Análisis de supervivencia mediante IA de los abonados a un centro deportivo

Andrés Espín Rearte
Grado de Ingeniería Informática
Área de Inteligencia Artificial

David Isern Alarcón
Susana Acedo Nadal

Enero de 2024

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de supervivencia mediante IA de los abonados a un centro deportivo.*
Nombre del autor:	*Andrés Espín Rearte*
Nombre del consultor/a:	*David Isern Alarcón*
Nombre del PRA:	*Susana Acedo Nadal*
Fecha de entrega (mm/aaaa):	01/2024
Titulación:	*Grado de Ingeniería Informática*
Área del Trabajo Final:	*Inteligencia Artificial*
Idioma del trabajo:	*Castellano*
Palabras clave	*Churn-rate, Gym, Survival Analysis*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

Según las estimaciones de la OMS hasta 5 millones de muertes anuales podrían ser evitadas si se aumentara el nivel de actividad física de la población mundial. Una gestión eficaz de los centros deportivos, que consiga incorporar el hábito de su práctica regular a la vida diaria, será clave para la mejora de estos índices.

Este proyecto tiene como objetivo mejorar la fidelización y retención de los abonados de un centro deportivo mediante el análisis de sus datos de membresía. Se trata de entender qué variables mejoran el tiempo de permanencia y cuales se relacionan con la baja de los clientes, así como de intentar obtener una predicción de cuáles están en riesgo de parar la práctica deportiva.

Una vez el conjunto de datos está previamente anonimizado y pretratado, se entrenan y aplican seis distintos algoritmos de aprendizaje automático supervisado adaptados al análisis de supervivencia para obtener predicciones de las funciones de supervivencia y riesgo acumulado de cada observación, que nos indican respectivamente la probabilidad de permanecer en el centro más de un periodo de tiempo *t*, y la tasa de cancelaciones' que suceden antes de *t*.

Se evalúa y compara resultados del rendimiento de los modelos aplicando las métricas apropiadas para el análisis de supervivencia y se obtiene el conjunto de características principales que los definen, de manera que del análisis de distintas cohortes se pueden obtener *insights* a partir de los que definir estrategias de retención para el centro.

Abstract (in English, 250 words or less):

According to estimates from the WHO, up to 5 million annual deaths could be prevented by increasing the level of physical activity in the global population. Effective management of sports centers that successfully incorporates the habit of regular physical activity into daily life will be key to improving these statistics.

This project aims to improve the loyalty and retention of subscribers to a sports center by analyzing their membership data. The goal is to understand which variables enhance the length of membership and which are related to customer attrition, as well as attempting to predict which individuals are at risk of discontinuing their sports activities.

Once the dataset is pre-anonymized and preprocessed, six different supervised machine learning algorithms adapted for survival analysis are trained and applied to obtain predictions of survival functions and cumulative hazard for each observation. These predictions respectively indicate the probability of remaining at the center for more than a period of time 't' and the rate of cancellations that occur before 't'.

Performance results of the models are evaluated and compared using appropriate survival analysis metrics, and the set of key features defining them is obtained. This allows for insights to be derived from the analysis of different cohorts, which can then be used to define retention strategies for the center.

Índice

Lista de figuras

2. Introducción

2.1 Contexto y justificación del Trabajo

Según las estimaciones de la OMS hasta 5 millones de muertes anuales podrían ser evitadas si se aumentara el nivel de actividad física de la población mundial, siendo entre 250 y 300 minutos semanales el tiempo recomendado. Su práctica regular, además de prevenir enfermedades cardiovasculares, óseas, algunos tipos de cáncer y psicológicas, es clave en la prevención del envejecimiento. Tomar las medidas adecuadas de promoción y adhesión a la práctica deportiva comporta tanto mejoras en los índices de salud, felicidad y bienestar de la población, como importantes ahorros en gastos sanitarios.

En España, si bien encontramos grandes diferencias en su práctica entre los territorios, las grandes ciudades como Barcelona, Madrid o Valencia se sitúan entre las 20 ciudades del mundo con mayor promoción de la actividad física según el último informe de la Encuesta de Hábitos Deportivos en España del INE. Sin embargo, se calcula que solo un 10% de los nuevos inscritos a un centro deportivo, al cabo de un año han conseguido incorporar ese hábito a su vida diaria.

Por otro lado, la viabilidad económica de estos centros depende en gran medida de su capacidad de fidelizar y retener a sus socios, precisamente por ese elevado porcentaje de rotación y abandono en el uso de su servicio. En un centro medio como el del objeto de este estudio, nos encontraremos con alrededor de un 5-8% cancelaciones mensuales sobre el total de clientes, esto son unas 200-320, que difícilmente son compensadas con las nuevas inscripciones. Cuando se aplican estrategias específicas de retención de clientes se consiguen reducciones de hasta el 12-15% de esa tasa de rotación, siendo las más eficaces las que actúan de manera anticipada y preventiva. Sin embargo, el volumen y la identificación temprana de las posibles bajas constituye los principales obstáculos para llevarlas a cabo.

Una gestión eficaz que facilite la tarea de incorporar la práctica deportiva regular a la vida diaria, será clave para la mejora de ambos índices. Para ello, como propone Gold, C. S. (2020) en su libro *Fighting Churn with Data: The Science and Strategy of Customer Retention* la incorporación de herramientas de Inteligencia Artificial y Aprendizaje Automático en la retención de clientes, aún en un estado muy temprano de aplicación en el sector deportivo, puede suponer nuevas oportunidades.

Actualmente la mayoría de los centros cuentan con soluciones de CRM y gestión de clientes específicas del sector, pero estas se encuentran especialmente orientadas a la captación de nuevos abonados y a las tareas administrativas, siendo la retención la gran asignatura pendiente en su gestión, y la que debería permitir conseguir un incremento en la adhesión a la práctica deportiva.

A partir de los sistemas de acceso y reserva de los centros deportivos, junto con los propios CRM, así como de las cámaras de vigilancia o incluso sensores de uso, se dispone de gran cantidad de datos sobre los hábitos de consumo de los abonados, a partir de los que es factible categorizarlos y analizar sus hábitos de consumo, además de prevenir los momentos de riesgo de desapego a la práctica.

Disponer de esta información será clave para los distintos departamentos a la hora de diseñar la oferta de servicio, y de definir campañas de comunicación adecuadas, que aumenten la vinculación del abonado con el centro, además de una mayor personalización del servicio y una mejor experiencia de usuario.

Nota: al cuaderno de trabajo de Google Colab, los datos del proyecto y al archivo README.txt para su instalación se puede acceder en el repositorio público https://github.com/AndresEspin/TFG24/tree/main.

2.2 Objetivos del Trabajo

En este proyecto de análisis de datos mediante aprendizaje automático se definen los siguientes objetivos generales y específicos que proporcionan una guía para el desarrollo y evaluación del proyecto.

<u>- Objetivo General:</u>

- Obj.G Desarrollar un sistema de análisis de datos basado en IA con el fin de predecir futuras bajas de abonados y determinar el tiempo de permanencia de los mismos.

<u>- Objetivos Específicos:</u>

- Obj.1 Anonimización y preparación de los datos de usos y membresía de los socios de un centro deportivo, asegurando la calidad y la integridad de los datos.
- Obj.2 Realizar un análisis exploratorio de datos para comprender las

características y tendencias de los miembros del gimnasio, incluyendo, por ejemplo, la distribución de edades, género y patrones de asistencia.

- Obj.3 Identificar posibles problemas de privacidad de datos, asegurando que se cumpla la regulación de protección de datos aplicable.

- Obj.4 Estudiar sobre los diferentes modelos estadísticos aplicables al llamado problema de Análisis de Supervivencia y ver cómo se ajusta al caso.

- Obj.5 Entrenar modelos de IA para predecir las variables respuesta baja de los clientes y tiempo de permanencia, utilizando algoritmos de aprendizaje supervisado. Evaluar y comparar la precisión y el rendimiento de los modelos de predicción con las métricas correspondientes.

- Obj.6 Realizar análisis de importancia de características para identificar qué variables influyen más en las predicciones de bajas y tiempo de permanencia.

- Obj.7 Incluir propuesta de visualización y paneles de control para comunicar los resultados del análisis de datos y las predicciones, y extrapolar recomendaciones basadas en los resultados del análisis, para mejorar la retención de clientes y la gestión de membresías.

- Obj.8 Proponer posibles ampliaciones del proyecto que puedan dar continuidad a la problemática estudiada, así como definir qué nuevas vías de obtención de datos sería interesante contemplar.

- Obj.9 Evaluar la viabilidad técnica y económica y el retorno de la inversión de la implementación del sistema propuesto.

2.3 Enfoque y método seguido

La cadena de centros deportivos en la que se quiere aplicar el proyecto dispone de diversos centros con características dispares. Se escoge un centro representativo por volumen de socios, precio, etc. y se trabaja con los datos de sus abonados y ex-abonados con diferentes ventanas de tiempo desde el inicio de su funcionamiento hasta la fecha actual.

La compañía actualmente no trabaja con sistemas predictivos y el uso de los datos se limita meramente a sencillos análisis descriptivos; un primer acercamiento a

la IA con resultados positivos en la reducción de bajas supondrá la apertura a su utilización en diversas problemáticas. Sin embargo, la incorporación de diferentes sistemas y soluciones a medida a lo largo del tiempo ha hecho que los datos se encuentren dispersos en diferentes bases de datos y, en gran medida, incompletos.

Como recientemente desarrolla también Sobreiro, Pablo. N. (2023) en su tesis doctoral, el estudio de la retención de clientes a un servicio se enmarca en el ámbito específico conocido como Análisis de Supervivencia, en el que la ocurrencia o no de un suceso y el tiempo en que sucede pueden ser una incógnita ya que se trabaja con lo que se conoce como datos censurados. Si bien aún son pocos los estudios y aplicaciones concretas en este ámbito, amplia literatura sobre el tema aborda su aplicación en sectores de características análogas como el de la salud, el mantenimiento de edificios y estructuras o la propia retención de clientes en otro tipo de servicios, como los trabajos publicados de Orozco, A (2012) o Morillo Leal, J (2023) o cuadernos disponibles online como el *"Customer Churn Project Daniela"* (2020) en kaggle. A partir del estudio de esta información y de las librerías disponibles sobre el tema se seleccionan y comparan aquellos modelos de aprendizaje automático que se ajustan mejor al caso de estudio.

Se sigue el esquema CRISP en el desarrollo del proyecto; se parte de una sólida comprensión del negocio y de las necesidades y carencias de los distintos departamentos implicados (Ventas, Marketing, Producto y Sistemas) así como de los datos y de la información relevante de y para cada uno de ellos. A ello, le seguirán de manera natural la comprensión y preparación de los datos, el modelado y la evaluación de su resultado. El despliegue de la solución obtenida quedará fuera del alcance de este proyecto y como propuesta de continuación.

Una de las mayores dificultades con las que se encuentra el proyecto, es la complejidad en la recolección, limpieza, unificación y análisis previo de los datos; se encuentran incompletos, originarios de diversas fuentes y en algunos casos con inconsistencias. Además, la naturaleza propia de la cuestión planteada en el proyecto hace que nos encontremos con muchos datos faltantes y la forma de imputación de estos es uno de los temas de estudio en el trabajo, así como la posible estrategia correctiva de balanceo de clases.

En el análisis exploratorio y descriptivo de los datos limpios y de sus variables, se estudia y su grado de correlación, descartando las que puedan ser irrelevantes y adaptando su formato a las necesidades del modelado.

Se quiere comparar el resultado de aplicar el modelo clásico de riesgos proporcionales de la *regresión de Cox* sobre los datos, con los obtenidos utilizando variantes específicas de los algoritmos de aprendizaje automático adaptadas al análisis de supervivencia *Random Survival Forest, Survival Support Vector Machine* y *Gradient Boosting Survival Analysis.* Para ello se utilizan las implementaciones disponibles en la biblioteca *scikit-survival* (Pölsterl, 2020) que constituyen el estado del arte en aplicación de aprendizaje automático a análisis de supervivencia; el proceso de entrenamiento se prueba con división sobre sets de entrenamiento, test y validación y validación cruzada.

Para llevar a cabo esta comparación será necesario utilizar métricas adecuadas y específicas de este tipo de análisis como *C-Índex*. A partir de la interpretación de los resultados obtenidos, se buscan *insights* que puedan mejorar las estrategias de retención de cada departamento, vinculandose con el negocio.

Una vez comprobada su eficacia, se propone el protocolo de evaluación y revisión continua del modelo, así como de despliegue de la solución, contrastando el coste del presupuesto con el rendimiento económico teórico de la ejecución del proyecto, en el estudio de viabilidad económica.

2.4 Planificación del Trabajo

2.4.1 Recursos materiales y humanos

Para la realización del proyecto se dispone de un ordenador portátil Airbook de Mac y un ordenador de sobremesa. Se trabaja principalmente en Python sobre Google Colab, con almacenamiento y seguimiento de versiones en Google Drive. Microsoft Office es utilizado para la edición de documentos.

Se estima en recursos humanos para los siguientes roles necesarios:

- 10 horas semanales del Jefe de Proyecto durante 16 semanas - 160 h JP
- 10 horas semanales de Senior Programer durante 12 semanas - 120 h SR

Total de 280 horas de RRHH

2.4.2 Fases del proyecto

La fase inicial del proyecto incluye la definición de los objetivos del proyecto, la revisión de la literatura relevante sobre el tema, centrada principalmente en el análisis

del *churn-rate* (o tasa de rotación clientes) como problema del tipo análisi de supervivencia, el análisis de los posibles riesgos existentes durante la realización del proyecto, así como la definición de las correspondientes acciones de mitigación, la planificación de los tiempos de desarrollo e hitos a conseguir, y la recopilación de los datos. Se lleva a cabo, también, la concreción del entorno de desarrollo y las tecnologías y lenguajes a utilizar.

La segunda fase se centra en el análisis preliminar y la preparación de estos datos. Incluye la limpieza, el análisis de datos incompletos, el estudio de las variables iniciales para comprender las características de los miembros del centro, la gestión de valores nulos y *outliers*, el balance de las diferentes clases, y la ingeniería de variables, si es necesario definir nuevas que ayuden a explicar el objeto de análisis. La privacidad de los datos puede ser una preocupación importante en esta etapa y se abordarán posibles problemas derivados.

En la tercera fase, se escogen y prueban distintos algoritmos de inteligencia artificial según el modelo estudiado para predecir las bajas de los clientes y el tiempo de permanencia de los mismos. Se dividen los datos en set de entrenamiento y test con los modelos se entrenan y se comparan, evaluando sus resultados con métricas relevantes al caso. Si es necesario, se hace reajuste de hiperparámetros y corrección de errores, intentando detectar en qué casos la predicción no es acertada.

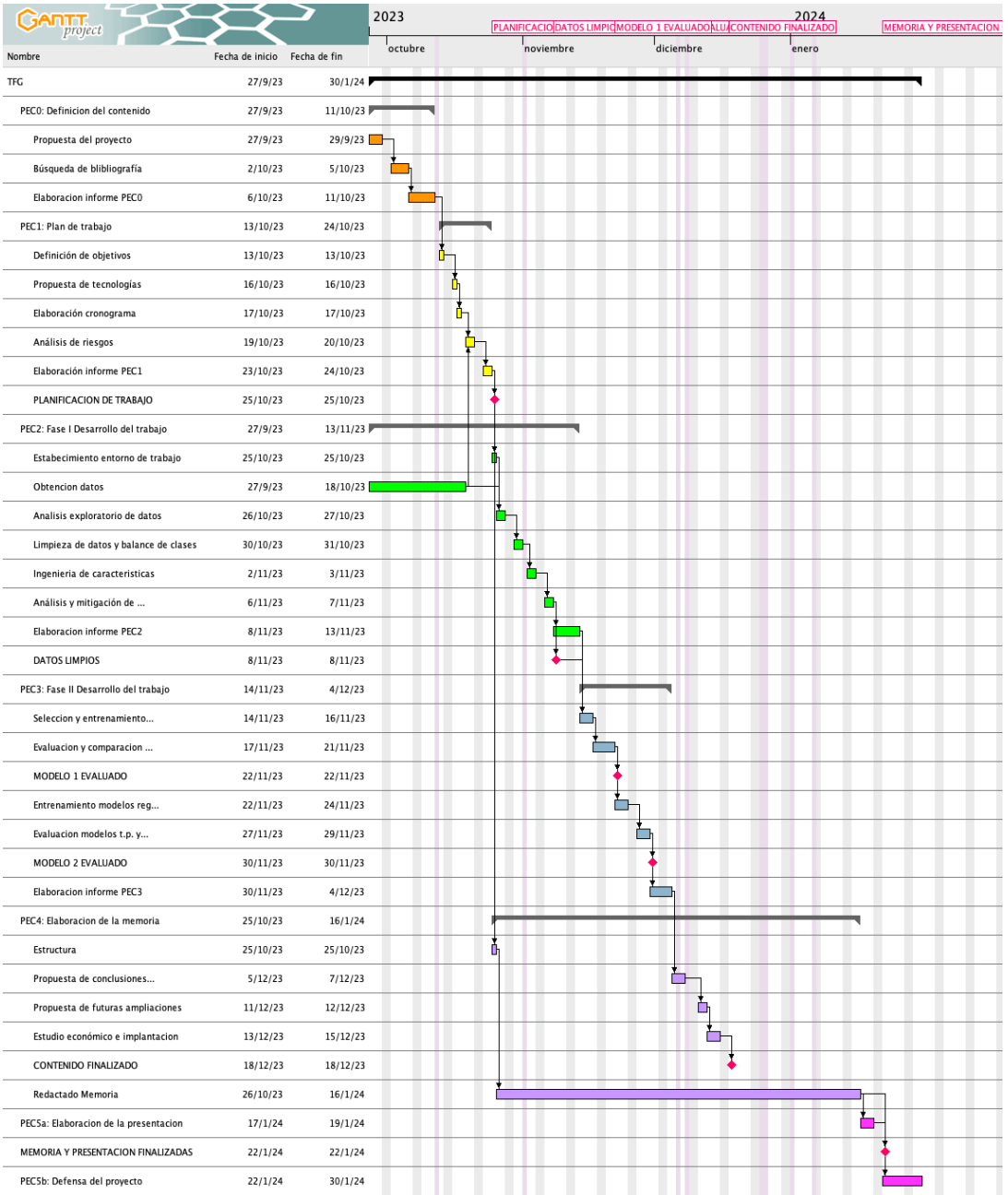
La cuarta y última fase se enfoca en la interpretación de resultados y detección y comunicación de *insights*. Se crean visualizaciones para comunicar los hallazgos del análisis y se formulan recomendaciones prácticas basadas en los resultados. Se evalúa la viabilidad económica de la implantación del sistema de predicción, se proponen futuras líneas de ampliación del proyecto, se comentan las dificultades encontradas y se documenta todo el proceso en la memoria. El proyecto culmina con la presentación formal y defensa del trabajo realizado.

En el lateral de la *Figura 1* se observa gráficamente la Estructura de Descomposición del Proyecto (EDP).

2.4.3 Diagrama de Gantt

Ajustando las tareas para la consecución de los diferentes objetivos y etapas del proyecto al calendario de entregas propuesto, se plantea la planificación del trabajo según se muestra en el diagrama de Gantt de la Figura 1.

Cada etapa está presentada en un color distinto y los hitos están marcados en rojo. Para mayor claridad se detallan también las fechas de inicio y fin de cada tarea. Se cuentan para la planificación, los días laborables, de lunes a viernes, excluyendo los festivos.



Nombre	Fecha de inicio	Fecha de fin
TFG	27/9/23	30/1/24
PEC0: Definicion del contenido	27/9/23	11/10/23
Propuesta del proyecto	27/9/23	29/9/23
Búsqueda de blibliografía	2/10/23	5/10/23
Elaboracion informe PEC0	6/10/23	11/10/23
PEC1: Plan de trabajo	13/10/23	24/10/23
Definición de objetivos	13/10/23	13/10/23
Propuesta de tecnologías	16/10/23	16/10/23
Elaboración cronograma	17/10/23	17/10/23
Análisis de riesgos	19/10/23	20/10/23
Elaboración informe PEC1	23/10/23	24/10/23
PLANIFICACION DE TRABAJO	25/10/23	25/10/23
PEC2: Fase I Desarrollo del trabajo	27/9/23	13/11/23
Estabecimiento entorno de trabajo	25/10/23	25/10/23
Obtencion datos	27/9/23	18/10/23
Analisis exploratorio de datos	26/10/23	27/10/23
Limpieza de datos y balance de clases	30/10/23	31/10/23
Ingenieria de caracteristicas	2/11/23	3/11/23
Análisis y mitigación de ...	6/11/23	7/11/23
Elaboracion informe PEC2	8/11/23	13/11/23
DATOS LIMPIOS	8/11/23	8/11/23
PEC3: Fase II Desarrollo del trabajo	14/11/23	4/12/23
Seleccion y entrenamiento...	14/11/23	16/11/23
Evaluacion y comparacion ...	17/11/23	21/11/23
MODELO 1 EVALUADO	22/11/23	22/11/23
Entrenamiento modelos reg...	22/11/23	24/11/23
Evaluacion modelos t.p. y...	27/11/23	29/11/23
MODELO 2 EVALUADO	30/11/23	30/11/23
Elaboracion informe PEC3	30/11/23	4/12/23
PEC4: Elaboracion de la memoria	25/10/23	16/1/24
Estructura	25/10/23	25/10/23
Propuesta de conclusiones...	5/12/23	7/12/23
Propuesta de futuras ampliaciones	11/12/23	12/12/23
Estudio económico e implantacion	13/12/23	15/12/23
CONTENIDO FINALIZADO	18/12/23	18/12/23
Redactado Memoria	26/10/23	16/1/24
PEC5a: Elaboracion de la presentacion	17/1/24	19/1/24
MEMORIA Y PRESENTACION FINALIZADAS	22/1/24	22/1/24
PEC5b: Defensa del proyecto	22/1/24	30/1/24

Figura 1. Diagrama de Gantt de la planificación del proyecto.

2.4.4 Análisis de riesgos

Durante el desarrollo del proyecto, así como en la propuesta de implantación del mismo, nos podemos encontrar desafíos derivados de los siguientes riesgos, para los que se proponen las consiguientes acciones de mitigación, esenciales para abordarlos:

- *Riesgo R1* Calidad y disponibilidad suficiente de datos: Los datos, al provenir de diversas fuentes, pueden contener errores, duplicados, valores incompletos u otros problemas que pueden afectar negativamente la calidad de los resultados. La cantidad y la calidad de los datos disponibles pueden ser insuficientes para entrenar modelos precisos. Riesgo alto e impacto alto.

  Acción de mitigación: Realizar una limpieza exhaustiva de los datos, identificando y corrigiendo errores, duplicados y valores faltantes. También, explorar la posibilidad de generar datos adicionales o formas de mejorar la calidad de los datos existentes si es necesario.

- *Riesgo R2* Sesgo en los datos: Los datos o la selección que de ellos se haga puede provocar sesgos inherentes que se reflejarán en los modelos de inteligencia artificial, lo que puede llevar a predicciones erróneas. Por ejemplo, cabe tener en cuenta la situación anómala durante el período de pandemia. Riesgo alto e impacto alto.

