación, se presenta cómo se hizo para cada modelo:

Tobit Model

predictfit2<-predict(fit2, data.test)

metrics_fit2<-Cindex(surv_obj, predicted = predictfit2)

KM Model

predictfit3<-predictSurvProb(fit3, data.test, dis_timefit3)

metrics_fit3 = Cindex(surv_obj, predicted = predictfit3[, med_indexfit3])

Cox Model

predictfit4<-predictSurvProb(fit4, data.test, dis_timefit4)

metrics_fit4 = Cindex(surv_obj, predicted = predictfit4[, med_indexfit4])

[1] https://cran.r-project.org/web/packages/SurvMetrics/vignettes/SurvMetrics-vignette.html

MTLR Model

```
predictfit7<-predict(fit7, data.test, type = "mean_time")
metrics_fit7 = Cindex(surv_obj, predicted = predictfit7)
```

6. Conclusiones y Discusión

Las técnicas de modelado de análisis de supervivencia más comunes son el modelo Kaplan-Meier (KM) y el modelo de riesgos proporcionales de Cox (CoxPH). El modelo KM proporciona una manera muy fácil de calcular la función de supervivencia de una cohorte completa, pero no lo hace para un individuo específico.

Por su parte, el enfoque del modelo CoxPH permite al usuario final predecir las funciones de supervivencia y riesgo de un individuo en función de su vector de características, pero presenta las siguientes limitaciones:

- Supone que la función de riesgo, o más precisamente el logaritmo de la relación de riesgo, está potenciada por una combinación lineal de características de un individuo.
- Se basa en la suposición de riesgo proporcional, que especifica que la función de riesgo de dos individuos tiene que ser constante en el tiempo.
- La fórmula exacta del modelo que puede manejar empates no es computacionalmente eficiente y es a menudo se reescribe utilizando aproximaciones.
- El hecho de que el componente de tiempo de la función de riesgo (es decir, la función de línea de base) permanece no especificado hace que el modelo CoxPH no sea adecuado para las predicciones de la función de supervivencia real.

Para superar estos inconvenientes, Yu (et al 2021) introdujo el modelo MTLR que puede calcular la función de supervivencia sin ninguna de las suposiciones o aproximaciones antes mencionadas.

En este trabajo nos hemos preguntado si *se puede mejorar el entendimiento de la supervivencia en los pacientes de Cáncer de pulmón de la Administración de Veteranos contenidos en la base "veteran" utilizando técnicas de machine learning.*

La evidencia que hemos encontrado a lo largo de este trabajo muestra como la mejora en la *peformance* en la utilización del MTLR tiene un salto significativo, de aproximadamente un 6% si lo comparamos con el modelo CoxPH, el cual es el que se desempeñó en forma más próxima.

Por ende, se podría concluir que la información que se deja de lado en los modelos KM o CoxPH es relevante para poder tener un mejor entendimiento en los análisis de supervivencia, y los modelos de *machine learning* permitirían avanzar en los estudios sobre estos temas. Obviamente estas conclusiones conceptuales y empíricas, para el caso

de este trabajo de TFM son aplicables en principio para la base de datos que hemos utilizado, con sus propias limitantes y características.

Esto permitiría a los tomadores de decisiones sobre posibles tratamientos del cáncer, entre otras tantas enfermedades, pudieran mejorar su toma de decisiones, para diferentes pacientes en diferentes momentos de un análisis. Es decir, con los mismos datos, con un modelo de *machine learning* como el que hemos presentado teórica y empíricamente, podemos asegurar que a nivel clínico las medidas y tratamientos sobre el cáncer (y otras patologías) pueden ser mejor analizadas y por ende mejor tratadas, dado que la toma de decisiones de los profesionales se hace con mejor información, más precisas.

Esto tiene suma relevancia en cómo entender mejor la supervivencia y el riesgo esperados. Como hemos analizado, evidenciado y demostrado, la censura es un tema central en el análisis de la supervivencia, y en este trabajo hemos visto cómo puede mejorarse.

En la próxima sección presentamos algunas líneas de investigación futura que ayudará a levantar estas limitantes, así como ampliar el conocimiento sobre el análisis de supervivencia en las diferentes áreas en las cuales se pueden aplicar.