  Acción de mitigación: Realizar un análisis cuidadoso de los datos para identificar y cuantificar cualquier posible sesgo presente. Estudiar y aplicar técnicas de mitigación de sesgo para asegurarse de que el modelo no los refleja.

- *Riesgo R3* Interpretabilidad de modelos: Los modelos de IA pueden ser complejos y difíciles de interpretar, lo que puede dificultar la comprensión de las conclusiones obtenidas y la toma de decisiones. Riesgo medio e impacto medio.

  Acción de mitigación: Utilizar modelos de IA interpretables si es posible. Si se utilizan modelos complejos, valorar técnicas de explicabilidad, como la importancia de características, para comprender mejor las decisiones del modelo.

- *Riesgo R4* Generalización insuficiente: Los modelos pueden funcionar bien en el conjunto de datos de entrenamiento y testeo, pero pueden no generalizar bien a nuevos datos o escenarios no vistos, como la aplicación del modelo a otros centros de la misma cadena, nuevas ofertas de cuota o cambios en los patrones

de comportamiento del cliente. Riesgo alto e impacto alto.

Acción de mitigación: Implementar técnicas de validación cruzada y ajuste de hiperparámetros para mejorar la generalización del modelo. También, considerar si es posible la recopilación de datos adicionales para abordar esos escenarios no vistos.

- *Riesgo R5* Desequilibrio de clases: En la predicción de bajas es posible que haya un desequilibrio significativo entre las clases (por ejemplo, muchas personas que no se dan de baja). Esto puede afectar la eficacia del modelo. Riesgo alto e impacto alto.

  Acción de mitigación: Utilizar técnicas de balanceo de clases, como el sobremuestreo o submuestreo estratificado o la generación de datos aleatorios, para abordar el desequilibrio y mejorar la eficacia del modelo.

- *Riesgo R6* Cambios en la tecnología y en el entorno: La tecnología y las herramientas de IA así como el entorno pueden evolucionar rápidamente, lo que podría hacer que los modelos propuestos queden obsoletos. Por ejemplo, la posibilidad de incorporación de nuevas fuentes de datos. Riesgo medio e impacto medio.

  Acción de mitigación: Revisar los últimos estudios y tendencias en tecnología y herramientas de IA aplicables al tema. Diseñar modelos que puedan adaptarse a cambios tecnológicos y del entorno. Proponer ciclos de revisión y mantenimiento del modelo en la implantación del proyecto.

- *Riesgo R7* Costo y recursos: El desarrollo y la aplicación de modelos de inteligencia artificial puede ser costoso en términos de recursos humanos y de hardware. Además, puede requerir inversión en capacitación y herramientas. Riesgo bajo e impacto bajo.

  Acción de mitigación: Realizar una evaluación de costos y recursos precisa y establecer un presupuesto adecuado, complementando con un análisis claro y detallado del ROI.

- *Riesgo R8* Falta de colaboración interdepartamental y resistencia al cambio: El éxito del proyecto de IA en el ámbito del centro deportivo puede depender de la colaboración entre expertos en datos, expertos en fitness y profesionales de marketing y del área de negocio. La falta de comunicación y colaboración ante el

cambio podría limitar el impacto del proyecto. Riesgo medio e impacto alto.

Acción de mitigación: Fomentar la colaboración entre diferentes departamentos. Comunicar a la dirección de manera clara los beneficios del proyecto y trabajar en la adopción del sistema con una propuesta clara de comunicación de resultados y paneles de control que ponga de manifiesto las ventajas del mismo. Establecer un proceso claro de implantación y formación del uso del sistema.

- *Riesgo R9* Dependencia de terceros: Si se utiliza software o servicios de terceros en el proyecto, la dependencia de estos proveedores puede ser un riesgo si se enfrentan a problemas de disponibilidad o confiabilidad. Riesgo medio e impacto medio.

  Acción de mitigación: Diversificar las fuentes de software o servicios de terceros si es posible. Tener planes de contingencia en caso de problemas con proveedores e incluir los costes en el presupuesto.

- *Riesgo R10* Privacidad y regulaciones: el manejo inadecuado de datos personales puede tener implicaciones legales. Así mismo, el celo en el acceso a datos por parte de la compañía puede limitar los recursos necesarios para el desarrollo del mismo. Riesgo bajo e impacto alto.

  Acción de mitigación: Asegurar que se cumplan las regulaciones de privacidad de datos y establecer políticas claras de su acceso. Consensuar con el departamento legal para garantizar el cumplimiento de las normativas.

2.4.5 Herramientas y tecnologías utilizadas

Para la realización del proyecto se pretende explorar y utilizar para aprender sobre las siguientes herramientas y tecnologías:

- Lenguaje *Python* para el desarrollo del modelo, por su uso extendido en modelos de aprendizaje automático y por la amplia gama de bibliotecas que ofrece.

- Algunas de las bibliotecas utilizadas serán *Pandas* y *Numpy*, esenciales para el preprocesamiento y análisis de datos. Como herramientas de aprendizaje automático, las bibliotecas *scikit-learn*, *XGBoost*, y *TensorFlow* para el entrenamiento de los modelos predictivos.

- *VSCode* como entorno de desarrollo integrado, por la integración de control de

versiones, las herramientas de depuración y el soporte para ambos lenguajes.

- *Google Colab* como entorno de ejecución, para tener acceso a recursos de hardware acelerado, necesarios para llevar a cabo tareas de aprendizaje automático que requieran GPU o TPU, insuficientes localmente, y acelerar el entrenamiento de los modelos.

- *PostgreSQL* como base de datos relacional por el volumen de los datos y la posibilidad de escalabilidad.

- A priori durante el desarrollo, el volumen de datos no sería tan grande como para ser necesario, pero se contempla estudiar el uso y necesidad de *Spark* para permitir la escalabilidad en la implantación del modelo.

- Docker como tecnología de contenedores para facilitar la implementación y la gestión de la aplicación de aprendizaje automático en los diferentes entornos.

- Tableau como herramienta para presentación de datos y diseño de paneles de control.

- GitHub como repositorio y control de versiones.

2.5 Sumario de productos obtenidos

Los productos que se entregan en el proyecto incluyen:

- Código Python. El código fuente utilizado para la obtención de los modelos ML se facilitará en un fichero comprimido.

- La Memoria del proyecto, la documentación del proyecto se entrega en formato PDF.

- El video de presentación y defensa del TFG.

- Una dirección en GitHub en la que se comparten los ficheros relacionados con el trabajo realizado (código fuente, datasheets, ficheros de configuración, etc.).

2.6 Breve descripción de los otros capítulos de la memoria

El capítulo 3 establece el contexto teórico del proyecto. Explora la literatura existente sobre modelos predictivos aplicados al análisis del *churn-rate* en distintos

sectores. Analiza los métodos, algoritmos y enfoques utilizados en proyectos similares para entender las mejores prácticas y la base teórica necesaria para desarrollar el modelo predictivo en este estudio.

El capítulo 4 detalla en las dos fases en que se divide la ejecución propia del proyecto, el desarrollo del sistema predictivo para evaluar el riesgo de baja de clientes en el centro deportivo. En la primera incluye la recopilación de datos, la selección y preparación de características relevantes mediante ingeniería de variables, la imputación de nulos y el balance de clases, así como el análisis descriptivo de datos. En la segunda, la elección de algoritmos de aprendizaje automático, la construcción y entrenamiento del modelo, así como la validación del mismo. Es el núcleo del trabajo y muestra cómo se implementa el conocimiento teórico en la práctica.

El capítulo 5 presenta los resultados obtenidos del modelo predictivo. Describe la efectividad del modelo según las métricas elegidas, la precisión de las predicciones y cómo se comparan los resultados esperados. Se analizan los aciertos y errores del modelo y se discute la interpretación de los resultados obtenidos.

En el capítulo 6 se discuten en detalle los resultados obtenidos en relación con el estado del arte, la literatura existente y las expectativas planteadas inicialmente. Se profundiza en las implicaciones de los hallazgos y se analiza la relevancia práctica de los resultados y las posibles acciones a emprender en base a estos.

En el capítulo 7 se resumen las conclusiones del proyecto. Se relacionan los resultados con los objetivos del proyecto y se ofrece una visión general de su impacto, viabilidad y relevancia para el centro deportivo. Se añaden propuestas y recomendaciones para posibles investigaciones o mejoras futuras en el modelo predictivo, se revisa el seguimiento de la planificación inicial en cuanto al cumplimiento de los objetivos y plazos establecidos, y se argumenta acerca de los objetivos no conseguidos y su causa.

El capítulo 8 o Glosario define términos técnicos o específicos utilizados en el documento para asegurar una comprensión de los conceptos clave.

La Bibliografía contiene la lista de las fuentes consultadas y referenciadas durante el desarrollo del proyecto.

Los anexos contienen información adicional que complementa el contenido principal del informe, como tablas detalladas, código fuente, gráficos adicionales, entre otros elementos relevantes para la comprensión de esta memoria.

3. Marco teórico y estado del arte del análisis de supervivencia

3.1 Introducción

El análisis de supervivencia es una rama estadística que se centra en estudiar el tiempo hasta que sucede un evento particular. Se utiliza para comprender y modelar la duración de un intervalo de tiempo antes de que se presente un suceso, como en el caso de estudio, la cancelación de una suscripción a un centro deportivo.

El análisis de supervivencia se origina en el ámbito médico con los estudios sobre el tiempo de vida y la supervivencia de pacientes enfermos. Uno de los primeros estudios notables que contribuyó al análisis de supervivencia en el contexto del cáncer fue el trabajo de George W. Corner, quien publicó "*Anatomic Basis of Medical Practice*" en 1958. En este libro, Corner introdujo conceptos como el tiempo de supervivencia y la tasa de supervivencia en este contexto; se quería entender el tiempo de vida después del diagnóstico y cómo diferentes tratamientos afectaron a esa duración.

Poco antes, el método de Kaplan-Meier, ampliamente utilizado en el análisis de supervivencia, fue desarrollado en la década de 1950. Edward L. Kaplan y Paul Meier publicaron independientemente sus trabajos sobre este método en 1958 y 1959, respectivamente. Este método es uno de los primeros y resulta especialmente útil para analizar datos de supervivencia censurados, donde no se observa el evento de interés en todos los individuos durante el período de estudio.

Algo más tarde, en 1972, Sir David Cox, estadístico británico cuyas contribuciones al análisis de supervivencia han sido fundamentales, propone el *Modelo de Regresión de Cox* (1972), también conocido como el *Modelo de Riesgos Proporcionales de Cox*. Este modelo es una herramienta estadística para analizar la relación entre las variables predictoras y el tiempo hasta un evento, como la muerte o la falla de un sistema. El modelo asume que el riesgo relativo entre dos grupos es constante con el tiempo, lo que hace que sea una herramienta versátil y ampliamente utilizada en la investigación biomédica y otros campos.

Hoy en día y con el desarrollo de algoritmos de aprendizaje automático específicos, el análisis de supervivencia se ha convertido en un componente fundamental en los estudios longitudinales de cohortes de pacientes, vitales para

comprender la progresión de enfermedades crónicas, evaluar la eficacia de tratamientos a largo plazo y predecir el riesgo de eventos de salud. Otros ámbitos importantes de aplicación son la investigación epidemiológica, el análisis de confiabilidad sobre el deterioro de estructuras y edificios según sus características, y el estudio del *churn-rate* (tasa de rotación) y el *time-life value* de clientes en el ámbito empresarial.

Así, el análisis de supervivencia aplicado a la retención de clientes, *churn-rate* y tiempo de permanencia en el contexto empresarial y de servicios es un área que ha ganado importancia en la gestión de clientes y marketing. Varios trabajos y enfoques como los de los autores Bruce Hardie o Paul D. Berger abordan la duración de la relación con el cliente y el análisis de churn y han contribuido al desarrollo de métodos de análisis de supervivencia en este ámbito.

El estudio tiempo de permanencia de un abonado a un centro deportivo se enmarca perfectamente en este tipo de análisis ya que este tiempo se verá condicionado por una serie de variables a analizar y la observación o no de la sucesión del evento, en este caso el abandono del centro, vendrá condicionada por la fecha de censura de la observación, es decir, sobre esta sucesión o no del evento, cobra importancia el cuándo.

En el contexto del caso de estudio se pretende aplicar y comparar los resultados obtenidos tras aplicar algoritmos de aprendizaje automático específicos del análisi de supervivencia como *Random Survival Forest*, *Fast Survival SVM*, *Fast Kernel Survival SVM* y *Gradient Boosting Survival Analysis* con los obtenidos tras la aplicación de las implementaciones *Coxnet Survival Analysis* y *Cox PH Survival Analysis* del método tradicional de análisis desarrollado por Cox.

3.1.1 Censura

La característica diferencial del análisis de supervivencia consiste en que, en las observaciones de este tipo de estudios, las variables respuesta, es decir, la ocurrencia o no del evento y el tiempo hasta el mismo, pueden ser desconocidas, ya sea porque el evento no ha tenido lugar antes de la finalización de la observación, porque sucedió en un instante desconocido, o porque el sujeto representado en la observación no ha continuado en el estudio. A estos datos desconocidos se les denomina censurados.

La censura, pues, ocurre cuando no se conoce el tiempo exacto en el que ocurre un evento para algunos individuos al final del estudio.

Hay tres tipos principales de censura:

- Censura por la derecha: Los datos se cortan después de un momento específico T_f a partir del cual se desconoce su evolución, típicamente por finalización del estudio antes de que se dé la ocurrencia. Puede ser por alcance de un tiempo pre-determinado o de un tamaño de muestra positiva necesario. También nos podemos encontrar con censura *aleatoria* cuando el tiempo de censura lo determina un fenómeno aleatorio que tiene lugar durante la consecución del estudio e impide seguir con la observación del individuo hasta el tiempo final; surge generalmente cuando los individuos salen del estudio sin presentar el evento por razones no controladas por el investigador y entonces se dice que se tiene un *análisis de riesgos competitivos*. Si el mecanismo de censura aleatoria es dependiente de los tiempos de falla, se dice que este es una censura *informativa*, ya que se tiene información de los tiempos de falla; en caso contrario es *no informativa*.

- Censura por la izquierda: Se desconoce el momento de ocurrencia del evento por suceder antes del momento T_i de inicio del estudio. Así, el tiempo de censura en este caso será el tiempo de inicio del estudio.

- Censura de intervalo: el evento de interés sucede en un momento desconocido entre dos instantes t_1 y t_2 de observación.

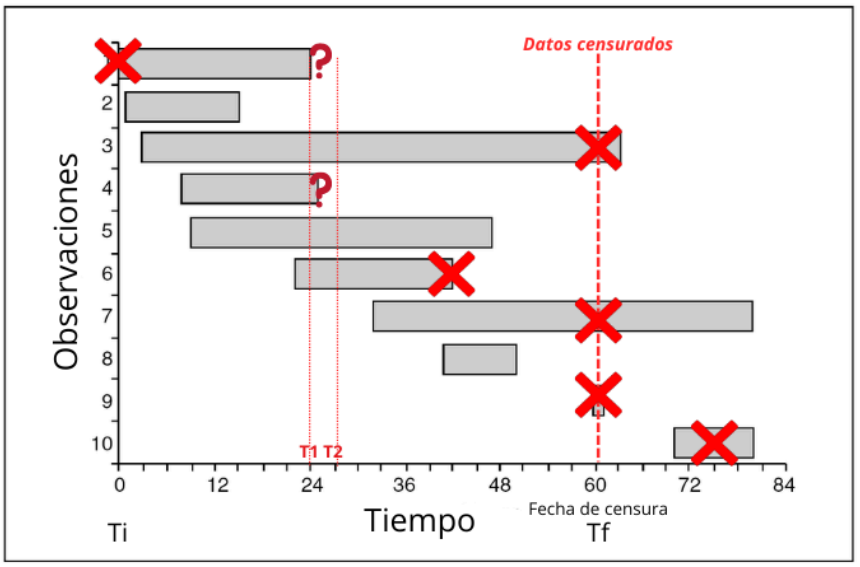


Figura 2. Esquema del tiempo de censura.

En el caso de estudio la censura de los datos es un aspecto clave del modelo escogido; por un lado las observaciones, es decir, cada una de las suscripciones al club, suceden con unos tiempos de inicio y fin distintos, y el dato relevante a estudiar será la duración de ellas; por otro, sea cual fuere el momento de realización del estudio, en aquellos abonados que aún están activos, el evento de baja del centro debe considerarse como censurado y también su momento de ocurrencia. Así, aún siendo el tiempo de permanencia total desconocido para algunos individuos, cada observación aporta al modelo unidades de supervivencia por individuo.

3.1.2 Truncamiento

El truncamiento tiene lugar cuando sólo aquellas observaciones que manifiestan el evento dentro de un intervalo observacional son observados, del resto no se realiza ningún seguimiento y, por tanto, no se obtiene información sobre ellos (no hay información parcial, a diferencia de los datos censurados, que si la hay).

En el estudio, se consideran truncados los ex-abonados que tramitaron la baja antes del periodo de inicio de observación como se explica más adelante, así como los abonados que se inscriban con posterioridad a la fecha de censura. No obstante, estos últimos pueden constituir empresarialmente el foco de interés y sirven como conjunto de datos de validación.

No se consideran aquí truncados los abonados que sin notificación abandonan el centro incumpliendo sus pagos, ya que a los tres meses son dados de baja automáticamente y como tal consta ese registro.

3.1.3 Función de supervivencia

La función de supervivencia $S(t)$ describe la probabilidad de que un evento no haya ocurrido antes de un tiempo específico t, es decir que el individuo sobreviva más allá de t.

Es una función decreciente que varía de 1 (la probabilidad de que no haya sucedido el evento es total y por tanto ningún evento ha ocurrido aún) a 0 (la probabilidad de que que no haya sucedido es nula y por tanto todos los eventos han ocurrido ya) y se puede estimar utilizando técnicas como el Método de Kaplan-Meier o modelos paramétricos como el de Cox.

Matemáticamente la función de supervivencia $S(t)$ se define como:

$$S(t) = P(T > t) = 1 - F_T(t) = \int_t^{\infty} f_T(u)du$$

Donde $F(t)$ se refiere a la función de distribución acumulativa del tiempo de supervivencia. Esta función proporciona la probabilidad de que un evento ocurra antes o exactamente en un tiempo específico según $F(t) = P(T \leq t)$.

Para el caso discreto, T es una variable aleatoria discreta que toma valores $0 < t1 < t2 <$… Entonces la función de supervivencia es:

$$S(t) = P(T > t) = \sum_{t<tj} f(tj)$$

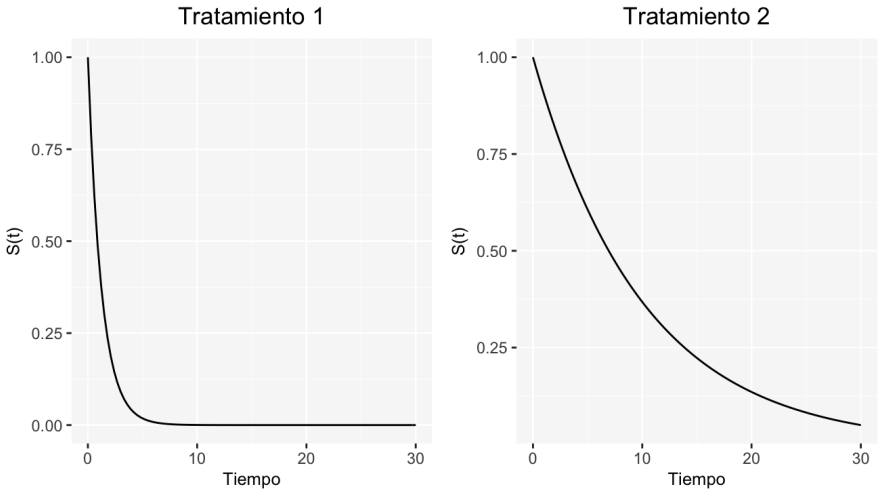


Figura 3. Ejemplo de curvas de supervivencia según distinto valor de la variable predictora.

Así pues, en el contexto de estudio, es posible tras la aplicación de los modelos elegidos conocer la función de supervivencia para cada uno de los abonados al centro, de manera que según su forma e inclinación se puede estimar cuáles de ellos tienen más probabilidad de permanecer en el centro para el periodo de interés. Será habitual, estudiar estas curvas de supervivencia por grupos o colectivos con características comunes y con un rango de variabilidad intra-grupal, de manera que la línea se limita por arriba y por abajo en un área de intérvalo de confianza.

3.1.4 Funciones de riesgo instantáneo y acumulado

La tasa o función de riesgo instantáneo $h(t)$ en un momento t dado representa la probabilidad condicional de que ocurra el evento, dado que no ha ocurrido antes. También llamada tasa de falla condicional en el análisis de confiabilidad, o tasa de mortalidad en demografía, se define pues como la probabilidad de que suceda el

evento durante un intervalo de tiempo infinitesimal suponiendo que el individuo ha sobrevivido hasta el inicio del intervalo.

$$h(t) = \lim_{\alpha \to 0} \frac{1}{\alpha} P(t < T \leq t + \alpha \mid T \geq t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} log(S(t))$$

La función de riesgo juega un papel importante en el análisis de supervivencia; Describe la forma en que cambia la tasa instantánea de fallo de un individuo al paso del tiempo (constante, lineal, exponencial, etc.) y conocerla puede darnos alguna idea sobre la selección del modelo para la distribución del tiempo de supervivencia. No hay un comportamiento "habitual" en la gráfica de $h(t)$, es decir, $h(t)$ puede crecer, decrecer, ser constante o cualquier otro comportamiento.

Integrando $h(t)$ se obtiene la función de riesgo acumulado $H(t)$ que se define como la probabilidad de que el evento suceda antes de un tiempo T y es importante en la medición de la frecuencia con que ocurren los fallos en el tiempo y en el análisis de residuos para el ajuste de algunos modelos. La gráfica de $H(t)$ será siempre creciente por definición. Podemos ver su relación con la función de supervivencia:

$$H(t) = \int_0^t h(u)\, du = -log(S(t))$$

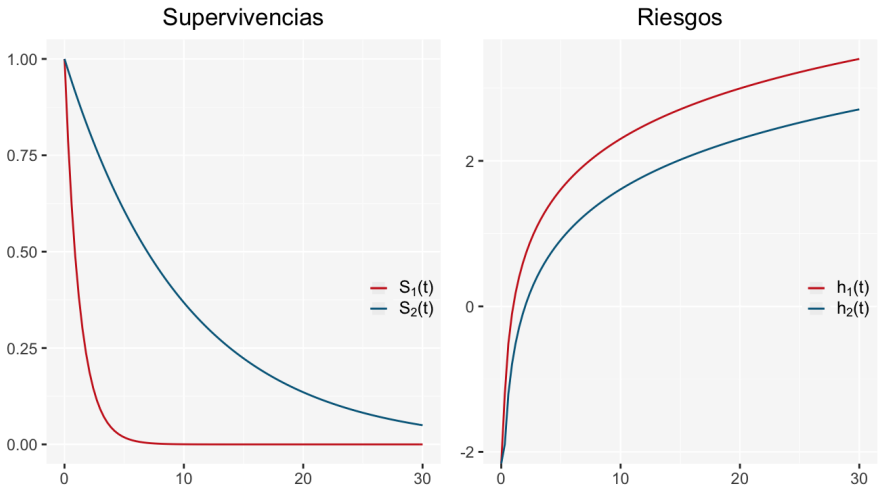


Figura 4. Ejemplo función de supervivencia frente a función de riesgo acumulado.