7. Trabajo futuro

Si bien en este trabajo de fin de maestría solamente se evidenció la mejora que se puede lograr con la utilización de modelos de *machine learning* sobre una determinada población, creemos que es importante seguir investigando y profundizando el análisis como lo planteamos a continuación.

En próximas tareas sería recomendable poder desarrollar 3 líneas de investigación diferentes y complementarias.

La primera de ellas poder trabajar sobre una base de datos con más observaciones, tal que permita que algunos de los modelos que tuvieron desempeños débiles puedan mejorar su *performance*. Incluso poder analizar otro tipo de limitantes cuando se utilizan base de datos con mayores niveles de NA o de comportamiento diferente de las covariables.

Una segunda instancia es poder seguir sumando modelos al proceso de análisis. Es sabido que, aunque el modelo MTLR proporciona resultados similares a los del modelo CoxPH sin tener que depender de los supuestos requeridos por este último, en su esencia sigue siendo impulsado por una transformación lineal. Por lo tanto, ambos modelos no logran capturar elementos no lineales de los datos y, en consecuencia, dejan de ofrecer

rendimientos satisfactorios. Por ende, seria relevante poder estudiar la regresión logística multitarea neuronal (N-MTLR) la cual, según algunos documentos analizados, ayudaría a resolver este problema.

Por último, poder mejorar los indicadores de comparación entre modelos dado los avances que existe actualmente gracias a la nueva disponibilidad de datos, y la complejidad de los modelos que se están desarrollando en base a análisis de *machine learning.* Esta es toda una línea de investigación en la cual el poder realizarnos nuevas preguntas, dado que disponemos de "nuevas herramientas" de análisis, permitiría mejorar los diagnósticos, y el desarrollo de conocimiento, sobre el análisis de supervivencia en pacientes con diferentes patologías.

Glosario

Sigla	Definicion
ADN	Ácido desoxirribonucleico
AFT	Tiempo de Fallo Acelerado
ANN	Artificial Neural Network (Redes Neuronales Artificiales)
ARN	Ácido ribonucleico
BN	Bayesian Network (Red Bayesiana)
CDC	Centros para el Control y Prevención de Enfermedades
C-Index	Confidence Index
DTG	Dolutegravir
ELV	Elvitegravir
ETS	Enfermedades de Transmisión Sexual
IBS	Integrated Brier Score
INI	Inhibidor de la Integrasa
IP	Inhibidor de la Proteasa
ITIAN	Inhibidor transcriptasa inversa análogo de nucleósido
ITINN	Inhibidor transcriptasa inversa no nucleósido
KGB	Comité para la Seguridad del Estado
MI	Mutaciones en la Integrasa
ML	Machine Learning
MP	Mutaciones en la Proteasa
MTI	Mutaciones en la Transcriptasa Inversa
MTLR	Multitask Logistic Regression
MTN	multi-task network
NAIVE	Se denomina así a aquéllos que no han tenido tratamiento antirretroviral previo.
NB	Naïve Bayes
NIPS	Neural Information Processing Systems (Sistemas de procesamiento de información neural)
NPC	nasopharyngeal carcinoma
ORS	overall risk score
PH	Proportional Hazards (Riesgos Proporcionales)
RNN	Recurrent Neural Network (Redes Neuronales Recurrentes)
RP	Reserve Price
RVM	Relevance Vector Machine
SIDA	Síndrome de Inmunodeficiencia Adquirida
SSVM	Survival Support Vector Machine
SVM	Support Vector Machine
VIH	Virus de Inmunodeficiencia Humana

Bibliografía

1. Bansal, A., & Heagerty, P. J. (2019). A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and prognostic research*, *3*(1), 1-13.
2. Bewick, V., Cheek, L., & Ball, J. (2004). Revisión de estadísticas 12: Revisión de análisis de supervivencia. *Biomed Central Ltd. Cuidados críticos*, 8(5), 389.
3. Breslow, N. (1974). Análisis de covarianza de datos de supervivencia censurados. *Biometría,* 30(1), 89–99.
4. Cao, H., Zhou, J., & Schwarz, E. (2019). RMTL: an R library for multi-task learning. *Bioinformatics*, *35*(10), 1797-1798.
5. Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., & Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in biology and medicine*, *134*, 104481.
6. Chai, H., Zhang, Z., Wang, Y., & Yang, Y. (2021). Predicting bladder cancer prognosis by integrating multi-omics data through a transfer learning-based Cox proportional hazards network. *CCF Transactions on High Performance Computing*, *3*(3), 311-319.
7. Chen, L., Shao, K., Long, X., & Wang, L. (2020). Multi-task regression learning for survival analysis via prior information guided transductive matrix completion. *Frontiers of Computer Science*, *14*(5), 1-14.
8. David R Cox. Modelos de regresión y tablas de vida. *En Avances en* estadísticas. Springer, 1992.
9. Efron, B. (1977). La eficiencia de la función de probabilidad de Cox para datos censurados*. Revista de la Asociación Americana de Estadística*, 72(359), 557–565.
10. Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*.
11. Gerds, T. A., & Schumacher, M. (2006). Estimación consistente de la puntuación de Brier esperada en modelos de supervivencia general con tiempos de eventos censurados a la derecha. *Revista biométrica. Biometrische Zeitschrift*, 48(6), 1029–1040.
12. Guo, B., Zhang, Y., Liu, J., Guo, T., Ouyang, Y., & Yu, Z. (2020). Which App Is Going to Die? A Framework for App Survival Prediction with Multi-Task Learning. *IEEE Transactions on Mobile Computing*.
13. Harrell, F. E., Jr, Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, *15*(4), 361–387.
14. Jin, P., Haider, H., Greiner, R., Wei, S., & Häubl, G. (2021). Using survival prediction techniques to learn consumer-specific reservation price distributions. *Plos one*, *16*(4), e0249182.
15. Jiang, Y., Jin, C., Yu, H., Wu, J., Chen, C., Yuan, Q., ... & Li, R. (2021). Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. *Annals of surgery*, *274*(6), e1153-e1161.

16. Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., ... & Li, R. (2021). Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications*, *12*(1), 1-11.
17. Jing, B., Deng, Y., Zhang, T., Hou, D., Li, B., Qiang, M., ... & Li, C. (2020). Deep learning for risk prediction in patients with nasopharyngeal carcinoma using multi-parametric MRIs. *Computer Methods and Programs in Biomedicine*, *197*, 105684.
18. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Sistema personalizado de recomendación de tratamiento que utiliza una red neuronal profunda de riesgos proporcionales de Cox. *Metodología de investigación médica BMC*, 18(1), 24.
19. Kaplan, E. L., & Meier, P. (1958). Estimación no paramétrica a partir de observaciones incompletas. *Revista de la Asociación Americana de Estadística*, 53(282), 457–481.
20. Kobayashi, K., Bolatkan, A., Shiina, S., & Hamamoto, R. (2020). Fully-connected neural networks with reduced parameterization for predicting histological types of lung cancer from somatic mutations. *Biomolecules*, *10*(9), 1249.
21. Lee, C., Zame, W., Yoon, J., & Van Der Schaar, M. (2018, April). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
22. Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.
23. Li, S., Deng, Y. Q., Zhu, Z. L., Hua, H. L., & Tao, Z. Z. (2021). A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging. *Diagnostics*, *11*(9), 1523.
24. Li, Y., Yang, T., Zhou, J., & Ye, J. (2018, May). Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data. In *proceedings of the 2018 SIAM international conference on data mining* (pp. 288-296). Society for Industrial and Applied Mathematics.
25. Li, Y., Wang, L., Wang, J., Ye, J., & Reddy, C. K. (2016, December). Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 231-240). IEEE.
26. Loya, H., Poduval, P., Anand, D., Kumar, N., & Sethi, A. (2020). Uncertainty estimation in cancer survival prediction. *arXiv preprint arXiv:2003.08573*.
27. Luck, M., Sylvain, T., Cardinal, H., Lodi, A., & Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*.
28. Meng, M., Gu, B., Bi, L., Song, S., Feng, D. D., & Kim, J. (2021). DeepMTS: Deep Multi-task Learning for Survival Prediction in Patients with Advanced Nasopharyngeal Carcinoma using Pretreatment PET/CT. *arXiv preprint arXiv:2109.07711*.
29. Navarrete Bellot, L. (2020). Aplicación de métodos de aprendizaje automático para el estudio del análisis de supervivencia en pacientes infectados por el VIH. *Trabajo de Final de Maestría, UOC/UB*