Los métodos implementados en los algoritmos de estudio permiten obtener y representar gráficamente los valores de funciones. El riesgo acumulado es el que definirá los grupos de clientes sobre los que actuar, identificando aquellos que tengan un riesgo mayor a un umbral establecido para el periodo T de estudio, por ejemplo, aquellos abonados que tienen más de un 80% de probabilidad de ser baja durante los próximos 90 días.

3.2 Metodologías y clasificación

Los modelos estadísticos clásicos de análisis de supervivencia se pueden clasificar en diferentes grupos: no paramétricos, paramétricos, semiparamétricos y. La principal diferencia entre ellos consiste en si se tiene o no en cuenta en la estimación el efecto de las covariables sobre la función de riesgo. Así pues, paramétricos y semiparamétricos especifican la relación entre el riesgo de experimentar el suceso y las covariables. Los no paramétricos, sin embargo, no plantean una relación concreta entre las covariables y la variable dependiente.

- Métodos no paramétricos: Son los métodos estadísticos más utilizados en el análisis de supervivencia y, a diferencia de los métodos paramétricos, no hacen suposiciones explícitas sobre la forma funcional de la función de supervivencia; por tanto, son más flexibles en términos de suposiciones, pero pueden requerir más datos para estimar la función de supervivencia con precisión. Por ejemplo, *Kaplan-Meier* es una técnica fundamental en el análisis de supervivencia; se utiliza para estimar la función de supervivencia en un conjunto de datos censurados, mostrando la probabilidad de que un sujeto sobreviva más allá de cierto tiempo dado,considerando los tiempos de ocurrencia y censura. Es muy útil cuando se estudia el tiempo hasta un evento sin hacer suposiciones sobre la distribución subyacente de los datos.

- Métodos paramétricos: Estos métodos asumen una forma funcional específica para la función de supervivencia. Suelen tener supuestos más estrictos sobre la distribución de los tiempos de supervivencia y pueden ser sensibles a la especificación de ese modelo.

- Métodos semiparamétricos: Combina elementos de ambos enfoques. Por ejemplo, el Modelo de Riesgos Proporcionales de Cox es semiparamétrico, ya que asume una forma específica para el riesgo, pero no especifica la forma de la función de supervivencia en sí misma. Este modelo asume una relación lineal entre las covariables (variables predictoras independientes) y la función de riesgo. El Modelo Addtivo de Aalen's en cambio, utiliza un enfoque aditivo para modelar la relación entre las variables predictoras y el riesgo de un evento a lo largo del tiempo y es una buena alternativa a Cox cuando los riesgos no son proporcionales en el tiempo.

La elección del método depende de varios factores, incluyendo la naturaleza de los datos, la cantidad de información disponible, el cumplimiento de los supuestos

necesarios para su aplicación y la información relevante de estudio. En el ámbito retención de clientes el estudio de interés es conocer la relación entre el tiempo de permanencia del cliente y las variables que lo parametrizan, por tanto serán utilizados los paramétricos y semiparamétricos.

En el modelo de propuesto por Cox o modelo de los riesgos proporcionales que se aplica en el proyecto, la función de riesgo instantáneo $h(t)$ se define por la función del tiempo t, y el conjunto de covariables X_i con la fórmula:

$$h(t, X_i) = h_0(t) \cdot e^{\sum_{I=1}^{n} \beta_i X_i}$$

Aquí a $h_0(t)$ se lo denomina riesgo base y corresponde al riesgo del evento cuando todas las variables tienen valor 0 y es la única parte de la expresión que depende del tiempo. $\beta = (\beta_1, \dots, \beta_n)$ es el vector de parámetros asociados a las covariables es decir, los coeficientes de regresión a estudiar.

$$\frac{hi(t)}{hj(t)} = e^{(X_i - X_j)'\beta}$$

El nombre de riesgos proporcionales se deriva del cociente de las funciones de riesgo de dos individuos; así, el instrumento fundamental de la Regresión de Cox es el *Hazard ratio* o riesgo relativo, que no es más que el cociente entre dos funciones de riesgo. A la expresión anterior se le conoce como riesgo relativo y es constante en el tiempo, cuyo valor depende únicamente de la diferencia entre valores de las covariables de los dos individuos.

El modelo resulta de interés en el estudio de la retención de clientes porque no busca tanto estimar la función $h_0(t)$, idéntica para todos los sujetos, como la relación entre los riesgos de baja entre individuos expuestos a características distintas. Para ello, el modelo parte de una hipótesis fundamental: que los riesgos son proporcionales; al utilizarlo es necesario verificar que se cumple dicha hipótesis y para ello es necesario comprobar que el efecto de cada variable es constante en el tiempo.

Las covariables se expresan como coeficientes en el modelo para determinar su influencia en el riesgo relativo y ello permite una interpretación más sencilla de los efectos de las variables en el tiempo de cancelación; un valor positivo de un coeficiente de regresión comportará un incremento del riesgo conforme aumenta el valor de esa variable, y sucederá inversamente para coeficientes negativos.

3.3 Análisis de supervivencia y aprendizaje automático

Con la incorporación de los métodos y algoritmos de aprendizaje automático al análisis de supervivencia se consigue mayor flexibilidad en los supuestos asumidos y se pueden manejar relaciones complejas entre las variables predictoras y el tiempo de supervivencia.

En el contexto de la inteligencia artificial, los algoritmos de aprendizaje automático utilizados, se pueden clasificar en dos tipos principales de aprendizaje (supervisado y no supervisado) según la existencia o no de etiquetas (información explícita de las salidas deseadas) en los datos de entrenamiento.

El proceso de aprendizaje supervisado implica entrenar al modelo con este conjunto de datos etiquetados, permitiendo que el algoritmo aprenda a mapear las entradas a las salidas correspondientes. El objetivo final es que el modelo pueda generalizar patrones y relaciones dentro de los datos de entrenamiento para realizar predicciones o tomar decisiones sobre nuevos datos que no ha visto previamente.

Hay dos tipos principales de problemas que se abordan con algoritmos de aprendizaje supervisado:

- Problemas de clasificación, en los que el objetivo es predecir a qué categoría o clase pertenece un nuevo ejemplo. En el caso concreto que nos aborda, por ejemplo, determinar si un cliente abandonará o no un servicio.

- Problemas de regresión, en los que el objetivo es predecir un valor numérico basado en variables de entrada. Por ejemplo, predecir el tiempo de permanencia de un cliente en el centro deportivo.

Los algoritmos clásicos de aprendizaje automático tienen problemas para trabajar con la censura propia de los datos del análisis de supervivencia. Por ello, recientemente se han ido desarrollando variantes específicas capaces de trabajar con estos datos. La biblioteca *sci-kit surviva*l utilizada en este trabajo tiene buenas implementaciones de ellos y se encuentra en constante desarrollo. En el proyecto se utilizan variantes de estos tres:

- *Random Survival Forests (RSF):* Random Survival Forests son una extensión de los árboles de decisión para problemas de supervivencia. Aprovechan la capacidad de los árboles para manejar interacciones no lineales y la aleatorización para construir múltiples árboles, lo que permite estimar la función de supervivencia y

evaluar la importancia de las variables predictoras. Tienen la ventaja de capturar relaciones no lineales y manejar interacciones complejas entre variables, sin embargo, la interpretación de estos modelos puede ser más desafiante debido a su complejidad.

- *Gradient Boosting Survival Analysis:* Es una extensión del algoritmo de *Gradient Boosting Machines (GBM)* para problemas de análisis de supervivencia. Estos modelos construyen múltiples árboles de decisión secuenciales, mejorando la predicción de la función de supervivencia. Utiliza técnicas de boosting para construir el modelo predictivo que puede predecir la supervivencia o el tiempo hasta un evento específico utilizando varias variables predictoras.

- *Support Vector Machines for Survival Analysis (SVM-SA*): Las Máquinas de Vectores de Soporte para Análisis de Supervivencia son una extensión de las SVM para manejar datos censurados. Estos modelos buscan encontrar un hiperplano óptimo que separe a los individuos que experimentan un evento del tiempo de supervivencia. Además de la implementación lineal FastSurvivalSVM, la no lineal FastKernelSurvivalSVM utiliza un kernel para calcular la similitud entre los datos. Esto permite aprovechar la información estructural de los mismos para mejorar el rendimiento del modelo.

Algunas otras opciones interesantes que quedan fuera del ámbito de estudio podrían ser:

- *Survival Neural Networks:* Las redes neuronales, especialmente las redes neuronales profundas, se han aplicado al análisis de supervivencia para capturar relaciones no lineales y complejas entre variables predictoras y el tiempo de ocurrencia del evento de interés. Técnicas más recientes involucran la aplicación de arquitecturas de aprendizaje profundo, como redes neuronales recurrentes (RNN) o redes neuronales convolucionales (CNN), para capturar patrones temporales complejos y realizar pronósticos en datos de supervivencia. Se descartan en parte por la falta de explicabilidad de sus resultados y, por tanto, de aplicabilidad para el caso.

- *Competing Risks Regression Models:* Se utilizan cuando hay múltiples eventos competitivos que podrían ocurrir, lo que implica la posibilidad de que un evento ocurra antes o en lugar de otro. Este tipo de modelos consideran y modelan múltiples resultados posibles. Recordemos los casos mencionados de censura aleatoria; como se comenta anteriormente, en el diseño de recolección de datos

para el caso se han descartado riesgos competitivos como la censura por otro motivo que no sea la tramitación de la baja.

- *Flexible Regression Models:* Tales como los modelos basados en splines, GAM (Generalized Additive Models), o modelos con penalizaciones (como LASSO o Ridge), permiten capturar relaciones complejas entre las variables y el tiempo de supervivencia.

3.4. Evaluación de modelos de análisis de supervivencia

Las características propias del análisis de supervivencia y la censura de sus datos hace que en sustitución de las métricas habituales de evaluación de rendimiento como la precisión, accuracy, recall, especificidad o F1-score, sea necesario definir métricas propias de estos modelos que reflejen apropiadamente la bondad de los mismos.

En 1982, Harrell, F.E propone como métrica de evaluación para modelos de análisis de supervivencia el conocido como índice de concordancia de Harrell o *C-Index*; una métrica útil para comparar diferentes modelos y determinar cuál tiene un mejor rendimiento predictivo en términos de capacidad para ordenar adecuadamente los tiempos de supervivencia. Se basa en la capacidad del modelo para ordenar adecuadamente los tiempos de supervivencia entre pares de individuos.

Para calcular el *C-Index* en el contexto del análisis de supervivencia, se comparan pares de individuos. Por cada par de individuos, el modelo predice cuál de los dos sobrevivirá más tiempo y se compara con los datos reales. Si el modelo predice correctamente el orden de supervivencia, se cuenta como un acierto. El *C-Index* se calcula dividiendo el número total de aciertos por el número total de comparaciones posibles entre pares de individuos. La validación cruzada o k-fold que se aplica en este proyecto divide el conjunto de datos en múltiples subconjuntos de entrenamiento y test y evalúa el *C-índex* en cada iteración; el promedio de los valores obtenidos permite la generalización del rendimiento del modelo.

La interpretación del *C-Index* es similar a la de otras métricas en problemas de clasificación. Un *C-Index* de 1.0 indica un modelo perfecto que ordena correctamente todos los pares de sujetos según su tiempo de supervivencia. Por otro lado, un *C-Index* de 0.5 indica un rendimiento similar al azar, donde el modelo no tiene capacidad para discriminar entre los tiempos de supervivencia.

En su artículo sobre evaluación de modelos de análisis de supervivencia, Cosimo Albanese, N. (2022) argumenta como en los proyectos de clasificación binaria, la curva ROC representa la tasa de verdaderos positivos (TPR = TP / (TP + FN)) frente a la tasa de falsos positivos (FPR = FP / (FP + TN)) en diferentes umbrales de clasificación. El Área Bajo la Curva (AUC) mide el área debajo de la curva ROC y se puede interpretar como la probabilidad de que el modelo clasifique mejor un ejemplo positivo aleatorio que un ejemplo negativo aleatorio. En el análisis de supervivencia, en cambio, las tasas de verdaderos y falsos positivos son dependientes del tiempo por diseño del caso; los clientes del gimnasio activos (verdaderos negativos en el evento 'darse de baja') experimentan un evento adverso con el tiempo (convirtiéndose en verdaderos positivos). Por lo tanto, la expresión del AUC se vuelve dinámica y se debe evaluar calculándose para varios momentos concretos de tiempo.

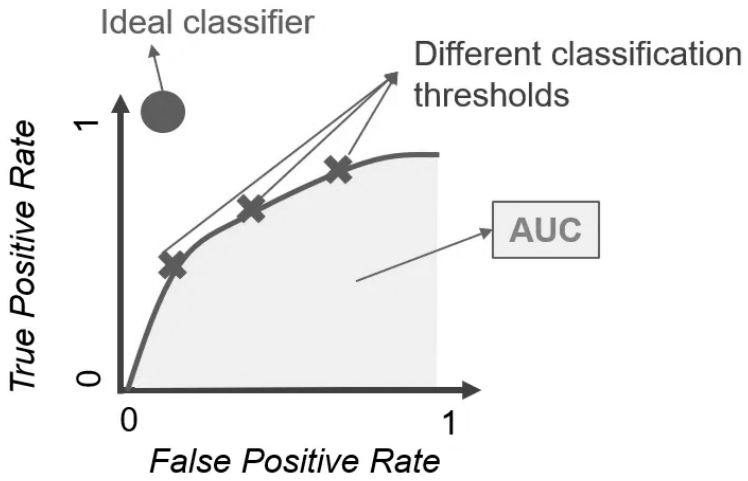


Figura 5. Ejemplo de aplicación de la gráfica AUC-ROC.

En el proyecto se asume que el coste de un falso negativo, es decir, la pérdida de un abonado que se hubiera podido detectar como susceptible de abandonar el centro, es superior a la de un falso positivo, es decir, al coste de medidas retentivas como descuentos o personalización en el servicio para un abonado que igualmente no iba a marcharse; en cualquier caso, supondrá seguramente un aumento en la percepción de la calidad del servicio recibido.

3.5 Análisis de supervivencia en el caso de estudio

El tiempo de permanencia de un cliente en un servicio (un centro deportivo en el caso de estudio) se puede modelar, pues, como un caso de análisis de

supervivencia, en el que el evento a estudiar será la cancelación de la membresía o abandono del centro por parte del cliente.

Para este análisis, las variables predictoras abarcan aspectos demográficos, como la edad, el género, la nacionalidad o el código postal de residencia (cercanía geográfica al centro); de uso, como la frecuencia de visita al centro, franja horaria y dia de la semana habitual, la actividad preferida o el porcentaje y anticipación de uso de la aplicación de reservas; y relacionados con la membresía, como el tipo de cuota, su precio mensual, las veces que estuvo apuntado anteriormente, el número de meses pagados por anticipado o el descuento aplicado.

En este proyecto se enfrenta la capacidad predictora de los resultados obtenidos tras la aplicación sobre estos datos de implementaciones del modelo de riesgos proporcionales de Cox, con los obtenidos mediante el uso de algoritmos de aprendizaje automático especializados de *Random Survival Forests (RSF), Gradient Boosting Survival Analysis y Support Vector Machines for Survival Analysis (SVM-SA*).

Para ello, se observan los datos de los datos en una ventana de tiempo igual al periodo desde el inicio de funcionamiento de la aplicación de reservas del centro hasta la fecha de captura de los datos, que se considera la fecha de censura. Todas las fechas se referencian a esta ventana. La sucesión del evento a estudiar será la baja del centro del abonado y el tiempo que transcurre hasta ella, evaluando tanto el riesgo de que suceda antes de un periodo de tiempo hasta T (riesgo acumulado) como la probabilidad de que el abonado permanezca inscrito más allá de T.

4. Desarrollo del proyecto

4.1 Fase 1a

4.1.1 Consideraciones sobre protección de datos

En España, la normativa principal que regula la protección de datos personales es la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD), que desarrolla el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, conocido como el Reglamento General de Protección de Datos (GDPR).

En los datos seleccionados para el proyecto se garantiza que se cumplen los aspectos clave de esta normativa a tener en cuenta al trabajar con datos personales en el ámbito comercial en España:

- Consentimiento del titular de los datos: los datos con los que se trabaja pasan el filtrado de haber firmado y aceptado en el momento de inscripción el consentimiento de uso y protección de los datos personales por parte del interesado, especificando el uso administrativo y de análisis que se hará de ellos.

- Finalidad y legitimidad del tratamiento: el documento tiene una base legal para el tratamiento de datos personales, y el cual se realiza para fines legítimos y específicos de análisis y mejora del servicio, y los abonados están informados de ello.

- Principio de minimización de datos: los datos que son utilizados son estrictamente demográficos y de consumos, eliminando de los conjuntos de datos antes de subirlos a la plataforma otros de carácter personal como el nombre de usuario o medios de pago, irrelevantes para el estudio.

- Derechos de los interesados: las personas cuyos datos se están procesando tienen el derecho de acceso, rectificación, supresión, oposición, limitación del tratamiento y portabilidad de datos de los que dispone el centro y en cualquier momento pueden hacer uso de ellos.

El centro cumple, además, con el resto de medidas relativas a la Seguridad de datos consistentes en implementar medidas técnicas y organizativas adecuadas para

garantizar la seguridad de los datos personales en su almacenaje y acceso, con una definición clara y restringida de permisos por roles, y supervisadas por la figura del Delegado de Protección de Datos de la compañía, con el conocimiento del cual se realiza este estudio.

4.1.2 Importación de archivos y modificaciones iniciales

De las bases de datos de CRM del centro y de la aplicación que gestiona la reserva de actividades se ha solicitado al departamento de IT la información con las variables que se quieren considerar para el estudio. Los trece archivos en formato .xls y .csv a partir de los que trabaja el proyecto son leídos e incorporados como dataframes pandas al proyecto en la plataforma de Google Colab.

Se cargan también las librerías que serán necesarias para su análisis; scikit-learn, que contiene las funciones necesarias para el entrenamiento, sci-kit survival con las implementaciones de los algoritmos estudiados, matplotlib y seaborn para la representación gráfica, math, pandas, numpy para las operaciones matriciales y timedelta para la correcta gestión de variables de tiempo.

Sobre los datasets iniciales se hacen operaciones básicas preliminares como eliminar registros no pertenecientes a clientes (personal), registros inconsistentes como con fecha de baja anterior a fecha de alta, renombrado y reformateado de variables para conseguir coherencia e interpretabilidad en el conjunto.

Se identifican también inicialmente dos variables globales que serán la fecha de censura, a partir de la cual las observaciones posteriores no son consideradas, y la fecha inicio de reservas, momento en el que se incorpora a la gestión del centro una aplicación de reservas, que permite disponer de más información sobre los usos de los abonados en el centro pero que introduce un elemento diferencial entre los registros anteriores y posteriores a esa fecha.

Cabe destacar que en una primera iteración de este proyecto, con la intención de incorporar la máxima información posible al proyecto, se utilizaron algunos datasets que presentaban datos parciales, únicamente de abonados inscritos, que añadían campos de información. Esto introduce sesgos en el modelo dado que los clientes que tenían un valor no-nulo en esos campos, por ejemplo, ‘documentación identificativa’ o ‘medio de pago’, eran precisamente registros que tenían valor en la variable respuesta 0 como no-sucesión del evento. Una vez detectado el error, se descartan estos dataset

y se recupera información adicional sobre la totalidad de registros, reformulando la construcción del dataset final a partir de solo 6 datasets pero con la información más relevante completa.

4.1.3 Interpretación de datasets y variables

Se dispone en el proyecto de trece datasets extraídos en el momento 'fecha de censura' del datawarehouse de la instalación. Los datasets contienen información relativa a abonados, compras, usos, y servicios. Se terminan utilizando siete de ellos, por contener el resto información sesgada.

En el anexo I se detalla la composición de estos datasets y sus variables, enumerando contenido, nombre y descripción, su tipo y la relación de variables de otros datasets que referencian el mismo objeto. El detalle de los formatos de cada variable, junto con su proceso de reformateo y renombrado, se puede ver en la sección específica del cuaderno del proyecto.

No se dispone de información completa de determinados registros y se toma una serie de decisiones en la consideración de valores nulos y en la ingeniería de variables para asegurar la consistencia de los datos basándose en el conocimiento del dominio en el que fueron tomado y que se detallan a continuación:

- El dataframe principal `data_clientes` constituye la base sobre la que se construye el conjunto de datos de trabajo. Contiene datos demográficos con un registro por cada inscripción con id de '`Abonat`', '`CentreOrigen`', '`DataAlta`' y '`DataBaixa`', '`Modalitat`' de cuota y '`ModalitatPagament`', '`EdatActual`', '`Nacionalitat`', '`Sexe`', '`CodiPostal`' e indicador de '`BaixaDefinitiva`'. Un mismo abonado puede haber realizado inscripciones sucesivas, y por lo tanto cada registro se identificará de manera única por la combinación del número de abonado y su fecha de alta.

- Ante la enorme variedad de tipos distintos de '`Modalitat`' de cuota que solo contribuye a generar ruido e introduciría excesiva variabilidad y sobreajuste en el modelo, las cuotas se categorizan en '`CategoriaModalitat`' por nombre según su tipo en *horario completo*, de *mañana*, de *tarde*, *VIP*, de *empresa*, etc. De la misma manera y por el mismo motivo se procede con el 'código' postal categorizándolo según '`distancia`' al centro.

- Las inscripciones y cancelaciones a un centro suelen ser recurrentes, siendo frecuente el retorno de antiguos abonados, y tienen un fuerte componente cíclico a lo largo del año. Para capturar esta información y dado que las fechas concretas posteriormente se transformaran en periodos, se construye una variable con el '**num_inscripcions**' anteriores para cada registro y se categoriza en una variable categórica nueva la fecha de inscripción según el 'Mes_de_Alta'.

- El dataframe data_entradesclaret contiene registros vinculados a cada acceso histórico al centro. Contienen 'DataHora' de acceso, e id de 'Abonat' como variables de interés. Las entradas de no abonados, abonados de otros centros o personal son descartadas al hacerse el merge por id de abonado. Incorporando la fecha de alta y baja desde el otro dataset y asegurando su consistencia, se extrae información en una tabla resumen con el recuento de '**numero_accesos**', '**freq_activa**' semanal, 'franja_habitual' y '**dia_habitual**' y el primer y último acceso registrados en cada inscripción, así como el número de accesos en los últimos 180, 90 y 30 días antes de la cancelación o la fecha de censura según corresponda. Posteriormente, estos recuentos de accesos son reinterpretados como frecuencias de acceso semanal en los periodos observados para facilitar su comparación y estudio.

- Los valores nulos de estas nuevas variables creados en los registros con el merge por la izquierda corresponden, por diseño y dado que el registro de entrada es obligatorio en el centro, a abonados que tras inscribirse nunca llegaron a acudir al centro, y como tal se les da esta interpretación a esos valores nulos, manteniendo una vez más la coherencia del conjunto; número de entradas y frecuencias son completados con 0, las fechas de primer y último acceso son equiparada a las de baja y alta respectivamente y las categóricas de día y franja habitual son completadas con una nueva de '*sin accesos'*.

- Los dataframes data_formaspago, data_altes_i_baixes y data_consum_altres añaden al conjunto por número de abonado y fecha de alta, variables adicionales de 'duración' y 'descuento aplicable' del pack o forma elegida de pago, 'fecha de solicitud de cancelación', e 'indicador de consumos de servicios extras' adicionales respectivamente. La 'fecha de solicitud de cancelación', obligatoria y previa a la baja del cliente, y por tanto al tiempo de sucesión del evento, se descartó del estudio ya que evidentemente está estrechamente correlacionada con la variable baja pero no aporta información útil.

Los nulos generados por la no aparición en el listado de consumidores son identificados mediante un valor 0 como no consumidores.

- A partir del dataframe `data_darrer_rebut` se incorpora la fecha de facturación e importe del último recibo bancario, a partir de los cuales se determinará el '`percent_consumit`' , el '`PreuBase`' y el '`CostMensual`' del abono consumido'. Los nulos generados en esta fusión se interpretan como abonados que pagan en recepción en el momento del alta y esa es la fecha a considerar.

- Según el importe del recibo, la duración del pack contratado y su descuento aplicable, se estimará el precio mensual de cuota, y, a partir de los abonados a los que no se les aplica ningún descuento por franja de edad se estiman los precios de las mensualidades y los distintos descuentos aplicables a cada franja de edad en cada modalidad de cuota. Así, en un proceso inverso y algo complejo, si la cuota mensual es desconocida en un registro determinado, se estima el precio tomando como valor el promedio de los valores de los registros con las mismas características de edad y cuota.

- De `data_reservas`, que contiene un registro por cada reserva de actividad, se extrae también una tabla resumen de las reservas dentro del periodo de inscripción con la '`antelación`' de la reserva, la '`CategoríaActividad`' preferida, el porcentaje de uso de cada uno de los cuatro tipos de servicios y el porcentaje de asistencias y cancelaciones.

- Puesto que la incorporación de la aplicación es muy posterior a la inauguración del centro, `data_reservas` tiene un periodo de observación distinto y más breve que el resto de datasets. Sin embargo, su información es relevante así que genera dos aproximaciones distintas al estudio: una en que se tiene en cuenta la totalidad de los registros con información parcial sobre reservas y un elevado número de imputados con knn para los registros faltantes, con la aleatoriedad que ello comporta; y una donde solo se tiene en cuenta aquellos registros de inscripciones que comparten total o parcialmente la convivencia con la app. Para perder la menor información posible en este truncamiento en el que dejan de ser considerados los abonados con 'fecha de baja' anteriores a la fecha de inicio de reservas, se añade una variable de 'porcentaje de app' que refleja el porcentaje de solapamiento entre el periodo de existencia de la app y la inscripción. Los valores nulos de los campos generados a partir de `data_reservas` en registros con este solapamiento son considerados como abonados que no han realizado

reservas de manera similar a como se hizo con las entradas. Gracias al elevado número de muestras disponibles este truncamiento no termina afectando al rendimiento del modelo como se observa al final del proyecto.

- Para el correcto funcionamiento de los modelos de *ML* (*machine learning* / aprendizaje automático) que serán aplicados, es necesario convertir el máximo posible de estas variables a formato numérico o categórico. Las variables con formato de fecha datetime pasan a ser interpretadas como períodos; `'dias_totales'` de inscripción referencia la fecha de baja a la de alta y constituye la variable respuesta del proyecto, `'dias_primer_uso'` referencia la fecha de primer uso también a la fecha de alta y `'dias_sin_venir'` referencia el último acceso a la fecha de baja o a la de censura según corresponda.

Estas decisiones son producto tanto de la interpretación del dominio del proyecto a partir de su conocimiento como de la necesidad de cumplir los requisitos necesarios para el buen funcionamiento del entrenamiento y la evaluación de los modelos.

4.1.4 Construcción del dataset final

Con el proceso detallado en el apartado anterior se obtiene un conjunto de datos `data_final` con 41844 registros y 65 variables.

```
print(data_final.shape)
print(lista_variables)
```

(41844, 65)

['Abonat', 'CentreOrigen', 'DataAlta', 'DataBaixa', 'BaixaDefinitiva', 'ModalitatCodi', 'ModalitatNom', 'ModalitatPagament', 'CaractModPagament', 'Sexe', 'EdatActual', 'CodiPostal', 'Nacionalitat', 'primer_acces', 'ultim_acces', 'numero_accesos', 'dia_habitual', 'franja_habitual', 'entrades_darrers_30', 'entrades_darrers_90', 'entrades_darrers_180', 'Data_registre', 'ServiciosExtra', 'DataFacturacio', 'ImportAbsolut', 'numero_reservas', 'antelacion_promedio', 'Attended', 'Cancelled', 'NotAttended', 'Fitness', 'PersTraining', 'GroupExercise', 'Nutrition', 'act_preferida', 'CategoriaModalitat', 'MesesModalidadPago', 'Descuento', 'DiasModalitatPago', 'Mes_de_Alta', 'dias_totales', 'period_primer_acces', 'period_solic_baixa', 'period_desde_fact', 'percent_consumit', 'dias_sin_venir', 'ratio_inactivo', 'ratio_primer_us', 'num_inscripcions', 'Edat_Alta', 'Distancia', 'freq_activa', 'freq_30dias', 'freq_90dias', 'freq_180dias', 'preuMB',

'EdatFacturacio', 'PreuBase', 'PreuBasePromig', 'DescEdat', 'DescEdatPromig', 'DescEdatPromigPromig', 'PreuBasePromigPromig', 'CostMensual', 'CostDiari']

Se eliminan las variables auxiliares para la construcción del conjunto para evitar la redundancia y correlación entre variables; se prescinde de las fechas y el resto de variables reinterpretadas en otras numéricas o categóricas y descritas anteriormente (código postal, nombre de la modalidad y de la forma de pago, número de accesos, importe del recibo, coste diario). El conjunto resultante `data_basic` constituye la base para una de las dos aproximaciones del estudio con los 41844 registros y 38 variables, que incluyen la totalidad de los registros de inscripciones y la información de reservas, aún contando con un porcentaje elevado de valores nulos en los registros que no convivieron con la aplicación.

```
print(data_basic.shape)
print(lista_variables)
```

(41844, 38)

['CentreOrigen', 'BaixaDefinitiva', 'Sexe', 'Nacionalitat', 'numero_accesos', 'dia_habitual', 'franja_habitual', 'ServiciosExtra', 'numero_reservas', 'antelacion_promedio', 'Attended', 'Cancelled', 'NotAttended', 'Fitness', 'PersTraining', 'GroupExercise', 'Nutrition', 'act_preferida', 'CategoriaModalitat', 'Descuento', 'DiasModalitatPago', 'Mes_de_Alta', 'dias_totales', 'period_primer_acces', 'period_solic_baixa', 'percent_consumit', 'dias_sin_venir', 'ratio_inactivo', 'ratio_primer_us', 'num_inscripcions', 'Edat_Alta', 'Distancia', 'freq_activa', 'freq_30dias', 'freq_90dias', 'freq_180dias', 'PreuBase', 'CostMensual']

El conjunto de datos `data_sin_reservas` prescinde de la información de reservas para la totalidad de los registros y queda excluido de este estudio pero disponible para futuras ampliaciones.

Finalmente, el conjunto de datos `data_reducido` constituye la segunda aproximación al ejercicio, prescindiendo, como se comenta anteriormente, de los registros con fecha de baja anterior a la existencia de la app pero sin la necesidad de introducir aleatoriedad en la variables de reservas. El entrenamiento con sus 9456 registros son suficientes para conseguir un buen rendimiento del modelo.

```
print(data_reducido.shape)
print(lista_variables)
```

(9456, 39)

['CentreOrigen', 'BaixaDefinitiva', 'Sexe', 'Nacionalitat', 'numero_accesos', 'dia_habitual', 'franja_habitual', 'ServiciosExtra', 'numero_reservas', 'antelacion_promedio', 'Attended', 'Cancelled', 'NotAttended', 'Fitness', 'PersTraining', 'GroupExercise', 'Nutrition', 'act_preferida', 'CategoriaModalitat', 'Descuento', 'DiasModalitatPago', 'Mes_de_Alta', 'dias_totales', 'period_primer_acces', 'period_solic_baixa', 'percent_consumit', 'dias_sin_venir', 'ratio_inactivo', 'ratio_primer_us', 'num_inscripcions', 'Edat_Alta', 'Distancia', 'freq_activa', 'freq_30dias', 'freq_90dias', 'freq_180dias', 'PreuBase', 'CostMensual', 'existe_app']

4.1.5 Análisis descriptivo de los datos

Con la función `df.describe()` se observa una visualización inicial de las estadísticas de los datos para comprobar su coherencia.

En los diagramas de barra se observa, tanto la distribución de las ocurrencias para cada valor posible de la categoría, como la relación de los valores que toma cada variable categórica con el porcentaje de valores 0 o 1 de la variable respuesta `BaixaDefinitiva`.

En la gráfica siguiente se observa la distribución de las ocurrencias según el valor de para cada una de las variables categóricas, y estos a su vez, ordenados decrecientemente por el porcentaje de aparición del valor 1.

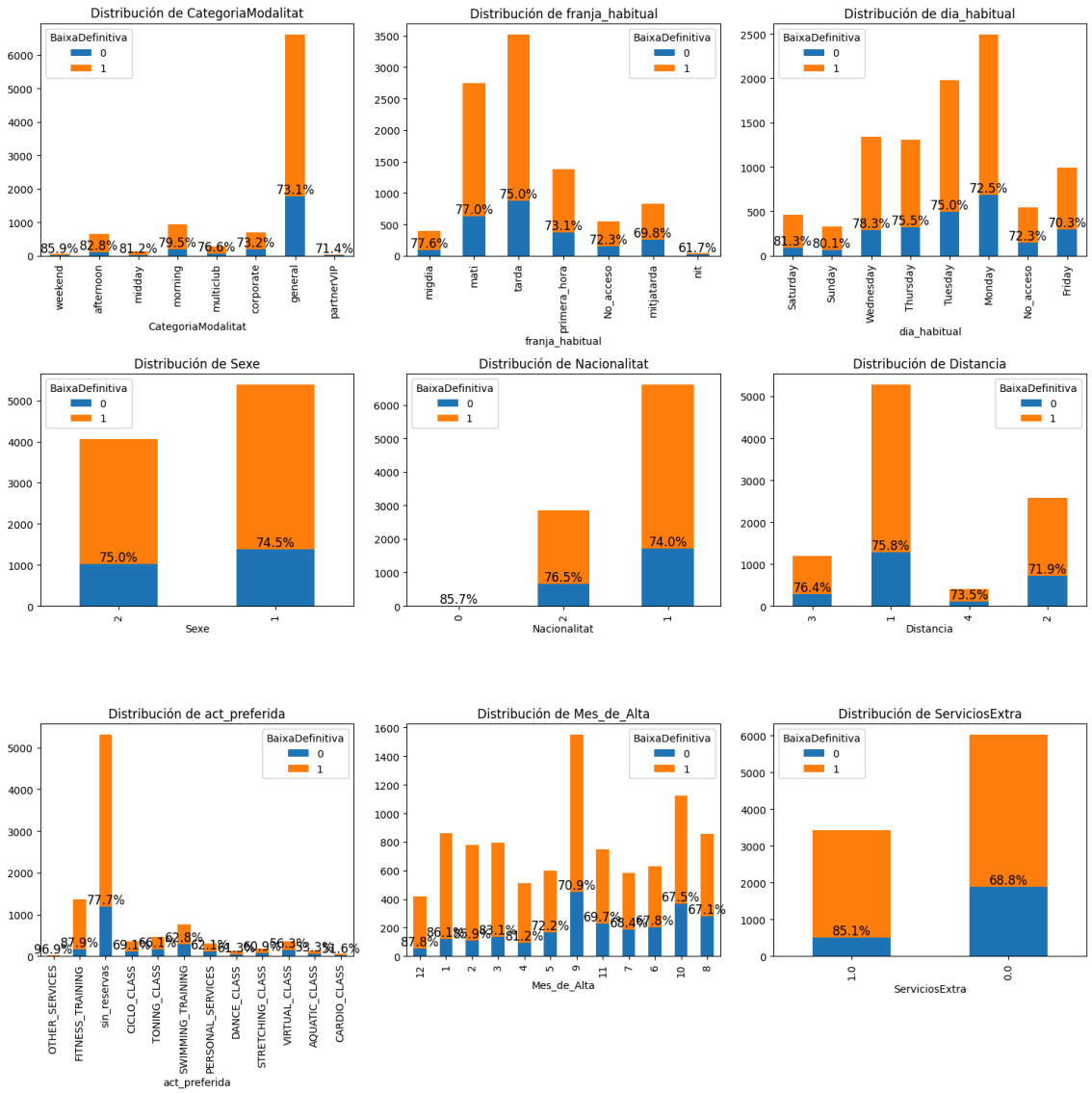


Figura 6. Diagramas de barras de las variables categóricas.

Así, por ejemplo, se observa que las mujeres, a las que se asigna el valor 1, porcentualmente cursan más bajas que los hombres, los usuarios de fitness más que los de clases o que los de otra nacionalidad tienen un mayor porcentaje de cancelaciones. Estos diagramas muestran también proporción de bajas entre los que toman en la categoría el valor para valores faltantes.

También el tipo de membresía parece tener una fuerte relación, y es obvio que abonados VIP difícilmente cursarán baja. Además las cuotas más limitadas en horario también disminuyen la vinculación con el centro. El horario o día de acceso también parece tener relación, siendo los que entrenan durante la semana o a última hora de la tarde los que menos rotan.

Para visualizar la relación entre los valores que toma `BaixaDefinitiva` y las variables continuas más relevantes se testean varias opciones como histogramas, gráficos de violín, de caja o curvas de densidad.

`kdeplot` es una función en la biblioteca Seaborn, que se utiliza para trazar la estimación de la densidad del kernel (KDE) de una variable. El KDE es una forma de estimar la función de densidad de probabilidad de una variable aleatoria continua de manera que cuando se aplica `sns.kdeplot` a un conjunto de datos, Seaborn ajusta una curva suave (kernel) sobre el histograma de los datos. Esta curva suave es una representación de la distribución de probabilidad continua de los datos.

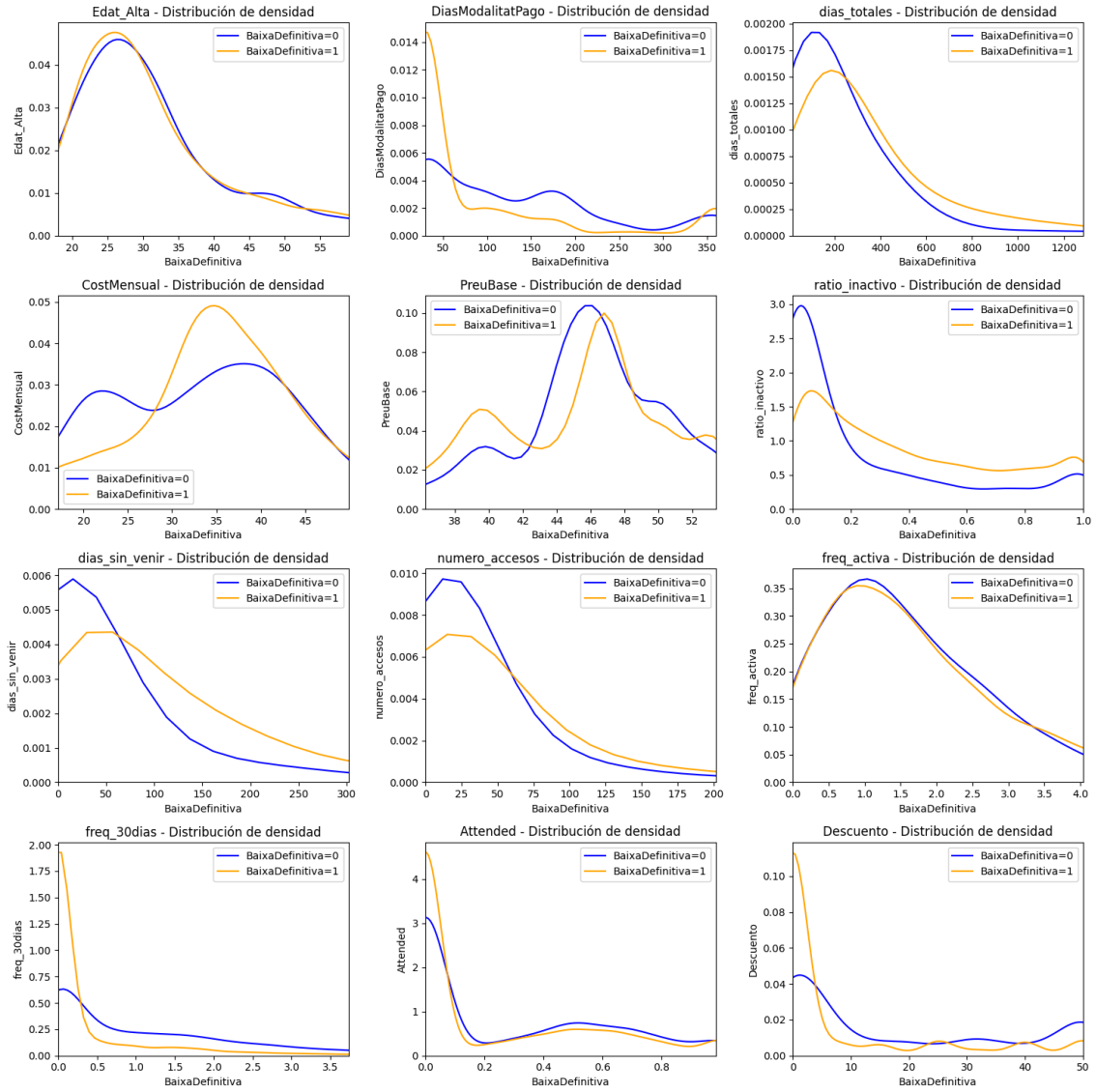


Figura 7. Distribuciones de probabilidad de las variables numéricas.

Para cada variable se observa en esta gráfica la relación entre el valor que toma `BaixaDefinitiva` y los valores de la variable. Así, por ejemplo, la frecuencia en activo que tuvo un abonado o la edad a la que inscribió no parecen ser factores relacionados con la decisión de darse de baja, y sin embargo se comprueba que los abonados con distintos descuentos o dias pagados tienen comportamientos distintos a la hora de darse de baja; los que pagan modalidad cortas y sin descuentos obviamente son mucho más rotacionales.

La matriz de correlación indica en un mapa de calor el grado de correlación entre las variables y, por tanto, el grado de información que es redundante. Colores claros indican un grado de correlación positivo, y colores oscuros, negativo. En la tabla de correlaciones se puede observar el valor concreto de los coeficientes de correlación

de cada par de variables, pero el propio análisis visual permite extrapolar conclusiones.

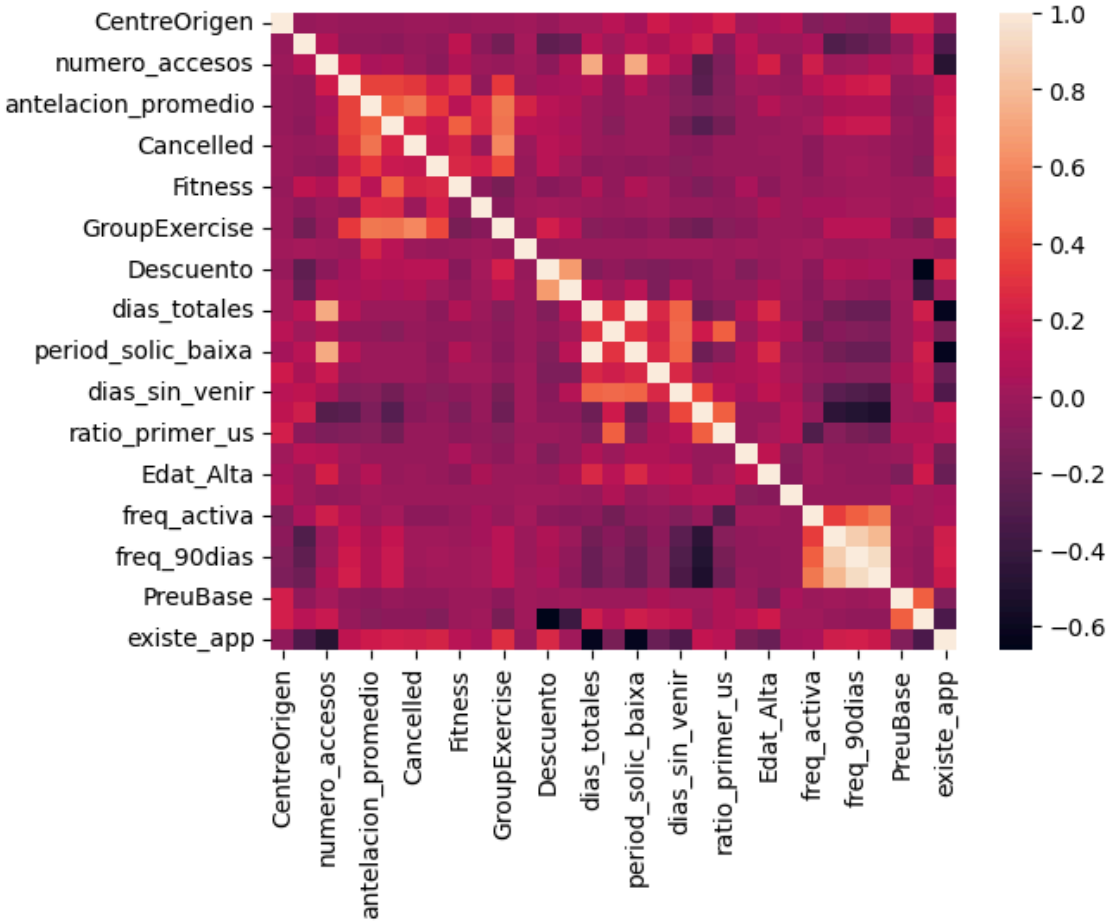


Figura 8. Mapa de calor de la matriz de correlación.

Concretamente, destacan los siguientes grupos de variables correlacionadas:

- La variable respuesta `'dias_totales'` tiene como se comenta antes un índice de correlación positivo fuerte y explicable con `'period_solic_baixa'` ; se ha mantenido aquí por el interés de su visibilización pero se prescinde de esta variable en el entrenamiento final. Tiene también lógicamente una correlación fuerte con `'numero_accesos'` ya que a más tiempo en el centro es lógico que consten mas accesos; es por ello que es más interesante como métrica de evaluación de usos la `'freq_activa'` semanal y se podría prescindir de la anterior si entorpece la convergencia del modelado. Las tres, tienen un grado de correlación negativa con el porcentaje de tiempo `'existe_app'` de convivencia con la aplicación, precisamente por la forma en que es calculada esta variable.

- Las frecuencias en los últimos 180, 90 y 30 días tienen una fuerte correlación

positiva entre ellas; se incorporaron pensando en que podría ser interesante considerar cambios en la frecuencia de usos indicativos de una baja inminente. Si no converge, podría prescindirse del par más correlacionado o sustituirlos por un coeficiente entre ellos. Además, y lógicamente tienen una correlación inversa con `'ratio_inactiva'` que es indicador de los días recientes sin venir respecto al total, precisamente en los que se está evaluando esas frecuencias de uso.