30. Peng, H., Dong, D., Fang, M. J., Li, L., Tang, L. L., Chen, L., ... & Ma, J. (2019). Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. *Clinical Cancer Research*, *25*(14), 4271-4279.
31. Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, *10*(1), 1-10.
32. Suo, Q., Zhong, W., Ma, F., Ye, Y., Huai, M., & Zhang, A. (2018, November). Multi-task sparse metric learning for monitoring patient similarity progression. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 477-486). IEEE.
33. Therneau, T., & Atkinson, E. (2020). The concordance statistic.
34. Therneau, T., & Atkinson, E. (2022). Concordance
35. Wang, P., Li, Y., & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, *51*(6), 1-36.
36. Yu, C.-N., Greiner, R., Lin, H.-C., & Baracos, V. (2011). Aprender las distribuciones de supervivencia del cáncer específicas del paciente como una secuencia de regresores dependientes. *En Advances in Neural Information Processing Systems* 24,1845-1853.
37. Yu, H., Zhang, X., Song, L., Jiang, L., Huang, X., Chen, W., ... & Fu, G. (2021). Large-scale gastric cancer screening and localization using multi-task deep neural network. *Neurocomputing*, *448*, 290-300.
38. Zhou, C., Zhou, L., Liu, F., Chen, W., Wang, Q., Liang, K., ... & Zhou, L. (2021). A Novel Stacking Heterogeneous Ensemble Model with Hybrid Wrapper-Based Feature Selection for Reservoir Productivity Predictions. *Complexity*, *2021*.

Anexo

```
head(veteran)
```

```
##   trt celltype time status karno diagtime age prior
## 1   1 squamous   72      1    60        7  69     0
## 2   1 squamous  411      1    70        5  64    10
## 3   1 squamous  228      1    60        3  38     0
## 4   1 squamous  126      1    60        9  63    10
## 5   1 squamous  118      1    70       11  65    10
## 6   1 squamous   10      1    20        5  49     0
```

Figura 1: análisis de los datos de la base *suvirval*[2]

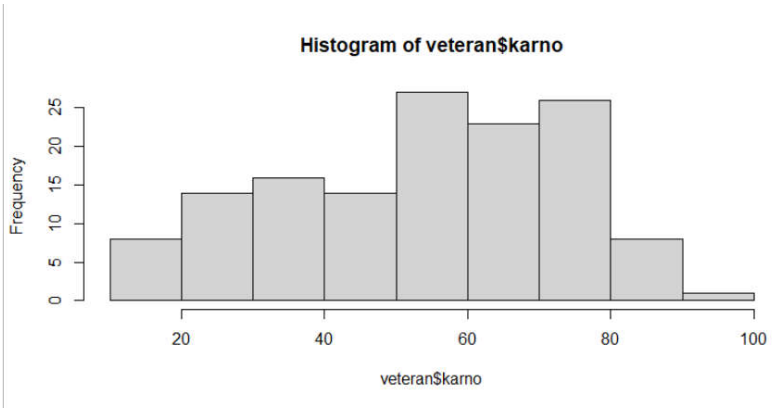

Figura 2: análisis de frecuencia de los pacientes en función del Karnofsky performance score

[2] Mas análisis descriptivos se presentan en el pdf resultante del informe *Markdown* que se presente adjunto a este reporte.

```
 Descriptive statistics by group
group: squamous
          vars  n   mean     sd median trimmed    mad min max range  skew kurtosis    se
trt          1 35   1.57   0.50      2    1.59   0.00   1   2     1 -0.28    -1.98  0.08
celltype*    2 35   1.00   0.00      1    1.00   0.00   1   1     0   NaN      NaN  0.00
time         3 35 200.20 248.23    111  152.41 142.33   1 999   998  1.91     3.35 41.96
status       4 35   0.89   0.32      1    0.97   0.00   0   1     1 -2.32     3.49  0.05
karno        5 35  60.86  20.49     60   62.07  14.83  20  90    70 -0.49    -0.69  3.46
diagtime     6 35  11.03  11.53      7    8.90   5.93   1  58    57  2.32     5.92  1.95
age          7 35  58.46  10.37     62   59.03  10.38  35  81    46 -0.40    -0.41  1.75
prior        8 35   4.00   4.97      0    3.79   0.00   0  10    10  0.39    -1.90  0.84
------------------------------------------------------------------------------------
group: smallcell
          vars  n  mean    sd median trimmed   mad min max range  skew kurtosis    se
trt          1 48  1.38  0.49    1.0    1.35  0.00   1   2     1  0.50    -1.79  0.07
celltype*    2 48  2.00  0.00    2.0    2.00  0.00   2   2     0   NaN      NaN  0.00
time         3 48 71.67 85.77   51.0   55.10 50.41   2 392   390  2.35     5.68 12.38
status       4 48  0.94  0.24    1.0    1.00  0.00   0   1     1 -3.50    10.49  0.04
karno        5 48 53.54 19.10   60.0   53.88 29.65  20  85    65 -0.15    -1.25  2.76
diagtime     6 48  9.25 13.91    4.0    6.58  2.97   1  87    86  3.85    17.80  2.01
age          7 48 59.88  9.92   62.5   60.83  8.90  35  72    37 -0.88    -0.25  1.43
prior        8 48  2.29  4.25    0.0    1.75  0.00   0  10    10  1.25    -0.45  0.61
------------------------------------------------------------------------------------
group: adeno
          vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
trt          1 27  1.67  0.48      2    1.70  0.00   1   2     1 -0.67    -1.61 0.09
celltype*    2 27  3.00  0.00      3    3.00  0.00   3   3     0   NaN      NaN 0.00
time         3 27 64.11 50.59     51   59.70 48.93   3 186   183  0.71    -0.50 9.74
status       4 27  0.96  0.19      1    1.00  0.00   0   1     1 -4.63    20.22 0.04
karno        5 27 58.11 22.12     60   58.70 29.65  10  99    89 -0.20    -0.83 4.26
diagtime     6 27  5.63  4.76      4    4.65  1.48   2  22    20  2.29     4.57 0.92
age          7 27 57.41 11.32     61   57.70  5.93  34  81    47 -0.43    -0.54 2.18
prior        8 27  1.85  3.96      0    1.30  0.00   0  10    10  1.53     0.36 0.76
------------------------------------------------------------------------------------
group: large
          vars  n   mean     sd median trimmed    mad min max range  skew kurtosis    se
trt          1 27   1.44   0.51      1    1.43   0.00   1   2     1  0.21    -2.03  0.10
celltype*    2 27   4.00   0.00      4    4.00   0.00   4   4     0   NaN      NaN  0.00
time         3 27 166.11 124.22    156  153.35 111.19  12 553   541  1.13     1.36 23.91
status       4 27   0.96   0.19      1    1.00   0.00   0   1     1 -4.63    20.22  0.04
karno        5 27  65.00  17.49     70   65.87  14.83  30  90    60 -0.66    -0.57  3.37
diagtime     6 27   8.15   4.99      8    7.96   5.93   1  18    17  0.27    -1.32  0.96
age          7 27  56.22  11.16     62   56.87   7.41  37  68    31 -0.57    -1.38  2.15
prior        8 27   3.70   4.92      0    3.48   0.00   0  10    10  0.51    -1.81  0.95
```