- Por último, es obvio también que el `'descuento'` aplicado tendrá una correlación negativa con el `'cost_mensual'`.

Si el modelo tiene muchas variables correlacionadas puede haber problemas como la multicolinealidad, que afecta a la interpretación del modelo y puede conducir a coeficientes poco fiables o afectar a la convergencia de algunos algoritmos de aprendizaje automático. Si se aplica análisis de componentes principales es posible reducir la dimensionalidad del conjunto reduciendo el número de variables a otras que contienen toda la información, pero para el caso de estudio se perdería explicabilidad de los resultados.

Tras el análisis de la correlación y ante fallos de convergencia en el entrenamiento, tal y como se explicó antes, se optó por la transformación de variables y una mínima selección de características ya que las cuatro familias de algoritmos escogidas en el trabajo son robustas ante la multicolinealidad.

También los diagramas de dispersión o scatter plots, son útiles para obtener información sobre la correlación entre dos variables concretas añadiendo la información de la variable respuesta. Si los puntos en el diagrama de dispersión siguen una tendencia ascendente (o descendente), esto sugiere una correlación positiva (o negativa). Si no hay una tendencia clara en los puntos y se distribuyen de manera aparentemente aleatoria, esto sugiere una ausencia de correlación. La concentración y dispersión de los puntos en torno a la línea de tendencia proporciona información sobre la fuerza de la correlación; un grupo más compacto de puntos alrededor de la línea sugiere una correlación más fuerte. Los diagramas de dispersión también pueden revelar la presencia de valores atípicos, que pueden afectar la interpretación de la correlación.

Se muestra a continuación el diagrama de dispersión de algunas variables numéricas levemente correlacionadas . Así en la primera gráfica se observa el comportamiento de la variable numérica respuesta `'dias_totales'` (días de

permanencia en el centro o hasta la fecha de censura) en función de la ‘Edat_Alta’; vemos que existe diferencia en el grado de aumento de la duración de las suscripciones conforme a la edad de los abonados que están activos respecto respecto a los que se dan de baja.

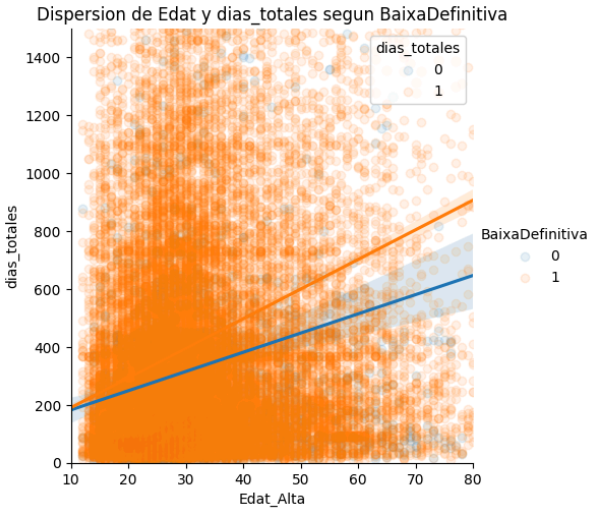


Figura 9. Diagrama de dispersión de dos variables numéricas.

En la segunda gráfica se observa un comportamiento distinto entre el par de variables ‘numero_reservas’ y ‘Edat_Alta’; el grado de correlación aquí más débil porque los puntos están más dispersos respecto a la línea de regresión y además aquí existe una relación opuesta entre ambas variables para los abonados que se dieron de baja; la gente más mayor que no se da de baja hace más uso del sistema de reservas, en cambio para le gente joven no parece ser tan determinante el uso de la aplicación, y eso es fácilmente observable en los hábitos de uso de las dos poblaciones.

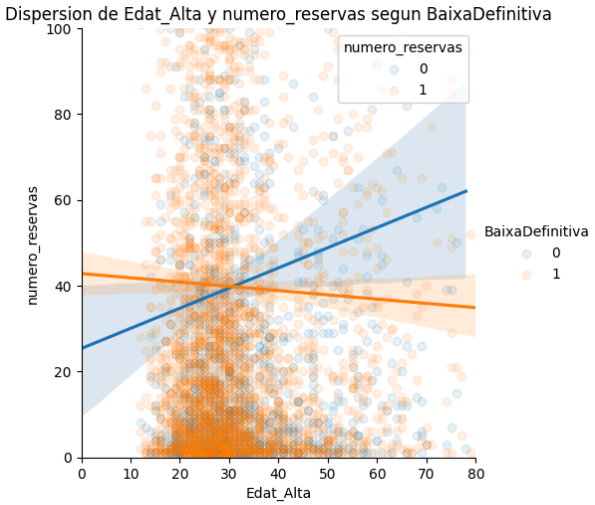


Figura 10. Diagrama de dispersión de dos variables numéricas.

4.1.6 Estudio e imputación de valores nulos

En la primera iteración del proyecto, por la forma de adquisición de los datos y de construcción del dataset, existía una elevada cantidad de valores nulos que se trataron según el caso.

En el diseño del conjunto final sobre el que se termina realizando el proyecto se tiene ya se tiene en cuenta:

- A los valores ausentes susceptibles de ser interpretados se les asigna un valor durante el diseño; por ejemplo, a aquellos resultantes del *merge* con el registro de entradas en las variables 'numero_accesos', que son nulos por no haber realizado ningún acceso el abonado, se entiende que deben tomar valor 0 e igualmente a las variables derivadas como freq_activa. En cambio, al periodo en días transcurrido hasta el primer acceso period_primer_acces se le asigna un valor concordante con su significado para el concepto como es el valor de dias_totales y de manera equivalente para periodo_inactivo. Para variables categóricas se crean categorías específicas como la categoría 'no_accesos' para la franja_habitual o dia_habitual. De manera similar para las variables derivadas de reservas.

- Los valores faltantes debidos a errores en la captura de alguna variable relevante como el precio de 'Cost_mensual' se pueden estimar para un porcentaje elevado de registros a partir de otras variables y otros registros como se detalló anteriormente.

- Al ajustar la ventana de tiempo a la de funcionamiento de la aplicación, se truncan gran parte de los registros con valores faltantes de las variables de reservas como 'numero reservas', 'Attended','GroupExercise',etc. Para los registros que permanecen con valores nulos se interpreta como 0 o como categoría 'sin reservas' de manera similar a como se procede con entradas.

- Por último, para tratar los valores faltantes en la única columna numérica 'Cost_mensual' del conjunto que aún contiene nulos, se hace codificación *onehot* de las variables categóricas, se realiza imputación de valores con dos métodos distintos, *MICE y KNNImputer* con estandarización y se comparan las estadísticas de los resultados.

```
La comparativa de estadisticas de CostMensual:
              antes         mice       knn_st
count   9451.000000  9456.000000  9456.000000
mean      35.057557    35.057558    35.057608
std       10.002940    10.000295    10.000864
min        0.000000     0.000000     0.000000
25%       28.998333    28.998333    28.998333
50%       35.268200    35.260000    35.264100
75%       41.417649    41.414167    41.414167
max       92.870000    92.870000    92.870000
```

Los resultados son muy similares debido al escaso porcentaje de nulos y el seleccionado es *KNN* con estandarización previa por su extensibilidad.

Para seleccionar el número de *k* vecinos óptimo se realizó similarmente al 'método del codo' una comparativa para diferentes valores *k* de la evolución de la diferencia entre las varianzas entre los datos originales y con la incorporación de imputados y se determina 5 como el valor óptimo.

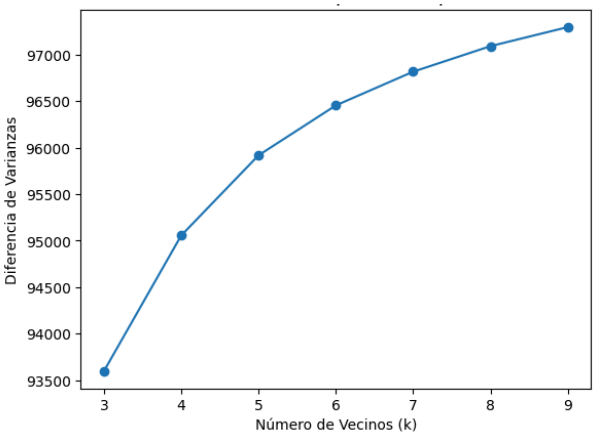


Figura 11. Diferencia de suma de varianzas según distintos valores de k.

Con los índices de la matriz de correlación entre ambos grupos de datos vemos la bondad del método ya que toman en todos los casos valores muy próximos a 1. Así mismo, se observa en la semejanza en las distribuciones de los valores comparadas antes y después de imputar.

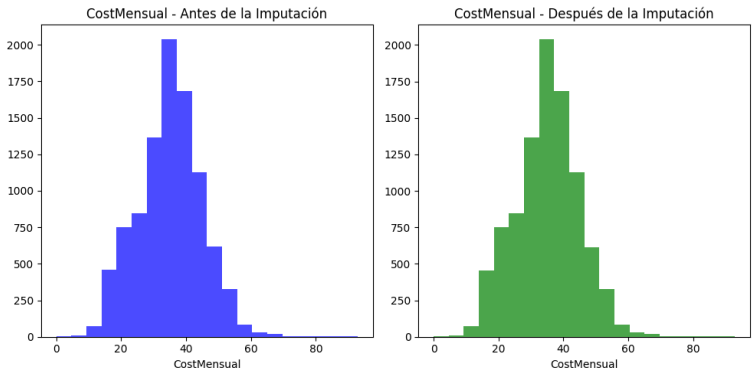


Figura 12. Distribución de la variable numérica antes y después de la imputación.

4.1.7 Tratamiento de outliers y balanceo de clases

En los boxplot se observa que en el conjunto de datos hay valores extremos, esto es, con poca frecuencia de aparición y valores mucho más altos o bajos que el grueso de ellos. Son las observaciones que aparecen fuera de los “bigotes” de las cajas y que dibujan colas largas en las curvas de distribución y violines.

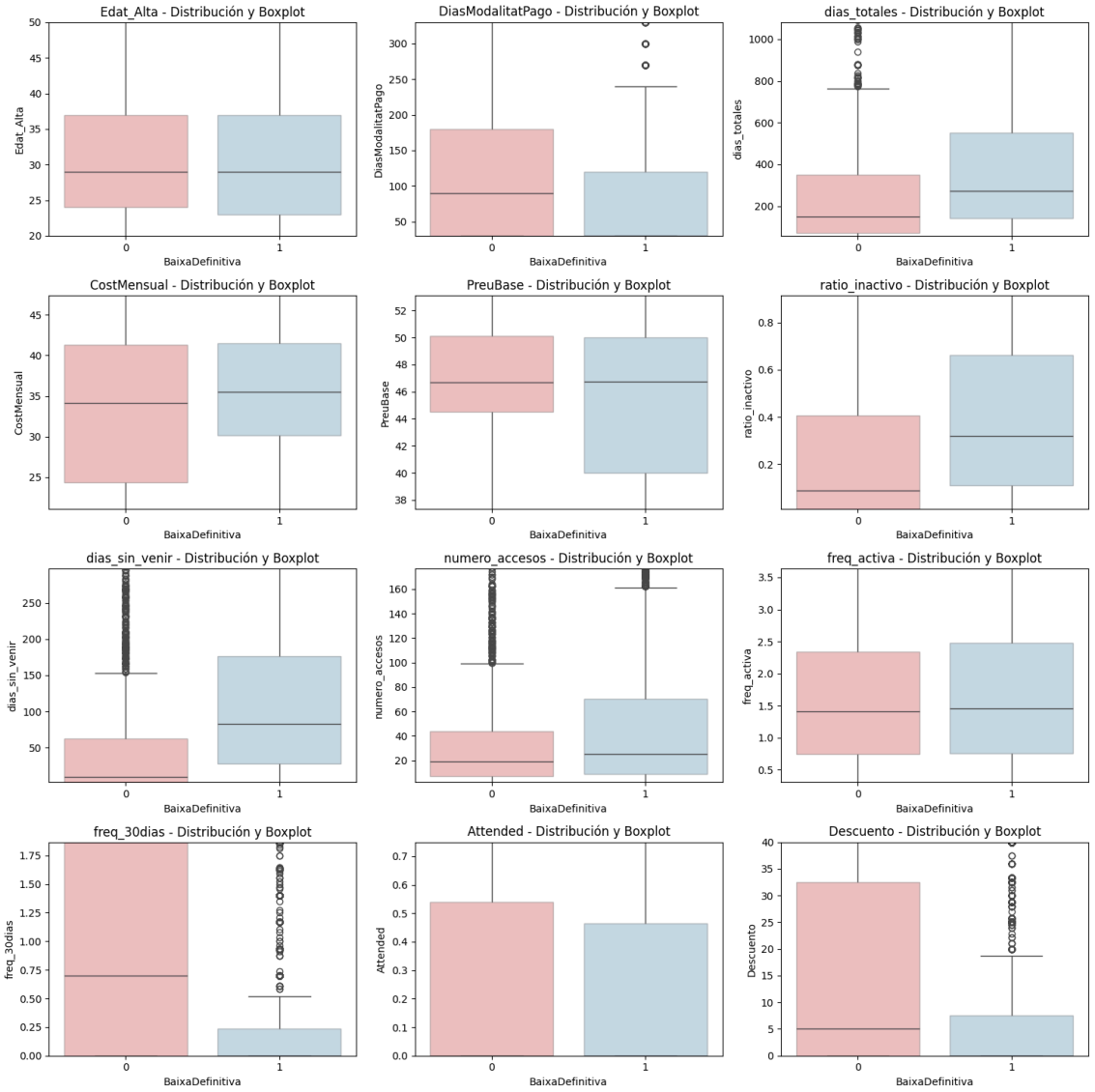


Figura 13. Diagramas de *box-plot* de variables numéricas para la identificación de *outliers*.

Se aplican dos algoritmos diferentes para la detección de outliers. *DBSCAN*, el más conocido de ellos, detecta como outliers todo el conjunto de datos a causa de su distribución irregular, así que su resultado se descarta.

Isolation Forest es un algoritmo de detección de *outliers* basado en árboles de decisión; la idea central detrás de Isolation Forest es que los *outliers* son excepciones

y pueden ser aislados más fácilmente que las instancias normales en un conjunto de datos. Tiene las ventajas de que es eficiente en tiempo y espacio, es robusto a la presencia de características irrelevantes y no asume ninguna distribución específica de los datos.

Aplicado al estudio, se identifica el 1% del total como *outliers*, 95 de 9456, que tras una revisión manual, la mayor parte corresponden a usuarios extremadamente antiguos o con importes elevados que difícilmente representarán la casuística habitual. De las tres opciones de tratamiento, ignorarlos, marcarlos o eliminarlos se opta por esta última como medida inicial.

```
print(outliers_IF.shape)
output
(95, 78)
```

Sobre el balance de clases, cuando se cuenta el porcentaje de registros que toma cada valor 0 y 1 de la variable respuesta `BaixaDefinitiva`, se observa que aparece algo desequilibrado en proporción de tres a uno para los usuarios con evento negativo. No obstante, no se considera necesario aplicar técnicas de balanceo de clases como *Oversampling* o generación sintética de datos, ya que en el análisis de supervivencia las técnicas son robustas ante el desequilibrio entre las clases.

```
Porcentaje de valores en BaixaDefinitiva:
1.0    74.735618
0.0    25.264382
Name: BaixaDefinitiva, dtype: float64
```

4.2 Fase 2a

Sobre el conjunto de datos final `data_reducido` que comprende los registros en la ventana de tiempo de funcionamiento de la app, sin variables auxiliares o redundantes, con valores nulos imputados y sin *outliers* se procede al entrenamiento y validación de los modelos estimadores.

4.2.1 Selección del modelo

Se seleccionan 6 modelos de la librería scikit survival que por sus características permitan ser comparados.

1. CoxPHSurvivalAnalysis (Modelo Proporcional de Riesgos de Cox): Asume una relación proporcional entre la función de riesgo y las variables explicativas. No asume una forma específica para la función de supervivencia, lo que lo hace flexible. Como resultados retornará coeficientes para cada característica, que representan el logaritmo del riesgo relativo.

- Ventajas: Flexibilidad, interpretabilidad de coeficientes, ampliamente utilizado y bien entendido.
- Desventajas: Supone riesgos proporcionales, lo cual no siempre es cierto en la práctica.

2. CoxnetSurvivalAnalysis (Regresión de Cox con Regularización): Extiende el modelo de Cox mediante la adición de regularización (penalización) para mejorar la selección de características y prevenir el sobreajuste. Como resultados también coeficientes regularizados para las características.

- Ventajas: Mejora la selección de características, útil para datos de alta dimensión, previene el sobreajuste.
- Desventajas: Puede ser más complejo de interpretar, requiere la selección de parámetros de regularización.

3. RandomSurvivalForest (Bosque Aleatorio de Supervivencia): Es una adaptación del algoritmo de bosques aleatorios para datos de supervivencia. Consiste en un conjunto de árboles de decisión, cada uno entrenado con un subconjunto de datos. Cada árbol da una estimación de la función de riesgo acumulativo, y la predicción final es el promedio de estas estimaciones.

- Ventajas: No requiere la suposición de riesgos proporcionales, maneja bien las interacciones complejas y los datos no lineales.
- Desventajas: Menor interpretabilidad, más complejidad computacional.

4. FastKernelSurvivalSVM (Máquina de Soporte Vectorial para Supervivencia): Es una variante de las máquinas de soporte vectorial (SVM) adaptada para la supervivencia, que utiliza una función de kernel para transformar los datos a un espacio de características de mayor dimensión donde se busca el hiperplano óptimo

que mejor separa los tiempos de evento. Retorna como resultados índices de riesgo para las observaciones.

- Ventajas: Buen rendimiento con datos no lineales, eficaz en conjuntos de datos de alta dimensión.

- Desventajas: Interpretación menos directa de resultados, selección de parámetros del kernel.

5. FastSurvivalSVM (Máquina de Soporte Vectorial Lineal para Supervivencia): Similar al FastKernelSurvivalSVM, pero utiliza un enfoque lineal en lugar de un kernel. Está diseñado para maximizar el margen entre los eventos y no eventos en el conjunto de datos, proporcionando índices de riesgo para cada observación.

- Ventajas: Más simple que el SVM con kernel, más más interpretable y computacionalmente eficiente.

- Desventajas: Menos flexible con datos no lineales.

6. GradientBoostingSurvivalAnalysis (Análisis de Supervivencia con Gradient Boosting): Combina múltiples modelos predictivos débiles, generalmente árboles de decisión, en un modelo fuerte mediante un enfoque de boosting. En cada iteración, se añade un nuevo modelo que corrige los errores cometidos por la suma de los modelos anteriores. Retorna estimaciones de la función de supervivencia.

- Ventajas: Alto rendimiento, maneja bien las interacciones no lineales y complejas.

- Desventajas: Puede ser propenso al sobreajuste, requiere cuidado en la selección de parámetros.

```
model1 = CoxPHSurvivalAnalysis(alpha=0.5)
model2 = CoxnetSurvivalAnalysis(fit_baseline_model=True)
model3 = RandomSurvivalForest(n_estimators=100, max_depth=5)
model4 = FastKernelSurvivalSVM()
model5 = FastSurvivalSVM()
model6 = GradientBoostingSurvivalAnalysis()

modelos=[model1, model2, model3, model4, model5, model6]
```

En la siguiente tabla se resumen sus características y la justificación de depor que RandomSurvivalForest será el modelo de referencia en el estudio.

Modelo	Relaciones Complejas	Robustez a la Censura	Reducción Sobreajuste	**Explicabilidad**	Flexibilidad y Generalización
CoxPHSurvivalAnalysis	Baja	Alta	Media	Alta	Media
CoxnetSurvivalAnalysis	Media	Alta	Alta	Media	Media
RandomSurvivalForest	**Alta**	**Alta**	**Alta**	**Alta**	**Alta**
FastKernelSurvivalSVM	Alta	Media	Media	Media	Alta
FastSurvivalSVM	Alta	Media	Media	Media	Alta
GradientBoostingSurvivalAnalysis	Alta	Alta	Alta	Media	Alta
Redes Neuronales	Alta	Media	Media	Baja	Alta

4.2.2 División en conjuntos de entrenamiento / test y entrenamiento.

Se divide el conjunto en variables predictora y de respuesta, siendo estas últimas `'BaixaDefinitiva', 'dias_totales'`. Se dejan de tener en cuenta las variables que presentan alta correlación.

```
# Seleccionar las variables predictoras que se desean incluir en el modelo
predictors = [col for col in df.columns if col not in ['BaixaDefinitiva',
'dias_totales','period_solic_baixa']]

# Definir las características (X) y las etiquetas (y)
X = df[predictors]
y = df[['BaixaDefinitiva', 'dias_totales']]
```

Se aplican dos enfoques para la división de datos en subconjuntos de entrenamiento y test, la primera, simple, con un 30% de reserva para el teste y que se utiliza por su rapidez y simplicidad para entrenar los modelos.

```
# Entrenamiento rapido con conjunto de entrenamiento y test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Lo smétodos esperan un array compuesta por evento y tiempo como respuesta
y_train_np = Surv.from_dataframe('BaixaDefinitiva', 'dias_totales', y_train)
y_test_np = Surv.from_dataframe('BaixaDefinitiva', 'dias_totales', y_test)
```

```
for model in modelos:
   model.fit(X_train, y_train_np)
```

La segunda, computacionalmente mucho más lenta, con validación cruzada en cinco iteraciones para obtener una aproximación más exacta de las métricas de evaluación. Se observa de todas maneras que los valores de las métricas con ambos métodos son muy aproximados.

```
# Dividir los datos en folds para Cross Validation
kf = KFold(n_splits=5, shuffle=True)

for model in modelos:
   cv_results = []
   for train_index, test_index in kf.split(df):
       X_train, X_test = X.iloc[train_index], X.iloc[test_index]
       y_train, y_test = y.iloc[train_index], y.iloc[test_index]
```

Con estos subconjuntos de entrenamiento y test se realiza el entrenamiento llamando al método `.fit`. Los algoritmos de supervivencia esperan como variable respuesta y una matriz donde la primera columna corresponde al indicativo del evento en formato booleano y la segunda el tiempo hasta la sucesión del evento; es necesario pre-formatear las variables respuestas con el método `.Surv` que la librería ofrece para ello.

```
# Lo smétodos esperan un array compuesta por evento y tiempo como respuesta
y_train_np = Surv.from_dataframe('BaixaDefinitiva', 'dias_totales', y_train)
y_test_np = Surv.from_dataframe('BaixaDefinitiva', 'dias_totales', y_test)

for model in modelos:
   model.fit(X_train, y_train_np)
```

Nota: En el estudio se prescinde únicamente de la variable `'period_solic_baixa'` por su alta correlación con la variable respuesta antes comentada. Aún con ello, la variante con Kernel de SVM, presentaba problemas de convergencia en algunas iteraciones. Se prueba a aumentar el porcentaje de datos de entrenamiento, a hacer un análisis de componentes principales y a realizar un ajuste de hiperparámetros sin éxito. Siendo aceptable su rendimiento para valores bajos de tiempo (generalmente el periodo de interés práctico) según las métricas obtenidas y no siendo este el modelo final escogido, se opta por dejar como trabajo de ampliación la re-aplicación del modelo sobre un conjunto de variables más reducido con un índice de correlación algo menor.

```
# Definir los hiperparámetros para la búsqueda en cuadrícula
param_grid = {
```

```
    'alpha': [0.1, 1, 10],# Parámetro de regularización, como C en otros SVM
    'kernel': ['rbf'],      # Otras opciones podrían ser 'linear', 'poly'
    'gamma': [0.01, 0.1, 1] # Relevante para algunos tipos de kernel como 'rbf'
}
# FastKernelSurvivalSVM
model4 = FastKernelSurvivalSVM(optimizer='rbtree', kernel='rbf')
grid_search = GridSearchCV(model4, param_grid, cv=5)
grid_search.fit(X_pca, y_np)
```

4.2.3 Resultados obtenidos

Cada uno de los modelos devuelve resultados ligeramente diferentes cuando se usa el método `.predict(X)` y es importante y parte de este estudio entender cuales son cada uno de esos resultados.