Figura 3: análisis descriptivo según el tipo de células que tiene la población.

```
 Descriptive statistics by group
group: 1
          vars  n   mean     sd median trimmed   mad min max range  skew kurtosis    se
trt          1 69   1.00   0.00      1    1.00  0.00   1   1     0   NaN      NaN  0.00
celltype*    2 69   2.35   1.05      2    2.32  1.48   1   4     3  0.40    -1.09  0.13
time         3 69 115.14 112.74     97   97.61 97.85   3 553   550  1.55     2.57 13.57
status       4 69   0.93   0.26      1    1.00  0.00   0   1     1 -3.23     8.54  0.03
karno        5 69  59.20  18.74     60   60.26 29.65  20  90    70 -0.40    -0.92  2.26
diagtime     6 69   8.65   8.76      5    7.11  4.45   1  58    57  3.02    12.79  1.05
age          7 69  57.51  10.81     62   58.21  8.90  34  81    47 -0.54    -0.74  1.30
prior        8 69   3.04   4.64      0    2.63  0.00   0  10    10  0.83    -1.33  0.56
------------------------------------------------------------------------------------
group: 2
          vars  n   mean     sd median trimmed   mad min max range  skew kurtosis    se
trt          1 68   2.00   0.00    2.0    2.00  0.00   2   2     0   NaN      NaN  0.00
celltype*    2 68   2.32   1.09    2.0    2.29  1.48   1   4     3  0.17    -1.30  0.13
time         3 68 128.21 193.83   52.5   87.16 55.60   1 999   998  2.95     9.56 23.50
status       4 68   0.94   0.24    1.0    1.00  0.00   0   1     1 -3.67    11.62  0.03
karno        5 68  57.93  21.40   60.0   58.39 29.65  10  99    89 -0.27    -0.89  2.59
diagtime     6 68   8.90  12.27    4.5    6.52  3.71   1  87    86  4.16    21.99  1.49
age          7 68  59.12  10.28   62.0   59.96  8.15  35  81    46 -0.70    -0.28  1.25
prior        8 68   2.79   4.52    0.0    2.32  0.00   0  10    10  0.96    -1.09  0.55
```

Figura 4: análisis descriptivo según el tipo de células que tiene la población.

```
##
##  0  1
## 68  8
```

```
fit2$y
```

```
##   14   50  118   43  137  135   90   91  130   57   92  121    9   93   99   72
##  25+  132   48   63   49  231   25 103+   15  216   21  186  314   13   99  87+
##   26    7   42  125   83   36   78   81  134  103   76   15   32  106  120  132
##   16   82    7   80  467  287  587   25  111   25  111   11  139   51  140  340
##   41   74   23   27   60   53  107  100  102   96   38   89   34   69  122  111
##   54  242  153  151   12    3   29    8   61   20   51   15   31  100   84   31
##   63   13   82   25   95   21   79  105   47  101   16    6  129   39   31  136
##  156  144  357  117    2 123+  389   80   92   99   30   10   53  122   18  378
##  124    4   88  127   86   52   22  109   70  112   35   40   48   30   12   75
##   45  126  283  164   30  162  97+   18  999   51   52   27   35   21    8  991
##  128   46   80   94  133   29   66  123    3   64  110   84   37    8   10  119
##   19    8   33   87  133   56  105   19  228 182+  83+  201   18  110 100+    7
```

```
fit4$y
```

```
##   14   50  118   43  137  135   90   91  130   57   92  121    9   93   99   72
##  25+  132   48   63   49  231   25 103+   15  216   21  186  314   13   99  87+
##   26    7   42  125   83   36   78   81  134  103   76   15   32  106  120  132
##   16   82    7   80  467  287  587   25  111   25  111   11  139   51  140  340
##   41   74   23   27   60   53  107  100  102   96   38   89   34   69  122  111
##   54  242  153  151   12    3   29    8   61   20   51   15   31  100   84   31
##   63   13   82   25   95   21   79  105   47  101   16    6  129   39   31  136
##  156  144  357  117    2 123+  389   80   92   99   30   10   53  122   18  378
##  124    4   88  127   86   52   22  109   70  112   35   40   48   30   12   75
##   45  126  283  164   30  162  97+   18  999   51   52   27   35   21    8  991
##  128   46   80   94  133   29   66  123    3   64  110   84   37    8   10  119
##   19    8   33   87  133   56  105   19  228 182+  83+  201   18  110 100+    7
```

Figura 5: análisis de los datos censurados por derecha

```
survdiff(Surv(time, status) ~ celltype, data = data.train)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ celltype, data = data.train)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype=squamous   26       23     38.2     6.046     13.02
## celltype=smallcell 36       33     21.0     6.923     10.14
## celltype=adeno      18       17     10.8     3.518      4.22
## celltype=large      16       15     18.0     0.506      0.66
##
##  Chisq= 21.1  on 3 degrees of freedom, p= 1e-04
```

```
survdiff(Surv(time, status) ~ prior + status, data = data.train)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ prior + status, data = data.train)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0, status=0   6        0     5.51   5.50599    6.0703
## prior=0, status=1  63       63    55.94   0.89111    2.5941
## prior=10, status=0  2        0     1.13   1.13051    1.1684
## prior=10, status=1 25       25    25.42   0.00707    0.0109
##
##  Chisq= 7.8  on 3 degrees of freedom, p= 0.05
```

Figura 6: análisis sobre el tipo de células, y sobre el estado + la terapia

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = data.train)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    20     77      20   0.7917  0.0414        0.7145        0.877
##    50     59      16   0.6234  0.0496        0.5333        0.729
##   100     39      18   0.4300  0.0511        0.3406        0.543
##   350      7      27   0.0916  0.0324        0.0458        0.183
```

Figura 7: principales resultados del modelo KM

```
## Call:
## coxph(formula = Surv(time, status) ~ ., data = data.train, x = TRUE)
##
##   n= 96, number of events= 88
##
##                coef  exp(coef)   se(coef)      z Pr(>|z|)
## trt        0.1428218  1.1535242  0.2565892  0.557  0.57779
## celltype1 -0.9161660  0.4000499  0.2341873 -3.912 9.15e-05 ***
## celltype2  0.4729743  1.6047601  0.1992676  2.374  0.01762 *
## celltype3  0.6220558  1.8627536  0.2321140  2.680  0.00736 **
## karno     -0.0394631  0.9613054  0.0064027 -6.164 7.11e-10 ***
## diagtime  -0.0050613  0.9949515  0.0104799 -0.483  0.62913
## age       -0.0159867  0.9841404  0.0119828 -1.334  0.18216
## prior      0.0007983  1.0007986  0.0288504  0.028  0.97793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## trt          1.1535     0.8669    0.6976    1.9074
## celltype1    0.4000     2.4997    0.2528    0.6331
## celltype2    1.6048     0.6231    1.0859    2.3715
## celltype3    1.8628     0.5368    1.1819    2.9358
## karno        0.9613     1.0403    0.9493    0.9734
## diagtime     0.9950     1.0051    0.9747    1.0156
## age          0.9841     1.0161    0.9613    1.0075
## prior        1.0008     0.9992    0.9458    1.0590
##
## Concordance= 0.759  (se = 0.029 )
## Likelihood ratio test= 60.59  on 8 df,   p=4e-10
## Wald test            = 55.67  on 8 df,   p=3e-09
## Score (logrank) test = 60.99  on 8 df,   p=3e-10
```

Figura 8: principales resultados del modelo Cox

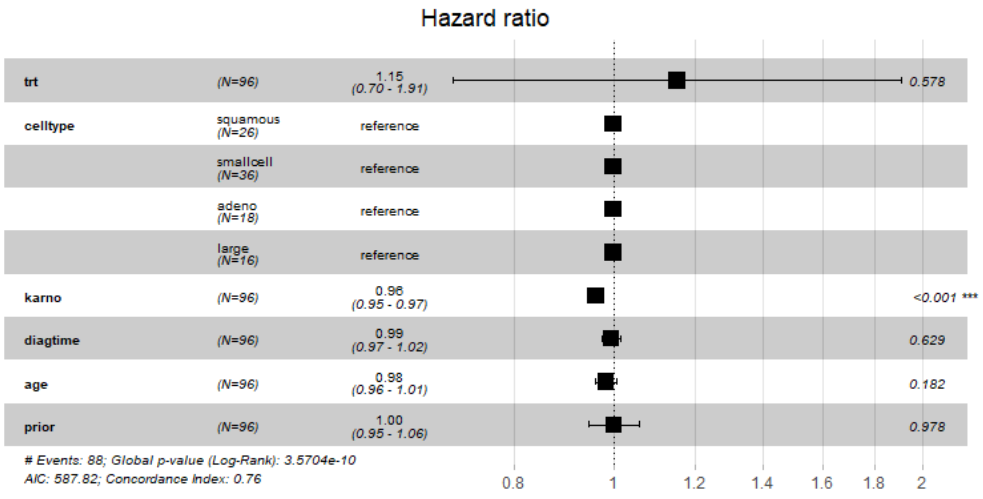

Figura 9: presentación gráfica de las tasas de riesgo del modelo Cox

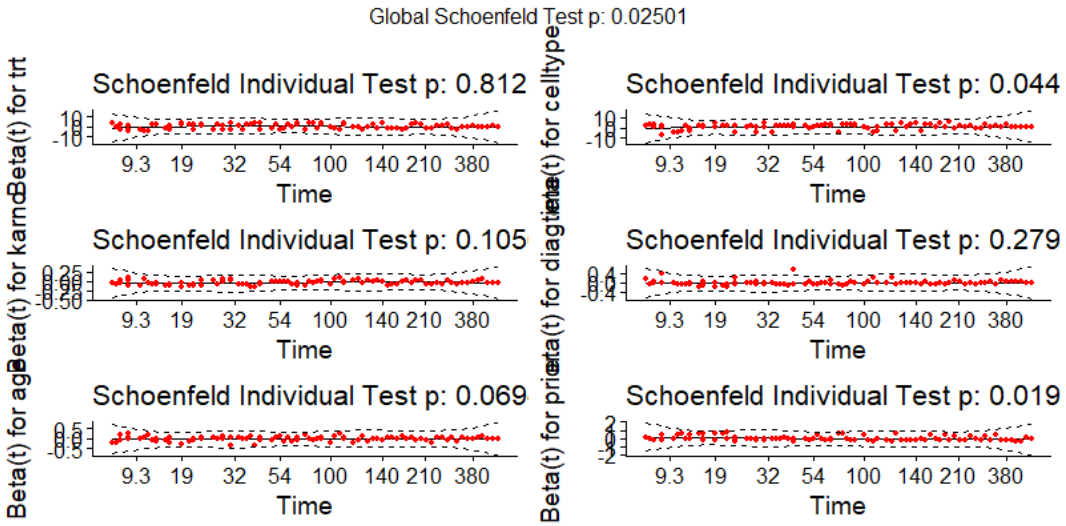

Figura 10: presentación gráfica de las tasas de riesgo proporcional del modelo Cox