CoxPHSurvivalAnalysis y CoxnetSurvivalAnalysis: devuelven el riesgo relativo de cada observación en X. Este valor es un índice de riesgo: un número más alto indica un mayor riesgo de experimentar el evento de interés. La comparación relativa de estos valores entre las observaciones puede ser útil. Se puede visualizar la distribución de los índices de riesgo o compararlos en subgrupos de datos.

RandomSurvivalForest: devuelve una estimación del riesgo acumulativo para cada observación. Estos valores representan la probabilidad acumulada de que el evento ocurra en un momento dado, siendo valores más altos indicativos de mayor riesgo. Se puede visualizar la distribución del riesgo acumulativo o comparar las curvas de supervivencia/riesgo estimadas con el método correspondiente.

FastKernelSurvivalSVM y FastSurvivalSVM: estos modelos devuelven un índice de riesgo, similar a los modelos de Cox, un valor más alto indica un mayor riesgo. La escala puede ser diferente, pero la relación relativa entre los valores es lo importante. Se puede visualizar la distribución de los índices de riesgo.

GradientBoostingSurvivalAnalysis: este modelo predice índices de riesgo Si la pérdida (loss) es 'coxph', las predicciones se pueden interpretar como el logaritmo de la razón de riesgo (log hazard ratio), similar al predictor lineal de un modelo de riesgos proporcionales de Cox. Si la pérdida es 'squared' o 'ipcwls', las predicciones representan el tiempo hasta el evento. Los valores son probabilidades de supervivencia en el tiempo, donde un valor más alto indica una mayor probabilidad de supervivencia en ese momento.

Así, se observa en la siguiente figura la curva de distribución de los índices de riesgo de los modelos 1 y 4.

```
for model in modelos:
    risk_scores = model.predict(X_test)
    sns.histplot(risk_scores, kde=True)
```

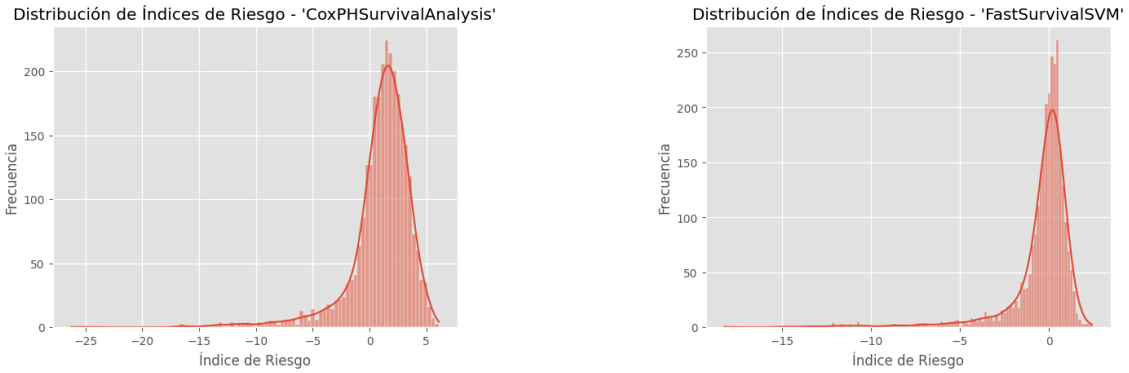


Figura 14. Distribuciones de índices de riesgo.

La librería ofrece también (no para modelos SVM) los métodos `.predict_survival_function` y `.predict_cumulative_hazard_function` para obtener directamente los valores para cada observación e instante de tiempo de las funciones de supervivencia y de riesgo respectivamente calculadas por los estimadores y que permiten su visualización. Debido al volumen de observaciones, se visualiza mejor seleccionando un número limitado de registros como muestra o definiendo grupos y representantes del total de registros. Así se puede comparar la consistencia de los resultados de ambos métodos.

```
# Obtener las funciones de supervivencia en el conjunto de prueba
survival_functions = model3.predict_survival_function(X_test)

# Por ejemplo, para las primeras 7 filas del conjunto de prueba
for i in range(7):
    time_points = survival_functions[i].x
    survival_probs = survival_functions[i](time_points)
    plt.step(time_points, survival_probs, label=f'Muestra {i + 1}')

(...)
survival_functions = model3.predict_cumulative_hazard_function(X_test)
(...)
```

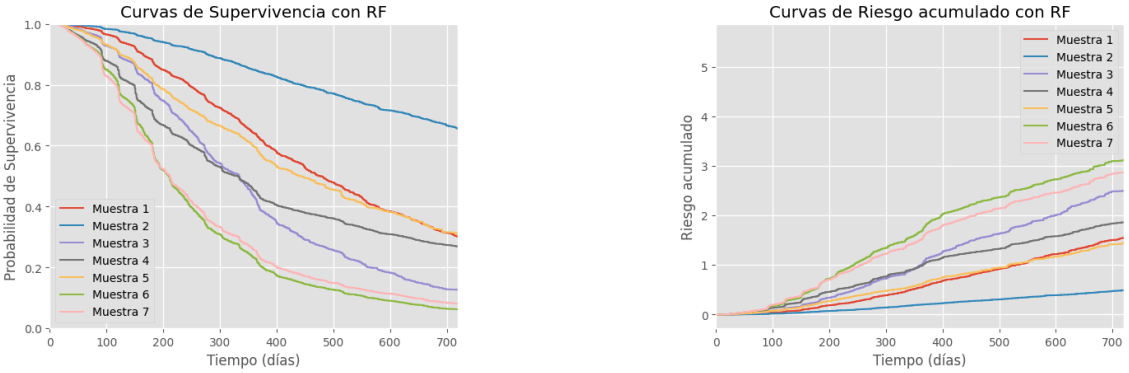


Figura 15. Curva de supervivencia y riesgo acumulado para las primeras siete observaciones.

Si se definen grupos por una variable, por ejemplo por `'Edat_Alta'` se puede obtener la variación de ambas funciones en función de esa característica.

```
# CURVAS DE SUPERVIVENCIA POR UNA VARIABLE
grupos_edad = {
    'menor_30': X_test['Edat_Alta'] < 30,
    'entre_30_60': (X_test['Edat_Alta'] >= 30) & (X_test['Edat_Alta'] <= 60),
    'mayor_60': X_test['Edat_Alta'] > 60}

models_funcion_surv = [model1, model2, model3, model6]
for model in models_funcion_surv:
    plt.figure(figsize=(8, 6))
    for grupo_nombre, grupo_mascara in grupos_edad.items():
        X_grupo = X_test[grupo_mascara]
        y_grupo = y_test[grupo_mascara]
    (…)
# Obtener las funciones de supervivencia
        surv_funcs = model.predict_survival_function(X_grupo)
        # Calcular la función de supervivencia promedio para el grupo
        curva_promedio = np.mean([func.y for func in surv_funcs], axis=0)
    (…)
# Obtener las funciones de riesgo acumulado por edad
        surv_funcs = model.predict_cumulative_hazard_function(X_grupo)
```

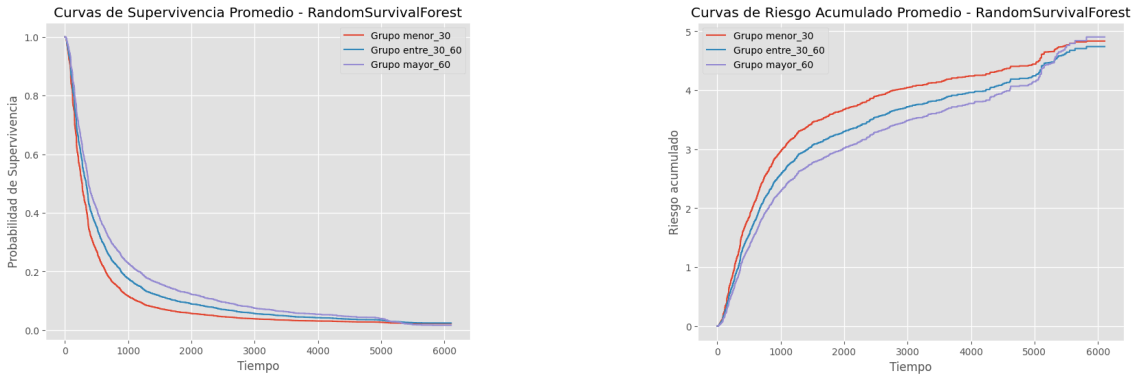


Figura 16. Curva de supervivencia y riesgo acumulado por grupos de edad.

O se puede agrupar las observaciones por percentil de probabilidad de supervivencia (o riesgo), y elegir un representante de cada grupo (en este caso la mediana) del que obtener una única representación. Esto permitirá, por ejemplo analizar las características intra-grupos

```
percentiles = np.percentile(df_predicciones['probabilidad_supervivencia'],
np.linspace(0, 100, 11))
    etiquetas_percentiles = ['0-10%', '10-20%', '20-30%', '30-40%', '40-50%',
'50-60%', '60-70%', '70-80%', '80-90%', '90-100%']

                                df_predicciones['grupo_percentil'] =
pd.cut(df_predicciones['probabilidad_supervivencia'], percentiles,
include_lowest = True, labels=etiquetas_percentiles)

      # Utilizar predict_survival_function para obtener las funciones de
supervivencia
       surv_funcs = model.predict_survival_function(X_test)

       representantes = df_predicciones.groupby('grupo_percentil')
['probabilidad_supervivencia'].median()
```

```
    # índice de la observación más cercana a la mediana en cada grupo
indices_representativos = df_predicciones.groupby('grupo_percentil').apply
(lambda x: (x['probabilidad_supervivencia'] -
representantes[x.name]).abs().idxmin())
```

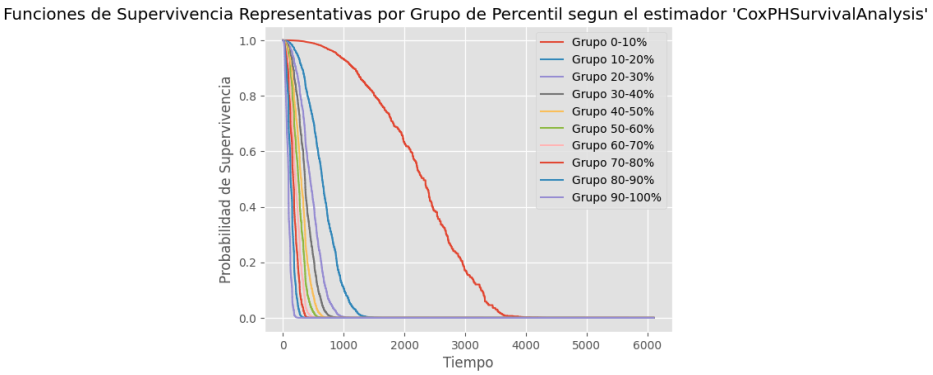


Figura 17. Curvas de supervivencia por percentiles de riesgo.

En los modelos SSVM , al no disponer de estos métodos, se puede utilizar el estimador `kaplan_meier_estimator` para visualizar las predicciones, por ejemplo, por `'Edat_Alta'`

```
# Calcular predicciones de riesgo para el grupo
riesgo_predicho = model4.predict(X_grupo)

# Calcular las curvas de supervivencia utilizando Kaplan-Meier
tiempo, prob_surv = kaplan_meier_estimator(y_grupo['BaixaDefinitiva'].
astype('bool'), y_grupo['dias_totales'])

# Graficar la curva de supervivencia
plt.step(tiempo, prob_surv, where="post", label=f'Grupo {nombre}')
```

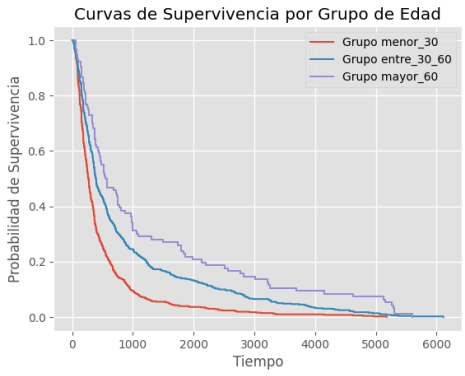


Figura 18. Curvas de supervivencia para SSVM con Kaplan-Meier.

A partir de los valores retornados, y, como objeto principal y final de interés del estudio, podemos identificar y ordenar aquellos registros del conjunto de test que tienen menor probabilidad de supervivencia más allá de un periodo de tiempo T de estudio (por ejemplo más de 2 meses) o mayor riesgo acumulado de cancelar suscripción antes de T y determinar umbrales de aceptación para definir los grupos de

riesgo sobre los que aplicar las medidas estratégicas de la compañía. Nótese que aunque en estos dos grupos aunque hay una evidente relación, no necesariamente aparecen las ocurrencias exactamente en el mismo orden.

```
survival_functions = model3.predict_survival_function(X_test)
cum_hazard_functions = model3.predict_cumulative_hazard_function(X_test)

# Índice donde el tiempo es igual o más cercano a 61
indice_tiempo = np.where(survival_functions[0].x >= 61)[0][0]

# valores de supervivencia y riesgo en el día 61
survival_ini = [fn.y[indice_tiempo] for fn in survival_functions]
hazard_ini = [fn.y[indice_tiempo] for fn in cum_hazard_functions]

survival_ini_pd = pd.DataFrame({"survival_initial": survival_initials})
hazard_ini_pd = pd.DataFrame({"hazard_initial": hazard_initials})

X_test_with_ini = pd.concat([X_test, survival_initials_pd, hazard_ini_pd], axis=1)

# Ordenado por supervivencia y riesgo
X_test_ord_by_surv = X_test_with_ini.sort_values(by="survival_ini", ascending=True)
X_test_ord_by_hazd = X_test_with_ini.sort_values(by="hazard_ini", ascending=False)

print(X_test_ord_by_surv[['survival_ini','hazard_ini']].head(5))
print(X_test_ord_by_hazd[['survival_ini','hazard_ini']].head(5))
```

```
survival_initial  hazard_initial
1153          0.632452        0.641760
263           0.642035        0.606051
647           0.643547        0.548537
1120          0.650403        0.569459
1337          0.663329        0.561689

      survival_initial  hazard_initial
1153          0.632452        0.641760
263           0.642035        0.606051
1455          0.694149        0.582878
1120          0.650403        0.569459
652           0.695919        0.565702
```

Así, en un ejemplo de uso, si se quiere diseñar estrategias de retención para la cohorte de hombres entre 40 y 50 años apuntados en enero, se podrá comparar las funciones específicas de esta cohorte con las del total de inscripciones:

```
# CURVAS DE SUPERVIVENCIA Y RIESGO CASO DE USO
grupos_interes = {
  'total': (X_test['Edat_Alta'] >= 10),
    'cohorte_estudio_ene':(X_test['Mes_de_Alta_2'] == 0) & (X_test['Mes_de_Alta_3']
==0) &
   (X_test['Mes_de_Alta_4'] == 0) & (X_test['Mes_de_Alta_5'] == 0) &
   (X_test['Mes_de_Alta_6'] == 0) & (X_test['Mes_de_Alta_7'] == 0) &
   (X_test['Mes_de_Alta_8'] == 0) & (X_test['Mes_de_Alta_9'] == 0) &
   (X_test['Mes_de_Alta_10'] == 0) & (X_test['Mes_de_Alta_11'] == 0) &
   (X_test['Mes_de_Alta_12'] == 0),
   'cohorte_estudio_hm': (X_test['Edat_Alta'] >= 40) & (X_test['Edat_Alta']<=50) &
   (X_test['Sexe_2'] == 1) & (X_test['Mes_de_Alta_2'] == 0)  &
   (X_test['Mes_de_Alta_3'] == 0) & (X_test['Mes_de_Alta_4'] == 0) &
   (X_test['Mes_de_Alta_5'] == 0) & (X_test['Mes_de_Alta_6'] == 0) &
   (X_test['Mes_de_Alta_7'] == 0) & (X_test['Mes_de_Alta_8'] == 0) &
   (X_test['Mes_de_Alta_9'] == 0) & (X_test['Mes_de_Alta_10'] == 0) &
   (X_test['Mes_de_Alta_11'] == 0) & (X_test['Mes_de_Alta_12'] == 0)
   }
models = [model3]
```

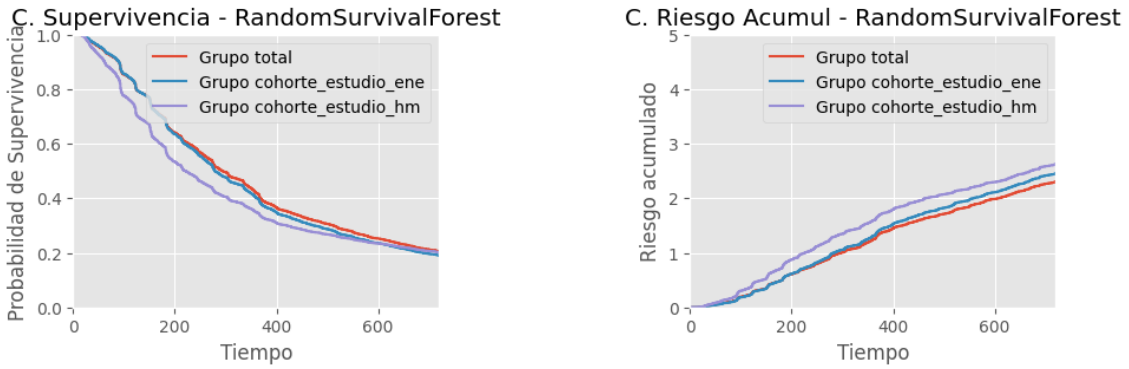


Figura 19. Curvas de supervivencia y riesgo acumulado para hombres entre 40 y 50 años apuntados en enero vs total de inscritos en enero vs total

Donde se observa que no hay una diferencia significativa entre la gente que se apunta en enero sobre el total de observaciones, pero si la hay en ese grupo específico de edad y género, siendo algo menores los índices de supervivencia para ellos. Esa diferencia, por tanto, estará mejor explicada por esas dos variables y, en caso de implementar estrategias específicas valdrá la pena hacerlo sobre todo ese colectivo de edad y género y no solamente sobre los inscritos en enero.

	Cohorte	Tiempo	P Superv Promedio	Índ de Riesgo Promedio
0	total	180	0.693619	0.493696
1	cohorte_estudio_ene	180	0.690358	0.491809
2	cohorte_estudio_hm	180	**0.587976**	**0.717053**

Así por ejemplo se debería comunicar una promoción de 3 meses de regalo abonando 9, destinada a ese colectivo en concreto y comunicarlo especialmente por teléfono a las primeras cinco observaciones ordenadas por menos probabilidad de supervivencia a los 6 meses.

.

```
mascara_cohorte = (
    (X_test['Edat_Alta'] >= 40) & (X_test['Edat_Alta']<=50) &
    (X_test['Sexe_2'] == 1) & (X_test['Mes_de_Alta_2'] == 0)  &
    (X_test['Mes_de_Alta_3'] == 0) & (X_test['Mes_de_Alta_4'] == 0) &
    (X_test['Mes_de_Alta_5'] == 0) & (X_test['Mes_de_Alta_6'] == 0) &
    (X_test['Mes_de_Alta_7'] == 0) & (X_test['Mes_de_Alta_8'] == 0) &
    (X_test['Mes_de_Alta_9'] == 0) & (X_test['Mes_de_Alta_10'] == 0) &
    (X_test['Mes_de_Alta_11'] == 0) & (X_test['Mes_de_Alta_12'] == 0)
)
X_cohorte = X_test[mascara_cohorte]
surv_funcs = model3.predict_survival_function(X_cohorte)
prob_surv_180_dias = [func(180) for func in surv_funcs]
```

Indice_Original	Probabilidad_Supervivencia_180_dias
3	**833**	**0.399167**
0	**1090**	**0.448450**
2	**6700**	**0.529572**
4	**4103**	**0.552564**
1	**6785**	**0.628360**
5	1509	0.969745

Nota: Los modelos CoxPHSurvivalAnalysis y GradientBoostingSurvivalAnalysis son modelos paramétricos, lo que significa que asumen una forma específica para la función de riesgo acumulativo. En el caso de CoxPHSurvivalAnalysis, la función de riesgo acumulativo se asume como una función lineal de los predictores. En el caso de GradientBoostingSurvivalAnalysis, la función de riesgo acumulativo se asume como una función no lineal de los predictores, pero que se aproxima mediante una serie de funciones lineales.

El modelo RandomSurvivalForest, por otro lado, es un modelo no paramétrico, lo que significa que no asume ninguna forma específica para la función de riesgo acumulativo. En su lugar, el modelo estima la función de riesgo acumulativo a partir de un conjunto de árboles de decisión.

En el caso de los datos utilizados, las funciones de riesgo acumulativo individuales para cada observación son bastante diferentes. Esto se debe a que los datos incluyen una variedad de valores diferentes para los predictores.

Los modelos CoxPHSurvivalAnalysis y GradientBoostingSurvivalAnalysis no son capaces de capturar esta variabilidad en las funciones de riesgo acumulativo individuales. En cambio, estos modelos estiman una función de riesgo acumulativo promedio para todas las observaciones. Por eso, la función de riesgo acumulativo promedio se observa superpuesta como una única función en las gráficas de los modelos CoxPHSurvivalAnalysis y GradientBoostingSurvivalAnalysis.

El modelo RandomSurvivalForest, en cambio si es capaz de capturar, calcular y representar gráficamente la variabilidad en las funciones de riesgo acumulativo individuales, y por ello, junto a sus buenas métricas, es el modelo utilizado.

No se ve este fenómeno en las funciones de supervivencia debido a la forma en que estas funciones se derivan del riesgo acumulativo. Aunque las curvas de riesgo acumulativo pueden ser similares, la transformación exponencial utilizada para obtener la función de supervivencia puede resultar en más variabilidad visible en estas curvas.

4.2.4 Evaluación del modelo

- 1. C-Índex

Como se comentaba en la introducción teórica, las métricas de evaluación del análisis de supervivencia son propias del método ya que en el contexto donde no es posible clasificar la observación como 0 o 1 porque esta clasificación es cambiante con el tiempo, se hace necesario definir nuevas maneras de valorar el rendimiento.

El C-índex evalúa qué proporción de pares de eventos ordenados suceden en el mismo orden que la predicción realizada y de esta manera se mide la correcta relación de temporalidad entre las mismas. Así, un índice de 1 es indicativo de una predicción perfecta , 0 una perfectamente errónea y 0.5 la misma precisión que una predicción al azar.