```
##
## Call:  mtlr(formula = Surv(time, status) ~ ., data = data.train, nintervals = 9)
##
## Time points:
##  [1]  11.6  19.0  26.8  46.6  56.9  87.0 107.3 133.5 177.1 296.8
##
##
## Weights:
##           Bias      trt celltype1 celltype2 celltype3   karno diagtime      age
## 11.64   0.1171 -0.03339   0.03081  0.000274   0.03571 -0.0197  0.03465 -0.01826
## 19     -0.0507 -0.03835   0.00333 -0.007206   0.00663 -0.0780  0.01243 -0.02928
## 26.82  -0.1685  0.01514  -0.00729  0.014363  -0.00533 -0.1320  0.00240 -0.03835
## 46.64  -0.0351  0.01607   0.01185  0.031709   0.03268 -0.1219 -0.02626 -0.02049
## 56.91   0.4227  0.01776  -0.03017  0.038213   0.02317 -0.1353  0.01017 -0.01955
## 87     -0.1226  0.03394  -0.02180  0.059672   0.05490 -0.1428  0.00457  0.00763
## 107.27 -0.2863  0.02987  -0.05675  0.059685   0.03929 -0.1099 -0.00350  0.00468
## 133.55 -0.0104  0.00147  -0.04636  0.049062   0.02854 -0.0682  0.00628 -0.00473
## 177.09 -0.0166 -0.04860  -0.06836  0.057887   0.03886 -0.0690 -0.00596  0.03760
## 296.82 -0.0524 -0.04662  -0.05469  0.035507   0.02726 -0.0490  0.01090  0.01674
##           prior
## 11.64   0.01958
## 19      0.05089
## 26.82   0.06572
## 46.64   0.01226
## 56.91   0.00542
## 87      0.01663
## 107.27  0.00601
## 133.55 -0.01814
```

Figura 11: principales resultados del modelo MTLR