En la tabla siguiente se observa que los resultados del modelo obtienen muy alto puntaje, siendo de 0,76 para FastKernelSSVM el más bajo (debido posiblemente a su no-convergencia) y 0,95 el mejor que ofrece GradientBoostSA. Cabe comentar también que un resultado tan positivo en la aplicación del modelo de Cox confirma la hipótesis de sus supuestos.

```
for model in modelos:
   cindex = model.score(X_test, y_test_np)
El c-index del algoritmo 'CoxPHSurvivalAnalysis' es: 0.8749683570692541
El c-index del algoritmo 'CoxnetSurvivalAnalysis' es: 0.8799080046935773
El c-index del algoritmo 'RandomSurvivalForest' es: 0.8719596727599783
El c-index del algoritmo 'FastKernelSurvivalSVM' es: 0.7605369098035011
El c-index del algoritmo 'FastSurvivalSVM' es: 0.9108577691212646
El c-index del algoritmo 'GradientBoostingSurvivalAnalysis' es:
0.9547241778794
```

- 2. Integrated Brier Score (IBS)

El *Brier Score* original propuesto por Frank E. Harrell Jr, es una medida de la precisión de las predicciones probabilísticas; en el contexto de la supervivencia, cuantifica qué tan bien las predicciones de supervivencia coinciden con los eventos observados (eventos o censuras) en un momento específico. El *Integrated Brier Score* que a continuación se evalúa es la media del Brier Score sobre todos los tiempos de seguimiento, ponderado por la estimación de la función de supervivencia; al integrar sobre el tiempo, el IBS captura la precisión de las predicciones a lo largo de todo el rango de tiempos, proporcionando una visión global de la precisión. Valores cercanos a 0 son positivos y cercanos a 1 son negativos.

```
Integrated Brier Score for CoxPHSurvivalAnalysis: 0.16034807337139584
Integrated Brier Score for CoxnetSurvivalAnalysis: 0.14479654969393985
Integrated Brier Score for RandomSurvivalForest: 0.12252872121701221
Integrated Brier Score for GradientBoostingSurvivalAnalysis: 0.12548748825038
```

Aunque la librería no ofrece la implementación del IBS para los modelos basados en SSVM es posible su interpretación mediante las curvas de Kaplan_meier pero esa ejecución queda fuera del abasto de este estudio. También aquí GradientBoostSA obtiene el mejor valor de la métrica.

- 3. AUC-ROC Dinámica

La curva AUC-ROC (Área Bajo la Curva del Receptor Operativo Característico) muestra el rendimiento de un modelo de clasificación para diferentes umbrales de clasificación. Esta curva traza dos parámetros:

- Tasa de Verdaderos Positivos (Sensibilidad): en el eje Y, representa la proporción de positivos reales que se identificaron correctamente.
- Tasa de Falsos Positivos (1 - Especificidad): en el eje X, representa la proporción de negativos reales que se clasificaron incorrectamente como positivos.

El AUC es una medida de la capacidad del modelo para distinguir entre las clases. Un AUC de 1.0 representa un modelo perfecto que clasifica todos los positivos y negativos correctamente. Un AUC de 0.5 sugiere un modelo sin capacidad discriminativa, equivalente a una clasificación aleatoria.

La AUC dinámica en análisis de supervivencia es una extensión de la AUC ROC tradicional para datos censurados y tiempos variables. Mientras que la AUC ROC estática evalúa la capacidad de un modelo para clasificar correctamente en un punto fijo en el tiempo, la AUC dinámica mide esta capacidad a lo largo de varios puntos temporales y permite evaluar cómo cambia la precisión predictiva del modelo a lo largo del tiempo, proporcionando una visión más completa de su rendimiento a lo largo de todo el período de estudio. Se puede realizar mediante el método `.cumulative_dynamic_auc` que la propia librería ofrece y constituye un buen indicador de la precisión del modelo.

```
for model in modelos:
    # Ajustar va_times para concentrarse en un rango con más eventos y menos censura
   va_times = np.arange(100, 5000, 200)
   cph_risk_scores = model.predict(X_test)
   cph_auc, cph_mean_auc = cumulative_dynamic_auc(y_train_np, y_test_np, cph_risk_scores, va_times)
```

```
El AUC del algoritmo 'CoxPHSurvivalAnalysis' es: 0.9672431162059512
El AUC del algoritmo 'CoxnetSurvivalAnalysis' es: 0.9681694844463583
El AUC del algoritmo 'RandomSurvivalForest' es: 0.9567000040469562
El AUC del algoritmo 'FastKernelSurvivalSVM' es: 0.797358546184552
El AUC del algoritmo 'FastSurvivalSVM' es: 0.9799122017973194
El AUC del algoritmo 'GradientBoostingSurvivalAnalysis' es: 0.9945964342756732
```

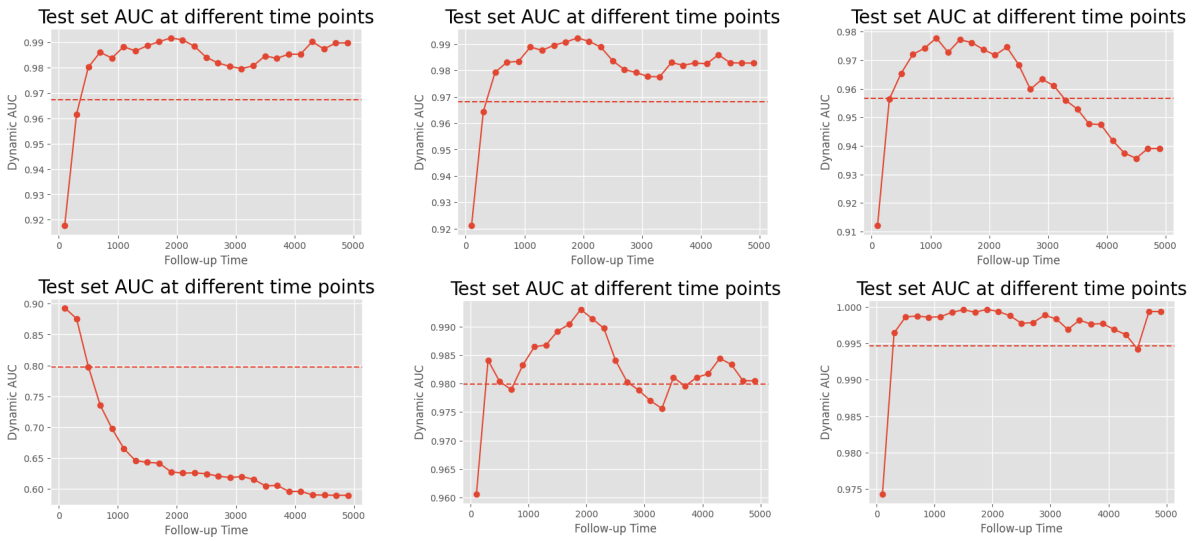


Figura 20. Curvas AUC-ROC dinámicas de los 6 modelos entrenados

Aquí también GradientBoostingSV, con casi un 0.99 de puntuación, FastSurvivalSVM y los modelos de Cox ofrecen el mejor resultado. Sin obtener una mala puntuación, se puede observar cómo Random Survival Forest pierde algo de capacidad predictiva con el paso del tiempo, provocado posiblemente por la menor cantidad de datos sin censura disponibles para momentos de tiempo más lejanos. En el modelo de SVM con Kernel se reflejan también los problemas de convergencia que antes se comentaron, igualmente en el contexto de escasez de referencias para periodos de tiempo largos.

- 4. Análisi de características principales

Sobre el análisis de características principales, es decir, la observación de qué características explican mejor las predicciones del modelo, la librería ofrece el método `.coef_` que retorna los coeficientes del modelo (no para RSF o GradientBoostSA). Si bien su valor relativo y signo nos da una idea de la relevancia de cada característica es importante tener en cuenta que en cada modelo tiene su interpretación propia; en los modelos basados en Cox, los coeficientes indican el logaritmo del riesgo relativo (un coeficiente positivo señala un aumento en el riesgo con el incremento de la variable correspondiente) mientras que en SVM para supervivencia los coeficientes no reflejan riesgo relativo sino la importancia de las variables en la separación de los datos de supervivencia; contribuyen a la definición de un hiperplano que distingue entre diferentes tiempos de supervivencia, enfocándose en la optimización del margen y no en una interpretación directa del riesgo.

Así, según la tabla adjunta, las variables de `ratio_primer_us`, `ratio_inactivo` y `existe_app`, muestran una relación positiva en el modelo de Cox con el incremento del riesgo de baja, mientras que las variables `dia_habitual_No_acceso`, `franja_habitual_No_acceso` y `CategoriaModalitat_partnerVIP` lo hacen en sentido opuesto.

```
coeficientes = model1.coef_.ravel()
# Calcular riesgo relativo
riesgo_relativo = np.exp(coeficientes)
```

	Factor de riesgo	Coefficiente [β]	Riesgo relativo [exp(β)]
17	ratio_primer_us	4.173390	64.935230
16	ratio_inactivo	3.403440	30.067355
27	existe_app	0.837206	2.309903
56	act_preferida_PERSONAL_SERVICES	0.632005	1.881378
24	freq_180dias	0.625020	1.868284
10	Nutrition	0.605498	1.832164
59	act_preferida_TONING_CLASS	0.522020	1.685428
60	act_preferida_VIRTUAL_CLASS	0.403055	1.496389
52	act_preferida_CICLO_CLASS	0.358954	1.431831
54	act_preferida_FITNESS_TRAINING	0.272773	1.313602

	Factor de riesgo	Coefficiente [β]	Riesgo relativo [exp(β)]
4	Attended	-0.223636	0.799606
37	CategoriaModalitat_weekend	-0.297795	0.742453
33	CategoriaModalitat_midday	-0.319976	0.726167
29	Nacionalitat_1	-0.388158	0.678305
47	franja_habitual_nit	-0.440559	0.643676
31	CategoriaModalitat_corporate	-0.515783	0.597033
8	PersTraining	-0.723730	0.484940
44	dia_habitual_No_acceso	-1.503310	0.222393
50	franja_habitual_No_acceso	-1.503310	0.222393
36	CategoriaModalitat_partnerVIP	-2.998147	0.049879

En la evaluación de coeficientes sobre el modelo de FastSurvivalSVM las variables `freq_30dias`, `Nacionalitat_1`, `franja_habitual_tarda`, `dia_habitual_Tuesday` y `Attended` tienen mayor relevancia en la definición del hiperplano que categoriza como tal los eventos, pero no necesariamente son indicativas de la sucesión del evento.

```
indices=pd.Series(model5.coef_, index=X_test.columns)
indices_ordenados = indices.sort_values(ascending=False)
```

```
freq_30dias                0.058606
Nacionalitat_1             0.046461
franja_habitual_tarda      0.043585
dia_habitual_Tuesday       0.038775
Attended                   0.031124
dia_habitual_Monday        0.023342
percent_consumit           0.019111
GroupExercise              0.018234
Mes_de_Alta_9              0.016440
Mes_de_Alta_5              0.010736
act_preferida_sin_reservas   -0.026987
act_preferida_sin_reservas   -0.026987
Nacionalitat_2               -0.046080
existe_app                   -0.048334
ServiciosExtra_0.0           -0.053888
freq_180dias                 -0.065137
dia_habitual_No_acceso       -0.072560
franja_habitual_No_acceso    -0.072560
ratio_primer_us              -0.080026
freq_activa                  -0.151192
ratio_inactivo               -0.167396
dtype: float64
```

En modelos basados en árboles, como RandomSurvivalForest y GradientBoostSA, la importancia de las características suele referirse a cuánto contribuye una característica a mejorar la capacidad predictiva del modelo. Un valor más alto de .feature_importances_ para una característica significa que tiene un mayor impacto en la precisión de las predicciones del modelo. Se interpreta como la "importancia" de la característica en la construcción del modelo.

En la tabla siguiente se observa que para GBSA aplicado al caso las variables numero_accesos, dias_sin_venir, freq_activa, ratio_inactivo, ratio_primer_us y percent_consumit son las que más contribuyen al modelo y entre las tres primeras explican un 75% del mismo.

```
importances = model6.feature_importances_
```

	Característica	Importancia
1	**numero_accesos**	**0.426134**
15	**dias_sin_venir**	**0.199817**
21	**freq_activa**	**0.145436**
16	**ratio_inactivo**	**0.079271**
17	ratio_primer_us	0.044206
14	percent_consumit	0.028938
24	freq_180dias	0.024016
13	period_primer_acces	0.022706
12	DiasModalitatPago	0.011478
11	Descuento	0.009977

La implementación de RandomSurvivalForest se basa en la de Random Forest de scikit-learn y hereda características como la construcción de árboles en paralelo, pero no incluye la .feature_importances_ debido a las limitaciones de la medición de impureza con datos censurados. Como alternativa, se puede aplicar la función de permutation_importance de scikit-learn, que es compatible. En la tabla siguiente vemos que los resultados son muy similares a los de GradientBoostingSa.

```
from sklearn.inspection import permutation_importance
X_test_array = X_test.values
result = pd.DataFrame()

result = permutation_importance(model3, X_test, y_test_np, n_repeats=5,
random_state=0)
```

	importances_mean	importances_std
numero_accesos	**0.275975**	**0.005363**
dias_sin_venir	**0.191860**	**0.007995**
freq_activa	**0.108167**	**0.001504**
ratio_inactivo	**0.057608**	**0.000368**
ratio_primer_us	0.021633	0.002704
freq_180dias	0.008719	0.000201
percent_consumit	0.006939	0.000399
freq_90dias	0.002874	0.000174

5. Análisis y discusión de resultados

Los resultados obtenidos del análisis de supervivencia muestran un rendimiento prometedor, destacando la eficacia de GradientBoostingSurvivalAnalysis, RandomSurvivalForest, FastSurvivalSVM y también de los modelos de Cox para abordar la complejidad de la retención de miembros en un gimnasio. La métrica clave, C-index, utilizado para evaluar la capacidad de discriminación temporal del modelo, arrojó resultados por encima del 95% subrayando la capacidad del mismo para definir la probabilidad de que un abogado esté mas de periodo un determinado de tiempo (o el riesgo de que marche antes), y por tanto, discerniendo entre que miembros que permanecen y cuales abandonan y en qúe plazos de tiempo lo hacen. El IBS muestra resultados similares y también la curva AUC-ROC, con una puntuación por encima del 99% obtenida con GradientBoostingSurvivalAnalysis y con resultados especialmente buenos en todos los algoritmos para períodos de tiempos más cortos, que a la práctica, suelen ser los de mayor interés en este ámbito .

Las curvas de supervivencia y riesgo que retornan los modelos permiten, además de confirmar tendencias previstas, descubrir *insights* interesantes para la gestión de los abonados al poderse estimar cómo evoluciona el riesgo de baja del centro (o la probabilidad de supervivencia) según valores concretos de las características de los clientes o para grupos predefinidos y así actuar específicamente sobre ellos. Se pueden así identificar momentos críticos de abandono y patrones de comportamiento; este análisis temporal no solo ayuda a la comprensión del comportamiento de los clientes, sino que también puede servir de base para decidir cuándo y qué estrategias de retención y fidelización aplicar.

Definiendo como algoritmo escogido RandomSurvivalFores por el resultado de sus métricas, al explicabilidad de sus predicciones y la capacidad de detección de relaciones complejas y según el resultado obtenido del análisis de características principales, en este centro el número de accesos, los días que lleva sin venir el abonado, la frecuencia con la que entrena mientras está activo, el ratio de inactividad (tiempo sin venir sobre tiempo total), el ratio de tiempo hasta el primer uso y el porcentaje de abono ya consumido, son las seis carísticas que mejor explican los tiempos y riesgos de baja del cliente. En base a esto, la compañía debería analizar qué medidas teniendo en cuenta sus costos, previsión de mejoras esperables en los distintos kpi y umbrales de aplicación. Entre estas se podrían dar recomendaciones como:

- Intentar acortar los periodos de inicio al entrenamiento una vez inscrito con sesiones de asesoramiento técnico agendadas en el momento de la inscripción.
- Aumentar las clases de aquellos instructores con ayor asistencia en sus sesiones, para aumentar la frecuencia de uso de los abonado
- Intentar aumentar la frecuencia de asistencia al centro con programas de obtención de puntos y fidelización según el número de visitas.
- Generar campañas automáticas de aviso o promoción de las actividades preferidas del abonado a través de la app del centro cuando se llega a un cierto umbral de días sin venir
- Ofrecer renovaciones de contrato ventajosas antes del consumo total y finalización del abono contratado

La combinación de resultados numéricos sólidos y el análisis detallado de los modelos aplicados respalda la aplicación práctica del enfoque de análisis de supervivencia en la gestión de gimnasios. Al considerar características clave de los abonados, relaciones no lineales y adaptabilidad temporal, el modelo se muestra como una herramienta estratégica para la toma de decisiones proactiva. Estos resultados, respaldados por la literatura especializada de expertos como P. E. Harrell o el propio Cox, consolidan la validez y relevancia del enfoque adoptado.

Por último, los datos utilizados en el proyecto fueron tomados por un centro mediano, de aproximadamente 4500 abonados, en una cadena con centros ubicados geográficamente en una misma ciudad. Es importante tener en cuenta que los resultados del proyecto pueden no ser generalizables a otros gimnasios, especialmente si estos se encuentran en diferentes contextos y que, una vez definido el modelo, su aplicabilidad puede depender de su re-entrenamiento y re-evaluación de manera que pueda extenderse a las características propias de los abonados de diferentes centros, compañías o ubicaciones.

6. Conclusiones

6.1 Conclusiones

El objetivo principal de este proyecto era analizar el tiempo de permanencia de los miembros de un gimnasio. Para ello, se han utilizado técnicas de análisis de supervivencia, que permiten estudiar la probabilidad de que un evento ocurra en un determinado momento. Los datos utilizados en el proyecto fueron proporcionados por el gimnasio y contenían información sobre los miembros, como su edad, sexo, tipo de contrato, etc. Los datos fueron limpiados y pretratados para eliminar valores faltantes y outliers.

Se entrenaron y evaluó el rendimiento de seis modelos de análisis de supervivencia diferentes:

- Regresión de Cox: Un modelo lineal que estima la probabilidad de ocurrencia del evento en función de las características de los individuos.
- Regresión de Cox robusta: Una versión de la regresión de Cox que es más robusta a *outliers*.
- Random survival forest: Un modelo de *ensemble* adaptado al trabajo con datos censurados que combina árboles de decisión.
- Survival SVM: Un modelo de aprendizaje automático supervisado que utiliza un clasificador SVM adaptado para predecir la probabilidad de ocurrencia del evento.
- Survival SVM con kernel: Una variante del anterior con kernel para detectar relaciones no lineales.
- Gradient boosting survival: Un modelo de *ensemble* que combina árboles de decisión mediante un enfoque de refuerzo.

El proyecto ha logrado avances significativos en la aplicación de técnicas avanzadas de machine learning en el contexto del análisis de supervivencia aplicado a la retención de clientes de un gimnasio. La limpieza y preparación de datos, junto con el análisis exploratorio, han sentado una base sólida para la modelización. El enfoque en la privacidad de los datos refleja una preocupación contemporánea crucial, garantizando la conformidad con las normativas vigentes y los modelos de IA

desarrollados y entrenados demuestran una alta precisión y rendimiento según las métricas establecidas.

A la finalización del proyecto se ha conseguido un modelo, utilizando algoritmos de Inteligencia Artificial, que ofrece resultados teóricos más que satisfactorios en el objeto de estudio de los tiempos de permanencia de clientes al gimnasio y que puede servir como herramienta clave para mejorar la retención del cliente. No solamente eso, si no que se ha conseguido tras un exhaustivo estudio de su modelado y a partir del estudio previo sobre casos similares y otras aplicaciones del análisis de supervivencia; los algoritmos propuestos modelan con eficacia el comportamiento de los clientes y pueden ayudar a las empresas a identificar a los miembros con mayor riesgo de darse de baja y a elegir qué medidas tomar para retenerlos.

Para que no quede en un marco meramente teórico a partir de este punto, en el siguiente apartado se proponen los pasos necesarios para su puesta a prueba en un entorno real, que permitan retroalimentar el proceso iterativo de perfeccionamiento del modelo final y comprobar su utilidad práctica.

Como primera lección aprendida durante el desarrollo del proyecto, queda patente la importancia de conocer en profundidad y disponer de un modelo teórico sólido, en este caso la base estadística del análisis de supervivencia, que dé respaldo a la solución que se desea proponer para el problema investigado. El conocimiento teórico debe ser la base sobre la que se construya el acercamiento a utilizar en su resolución.

Durante el desarrollo del proyecto se puso de manifiesto además la necesidad de la flexibilidad y adaptabilidad en la gestión de proyectos; la capacidad de reevaluar y ajustar la planificación en respuesta a dificultades inesperadas es fundamental para el éxito del proyecto. Concretamente es importante no subestimar los tiempos de preparado y trabajo previo de las fases iniciales.

También queda patente la relevancia de tener un conocimiento sólido del entorno y el lenguaje de programación o de considerar correctamente los tiempos de aprendizaje; invertir tiempo en formación y desarrollo de habilidades antes de la ejecución del proyecto puede mitigar retrasos y mejorar la eficiencia general.

Por último, la priorización estratégica de objetivos y tareas ha sido clave para poder finalizar este proyecto. Sin embargo, es crucial considerar el impacto a largo

plazo; en futuros proyectos sería recomendable realizar una evaluación continua de las consecuencias de dejar de lado ciertos objetivos en los resultados finales y su impacto en la aceptación del proyecto por parte de los *stakeholders*.

6.2 Líneas de futuro

Concluido el proyecto actual, es importante identificar futuras líneas de ampliación del trabajo que no solo aborden las limitaciones encontradas, sino que también amplíen el alcance y la eficacia del sistema desarrollado:

- Desarrollo de herramientas de visualización interactiva; la visualización de datos es un área que quedó pendiente en el proyecto actual. El desarrollo de herramientas de visualización interactiva y paneles de control que permitan a los usuarios finales, como gerentes de gimnasios o equipos de marketing, interactuar con los datos y obtener insights en tiempo real, sería un paso fundamental para mejorar la interpretabilidad y la usabilidad de los resultados obtenidos.

- Integración de nuevas fuentes de datos y corrección de variables; esto podría además de mejorar la calidad de los datos disponibles sobre precios y servicios consumidos, integrar datos de redes sociales, dispositivos de seguimiento de fitness o encuestas de satisfacción del cliente, proporcionando una visión más completa de las preferencias y comportamientos de los clientes, y de sus interacciones con la compañía. El análisis de estos datos puede revelar patrones y tendencias no detectados, mejorando la precisión de los modelos de predicción.

- Aplicación de técnicas de Deep Learning; ligada con el punto anterior que comporta la recolección de grandes volúmenes de datos no estructurados, la aplicación de técnicas avanzadas de deep learning y redes neuronales profundas con mayor capacidad de generalización para mejorar los modelos predictivos es una dirección interesante de investigación. Quedó fuera de este proyecto por la dificultad en su interpretabilidad y porque los modelos usados demostraron rendimiento más que suficiente.

- La evaluación de la viabilidad técnica y económica, uno de los objetivos pendientes, es esencial para la implementación exitosa del proyecto en un entorno real. Esto debería incluir un análisis detallado del retorno de la inversión, los costos operativos y la escalabilidad del sistema que proporcione una base sólida para argumentar a favor de la implementación del sistema.

- La realización de estudios de caso y pruebas piloto en gimnasios específicos debería proporcionar retroalimentación práctica y datos de rendimiento en entornos reales. Esto no solo ayudaría a refinar el sistema propuesto, sino que también podría revelar desafíos y oportunidades únicos en distintas ubicaciones.

- A nivel teórico, sería interesante también incorporar el estudio previo del cumplimiento de hipótesis del modelo estadístico a pesar de que se ha obviado por el buen rendimiento del modelo.

Las futuras líneas de trabajo en el proyecto deberían centrarse pues en mejorar la accesibilidad y comprensión de los datos, expandir el alcance de los análisis mediante la integración de nuevas fuentes de datos y técnicas de modelado más avanzadas, y validar la efectividad y eficiencia del sistema en entornos reales. Esta ampliaciones no solo mejorarían la capacidad del sistema para predecir y manejar la retención , sino que también asegurarían su relevancia y aplicabilidad en el sector.

6.3 Seguimiento de la planificación

Este análisis se centra en cómo las dificultades encontradas en las etapas iniciales del proyecto impactaron la planificación y condujeron a una reestructuración de las prioridades y objetivos.

El proyecto enfrentó en una primera fase desafíos significativos debido a la complejidad en la interpretación, análisis y fusión de los conjuntos de datos iniciales. Los datos de clientes del centro de que se disponía no estaban identificados, mostraban incoherencias y ventanas de observación distintas. En la limpieza y preparación de datos se probaron distintos enfoques para intentar conseguir un conjunto de datos sólido sobre el que trabajar, pero resultó ser más complicado de lo anticipado por su inconsistencia, lo que afectó a la planificación original y a la duración estimada de esta fase, que en lugar de durar unas cuatro semanas se alargó más de dos meses. Se tuvo que re-obtener los datos y ajustar el proceso de limpieza al nuevo formato.

Otro obstáculo importante fue el desconocimiento inicial del entorno y Python como lenguaje de programación utilizado. Una curva de aprendizaje para adquirir competencia en herramientas y lenguajes de programación más lenta de lo esperada dificultó el avance y la progresión especialmente en las etapas iniciales del proyecto.

Finalmente, también hubo dificultad en identificar y tratar el problema en cuestión del proyecto, el análisis de retención de clientes, como un problema sujeto de ser tratado como problema de análisis de supervivencia, así como la necesidad de un periodo para estudiar y entender este modelo, qué enfoques y herramientas se podrían utilizar y cómo debían ser tratados los datos con censura.

Ante estos desafíos, se tuvo que reformular la planificación original, asignando más tiempo del previsto a las tareas de preparación y análisis de datos. Esta reasignación de tiempo y recursos tuvo un efecto dominó en las etapas posteriores del proyecto, en particular en la etapa de entrenamiento y evaluación de modelos, que se acortó a dos semanas. Para compensar el tiempo adicional dedicado a las primeras etapas, se tomaron decisiones estratégicas paralelizando otras tareas, como el estudio teórico sobre análisis de supervivencia y la redacción de la memoria del proyecto. Esta adaptación quiso reflejar una gestión de proyecto ágil y centrada en mantener el progreso hacia los objetivos generales.

Afortunadamente, y gracias al estudio anterior, la elección de los modelos de Regresión de Cox, Random Survival Forest, Survival SVM y Survival Gradient Boosting fue acertada y su entrenamiento y evaluación se desarrolló sin demasiados contratiempos en la segunda fase del proyecto consiguiendo resultados en su desempeño muy positivos.

Sin embargo, como se desarrolla en el siguiente punto, esta reestructuración también llevó a la decisión de dejar de lado los objetivos específicos 7 (desarrollo de visualizaciones y paneles de control) y 9 (evaluación de la viabilidad técnica y económica); esto permitió mantener el enfoque en los objetivos prioritarios, aún prescindiendo de componentes del proyecto que podrían haber añadido un valor considerable en términos de usabilidad y justificación del proyecto y que se proponen como futuras líneas de ampliación.

En resumen, el seguimiento de la planificación en este proyecto pone de manifiesto la necesidad de un equilibrio entre rigurosidad y flexibilidad en la gestión de proyectos de IA. A pesar de los retos y cambios en la planificación, se logró progresar significativamente en los objetivos propuestos de mayor relevancia.

6.4 Objetivos no conseguidos

Este análisis detalla los logros alcanzados en la búsqueda del proyecto de un sistema de análisis de datos basado en inteligencia artificial (IA) para predecir la baja de abonados en un centro deportivo y determinar su tiempo de permanencia y las áreas pendientes.

Entre los objetivos cumplidos se encuentran:

- El Objetivo General (Obj.G) se ha abordado de manera efectiva, estableciendo un sistema para la predicción de bajas de abonados y la determinación de su tiempo de permanencia sólidamente fundamentado en el modelo teórico de análisis de supervivencia y con la aplicación exitosa de cinco algoritmos de aprendizaje automático

- EL Obj.1 de anonimización, recolección y limpieza de datos se ha completado satisfactoriamente tras la recopilación y limpieza de datos de membresía y uso del centro deportivo así como la investigación y aplicación de técnicas de tratamiento de outliers e imputación de valores nulos como KNN o MICEK para asegurar la calidad e integridad de los datos.

- El Obj.2 de realización de análisis exploratorio de datos (EDA) se llevó a cabo de manera exhaustiva, analizando características categórica sy numéricas mediante las estadísticas correspondientes y obteniendo las representaciones gráficas que mejor mostraban los aspectos relevantes a considerar del conjunto de datos como el comportamiento estadísticamente distinto de las características de los clientes que según se habían dado o no de baja.

- Se identificaron y abordaron los posibles problemas de privacidad de datos que podían surgir, garantizando la conformidad del estudio con las regulaciones de protección de datos aplicables para el cumplimiento del Obj.3.

- Se realizó un estudio detallado (Obj.4) de los modelos estadísticos apropiados para el análisis de supervivencia, incluyendo regresión de Cox, Random Survival Forest, Survival Gradient Boosting y Survival SVM, adecuándose al contexto del proyecto.

- El entrenamiento de los modelos mencionados se ejecutó utilizando técnicas de validación cruzada (K-Fold) y las métricas de C-Index y AUC-ROC se emplearon para evaluar la precisión y el rendimiento, consiguiendo resultados prometedores durante la consecución del Obj.5 de entrenamiento y evaluación de modelos de IA.

- El análisis de importancia de características (Obj.6) se llevó a cabo para conseguir identificar las variables más influyentes en la predicción de las bajas de clientes y el tiempo de permanencia.adaptando la identificación a las características del modelo; relacionándolas con los coeficientes de regresión el modelo de regresión de Cox, …

Se ha documentado meticulosamente el desarrollo del proyecto desde la adquisición de datos hasta la implementación de modelos, y se ha trabajado siguiendo la metodología CRISP propuesta, adaptando la planificación prevista a las necesidades y eventualidades del proyecto.

Como objetivos pendientes de cumplimiento quedan:

- El desarrollo perteneciente al Obj.7 de herramientas de visualización de datos y paneles de control del para la comunicación de los resultados y la extrapolación de recomendaciones para mejorar la retención de clientes permanece inconcluso. Este aspecto es crucial para la posibilidad de implementación del proyecto en un entorno real y la interpretación y aplicación práctica de los hallazgos. La consecución objetivo se descarta en la segunda fase durante el reajuste de planificación necesario para la consecución del resto de objetivos.

- La realización del Obj. 9 de la evaluación de la viabilidad técnica, económica y el retorno de la inversión del sistema propuesto. Esta evaluación sería igualmente esencial para determinar la factibilidad del despliegue del sistema en un entorno operativo y para asegurar el apoyo de las partes interesadas pero se descarta en una etapa temprana por considerarse no necesario en un proyecto aún en ámbito académico.

El proyecto ha alcanzado la mayoría de sus objetivos con éxito, estableciendo un camino prometedor hacia el uso de algoritmos de IA en la predicción efectiva de la baja de clientes en centros deportivos mediante el uso de. Quedan pendientes propuestas de ampliación inherentes a su implantación en un entorno real en las que se debería invertir esfuerzo para maximizar su impacto y relevancia en el sector deportivo.

7. Glosario

Árbol de decisión: Modelo de aprendizaje automático basado en reglas de decisión representadas en una estructura de árbol.

Aprendizaje automático: Campo de la inteligencia artificial que desarrolla algoritmos que permiten a las máquinas aprender patrones a partir de datos.

Aprendizaje supervisado: Tipo de aprendizaje automático con modelos entrenados usando datos etiquetados.

Censura: Técnica que se utiliza cuando no se observa la ocurrencia del evento de interés para todos los individuos en el estudio.

C-Index: Índice de concordancia utilizado en el análisis de supervivencia para evaluar la capacidad predictiva de un modelo.

Churn-rate: Tasa de abandono, mide la proporción de clientes que dejan de utilizar un servicio en un período.

Coeficiente de Regresión: Representa la relación entre la variable de respuesta y las variables predictoras en un modelo de regresión.

Curva AUC: Métrica que evalúa la capacidad de discriminación de un modelo, especialmente en clasificación binaria.

Dataset: Conjunto de datos utilizado para entrenar y evaluar modelos de aprendizaje automático.

Especificidad: Capacidad del modelo para identificar correctamente los casos negativos.

Función de Riesgo: En el análisis de supervivencia, indica la tasa instantánea de falla en un momento dado.

Función de Supervivencia: Proporciona la probabilidad de que un evento no ocurra antes de un tiempo específico.

Gradient Boosting: Técnica de aprendizaje automático que construye un conjunto de modelos secuencialmente.

K-Fold: Técnica de validación cruzada que divide el conjunto de datos en k partes para entrenamiento y validación.

KNN (K-Nearest Neighbors): Algoritmo de clasificación basado en la mayoría de los vecinos más cercanos.

MICE (Multiple Imputation by Chained Equations): Método de imputación para tratar valores faltantes en datos.

Precisión: Métrica que mide la proporción de verdaderos positivos entre todos los casos positivos predichos por el modelo.

Random Forest: Método que combina múltiples árboles de decisión para mejorar la precisión.

Regresión de Cox: Modelo para evaluar el impacto de múltiples variables en el riesgo de un evento.

Scikit-learn: Biblioteca de aprendizaje automático en Python.

Scikit-survival: Extensión de scikit-learn para problemas de supervivencia.

Sensibilidad: Capacidad del modelo para identificar correctamente los casos positivos.

SVM (Support Vector Machine): Algoritmo de aprendizaje automático para clasificación y regresión.

Time-life Value: Valor monetario esperado que un cliente aportará durante su relación con un negocio.

8. Bibliografía

Albanese, N. C. (2022). How to Evaluate Survival Analysis Models. Towards Data Science. Recuperado el 8 de enero de 2024, de https://towardsdatascience.com/how-to-evaluate-survival-analysis-models-dd67bc10caae#e5ec

Analyzing Customer Churn With Lifelines. (2018). GitHub. Recuperado el 8 de enero de 2024, de https://github.com/zangell44/survival-analysis-lifeline-basics/blob/master/customer_churn.ipynb

Bonaplata, A. (2019). Análisis exploratorio de datos con Python. LinkedIn. Recuperado el 8 de enero de 2024, de https://es.linkedin.com/pulse/análisis-exploratorio-de-datos-con-python-almudena-bonaplata

Castrillo, M. (2023). Cómo identificar y tratar outliers con Python. Medium. Recuperado el 8 de enero de 2024, de https://medium.com/@martacasdelg/cómo-identificar-y-tratar-outliers-con-python-bf7dd530fc3

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B, 34, 187-202.

Customer Churn Project Daniela. (2020). Kaggle. Recuperado el 8 de enero de 2024, de https://www.kaggle.com/code/darango94/churn-de-clientes-project-daniela/

De la Calle, J. E. (2022). Análisis de supervivencia: una herramienta oculta pero clave para el marketing. Medium. Recuperado el 8 de enero de 2024, de https://juandelacalle.medium.com/análisis-de-supervivencia-una-herramienta-oculta-pero-clave-para-el-marketing-e5920bcfdd8f

Ejemplo de uso de DBSCAN en Python para detección de outliers. (2019). Exponentis. Recuperado el 8 de enero de 2024, de http://exponentis.es/ejemplo-de-uso-de-dbscan-en-python-para-deteccion-de-outliers

Fernández, C., Ramirez Teodoro L. A., & Sofía Villers Gómez, S. (Fecha no especificada). Modelos de supervivencia. Recuperado el 8 de enero de 2024, de https://carlosfernandovg.github.io/supervivencia_y_series_FC2021-1/

Gold, C. S. (2020). Fighting Churn with Data: The Science and Strategy of Customer Retention. Manning Ed.

Handling Missing Values for Machine Learning. (2023). Kaggle. Recuperado el 8 de enero de 2024, de https://www.kaggle.com/discussions/questions-and-answers/353100

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. JAMA, 247(18), 2543-2546.

Identificación de valores perdidos en Python. (2021). El Mundo de los Datos. Recuperado el 8 de enero de 2024, de https://elmundodelosdatos.com/identificacion-valores-perdidos-python/

Modelo Predictivo: Abandono de Clientes. (2021). RPubs. Recuperado el 8 de enero de 2024, de https://rpubs.com/arojasmor17/abandono

Morillo Leal, J. (2023). Integración de un modelo de análisis de supervivencia para la predicción de la pérdida de clientes en seguros. Universidad Politécnica de Madrid. Recuperado el 8 de enero de 2024, de https://oa.upm.es/76129/1/TFM_JAIME_MORILLO_LEAL.pdf

Orozco, A. (2021). Modelo Predictivo: Abandono de Clientes. RPubs. Recuperado el 8 de enero de 2024, de https://rpubs.com/arojasmor17/abandono

Pölsterl, S. (2020). scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research, 21(212), 1-6. Recuperado el 20 de enero de 2024, de https://scikit-survival.readthedocs.io/en/stable/index.html

Rojas, P. (2018). Análisis de supervivencia. RPubs. Recuperado el 8 de enero de 2024, de https://rstudio-pubs-static.s3.amazonaws.com/438542_64aa278b60684f69be38236a5b58711f.html

Segura Moreno, J. A. (2022). Desarrollo de un modelo de predicción de fuga de clientes y diseño de experimento para la aplicación de estrategias de fidelización en factoring. Universidad de Chile. Recuperado el 8 de enero de 2024, de https://repositorio.uchile.cl/bitstream/handle/2250/186972/Desarrollo-de-un-modelo-de-prediccion-de-fuga-de-clientes-y-diseno-de-experimento-para-la.pdf

Silva Allende, A. (Fecha no especificada). Experimental Comparison of Semi-parametric, Parametric, and Machine Learning Models for Time-to-Event Analysis Through the Concordance Index. GitHub. Recuperado el 8 de enero de 2024, de https://colab.research.google.com/github/alonsosilvaallende/Random-Survival-Forest-GBCSG2/blob/master/Cox_PH_and_RSF_colab.ipynb

Sobreiro, P. N. (2023). Predicción de abandono de clientes mediante modelos de aprendizaje automático de supervivencia híbridos. Tesis doctoral. Universidad de Extremadura. Recuperado el 8 de enero de 2024, de https://dehesa.unex.es/bitstream/10662/16726/1/TDUEX_2023_Sobreiro_PN.pdf

Survival Analysis for Deep Learning. (2023). GitHub. Recuperado el 8 de enero de 2024, de https://colab.research.google.com/github/sebp/survival-cnn-estimator/blob/master/tutorial_tf2.ipynb

Tutorial: Creación, evaluación y puntuación de un modelo de predicción de renovación. (2023). Microsoft Fabric. Recuperado el 8 de enero de 2024, de https://learn.microsoft.com/es-es/fabric/data-science/customer-churn

Vaquerizo, R. (2019). El análisis de supervivencia para segmentar el churn. Análisis y Decisión. Recuperado el 8 de enero de 2024, de https://analisisydecision.es/el-analisis-de-supervivencia-para-segmentar-el-churn/

Visualización de datos con Python. (2023). GitHub. Recuperado el 8 de enero de 2024, de https://joserzapata.github.io/courses/python-ciencia-datos/visualizacion/

Zapata, J. R. (2023). Visualización de datos con Python. GitHub. Recuperado el 8 de enero de 2024, de https://joserzapata.github.io/courses/python-ciencia-datos/visualizacion/

Analizando el abandono de clientes con Python. (2020). Blog Visionarios. Recuperado el 8 de enero de 2024, de https://blogvisionarios.com/articulos-data/analizando-el-abandono-de-clientes-con-python/

Anexos

ANEXO I. Catálogo de variables

Dataset data_clientes

Contiene información acerca de los abonados y exabonados del centro y sus suscripciones, sin identificarlos, con varias columnas con información tal como fecha de alta y baja, tipo de cuota y forma de pago, edad, sexo, nacionalidad, etc.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
1.1	Abonat	Número identificador del abonado	Entero	Num Discreta	2.3 - 3.1 - 4.1 - 5.1 - 6.1
1.2	CentreOrigen	Código del centro de origen	Entero	Num Discreta	-
1.3	DataAlta	Fecha de alta del abonado	Fecha y hora	Temporal	-
1.4	DataBaixa	Fecha de baja del cliente	Fecha y hora	Temporal	5.4
1.5	BaixaDefinitiva	Indicador de baja definitiva	Entero	Cat Nominal	5.3
1.6	Modalitat	Código de la cuota de inscripción	Entero	Num Discreta	2.5 - 5.5
1.7	Descripcio	Nombre de la cuota de suscripción	Objeto	Cat Nominal	2.6 - 5.6
1.8	ModalitatPagament	Modalidad de pago de la cuota	Objeto	Cat Nominal	2.7 - 7.1
1.9	Sexe	Género del abonado	Entero	Cat Nominal	-
1.10	EdatAra	Edad actual del abonado	Entero	Num Continua	-
1.11	CodiPostal	Código postal del abonado	Objeto	Cat Nominal	-
1.12	Nacionalitat	Nacionalidad del cliente	Entero	Cat Nominal	-

Dataset data_altes_i_baixes

Contiene información sobre altas y bajas de abonos en el centro deportivo. Registra datos como fechas, identificadores de abonados y ex-abonados, operadores, códigos y nombres de modalidades, así como los métodos de pago utilizados.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
2.1	Tipus	Tipo de registro de ALTA o BAJA	Objeto	Cat Nominal	-

Id	Nombre	Descripción	Formato	Tipo	Relacionada
2.2	Data	Fecha y hora de solicitud del registro	Fecha y hora	Temporal	-
2.3	Abonat	Número identificador del abonado	Entero	Num Discreta	1.1 - 3.1 - 4.1 - 5.1 - 6.1
2.4	Operador	Identificador del operador	Entero	Num Discreta	-
2.5	ModalitatCodi	Código de la cuota de suscripción	Entero	Num Discreta	1.6 - 5.5
2.6	ModalitatNom	Nombre de la cuota de suscripción	Objeto	Cat Nominal	1.7 - 5.6
2.7	ModalitatPagament	Modalidad de pago de la cuota	Objeto	Cat Nominal	1.8 - 7.1

Dataset data_consum_altres_serveis

Los abonados, exabonados y no abonados que aparecen en este dataset, identificados por su id u otra documentación, han sido consumidores de otros servicios en el centro.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
3.1	NUM_CLIENTE	Número de cliente	Objeto	Num Discreta	1.1 - 2.3 - 4.1 - 5.1 - 6.1

Dataset data_darrer_rebut

Contiene información acerca del último recibo emitido a los abonados y exabonados del centro. No incluye, por tanto, los pagos en caja.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
4.1	Abonat	Número identificador del abonado	Entero	Num Discreta	1.1 - 2.3 - 3.1 - 5.1 - 6.1
4.2	DataFacturacio	Fecha de facturación del último recibo emitido	Fecha y hora	Temporal	-
4.3	ImportTotal	Importe del recibo	Real	Num Continua	-

Dataset data_entradesclaret

Registra entradas de abonados y exabonados desde la apertura del centro, con detalles como el número identificador del abonado, fechas y horas de entrada y salida, fecha de baja del abonado y la cuota asociada en el momento de la entrada.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
5.1	Abonat	Número identificador del abonado	Entero	Num Discreta	1.1 - 2.3 - 3.1 - 4.1 - 5.1 - 6.1
5.2	DataHora	Fecha y hora de la entrada	Objeto	Temporal	-
5.3	BaixaDefinitiva	Indicador de baja definitiva	Real	Cat Nominal	1.5
5.4	Databaixa	Fecha de baja	Objeto	Temporal	1.4
5.5	codimodalitat	Código de la cuota de suscripción en el momento de la entrada	Real	Num Discreta	1.6 - 2.5
5.6	Modalitat	Nombre de la cuota en el momento de la entrada	Objeto	Cat Nominal	1.7 - 2.6

Dataset data_reservas

Almacena información detallada sobre reservas de actividades y servicios realizadas a través de la app del centro, puesta en marcha a partir de la pandemia. Incluye nombres de clientes, identificadores de abonados, fechas de inicio de uso de la app, estado del cliente, detalles del nombre, tipo de ubicación, fecha y hora, asistencia y facturación de la actividad y su profesorado, así como costos y características de servicios de pago.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
6.1	Client Alt ID	Número identificador del abonado	Entero	Num Discreta	1.1 - 2.3 - 3.1 - 4.1 - 5.1 - 6.1
6.2	Joined Date	Fecha de inicio de alta en la app	Fecha y hora	Temporal	-
6.3	Client Status	Estado del cliente	Objeto	Cat Nominal	-
6.4	Sub Location	Lugar de la actividad	Objeto	Cat Nominal	-
6.5	Service Category	Categoría del servicio o actividad	Objeto	Cat Nominal	-
6.6	Service Title	Nombre del servicio o actividad	Objeto	Cat Nominal	-
6.7	Activity	Tipo de actividad	Objeto	Cat Nominal	-
6.8	Trainer Alt ID	Identificación del entrenador	Real	Num Discreta	-
6.9	Date of Session	Fecha de la sesión	Fecha y hora	Temporal	-

Id	Nombre	Descripción	Formato	Tipo	Relacionada
6.10	Time of Session	Hora de la sesión	Objeto	Temporal	-
6.11	Attendance	Estado de asistencia	Objeto	Cat Nominal	-
6.12	Billed	Si el servicio ha sido cobrado	Objeto	Cat Nominal	-
6.13	Booked on	Fecha de reserva	Fecha y hora	Temporal	-
6.14	Booked at	Hora de reserva	Objeto	Temporal	-
6.15	Device	Dispositivo desde el que se realiza	Objeto	Cat Nominal	-
6.16	Device.1	Otro dispositivo desde el que se accede	Objeto	Cat Nominal	-
6.17	Ref ID	Identificador del registro de reserva	Entero	Num Discreta	-
6.18	Client ID	Identificador del cliente en la app	Entero	Num Discreta	-
6.19	CSP ID	Identificación del paquete de servicios	Real	Num Discreta	-
6.20	CSP Package AltID	Código Identificador del tipo de paquete de servicios	Objeto	Cat Nominal	-
6.21	CSP cost	Costo del paquete de servicios	Real	Num Continua	-

Dataset data_formaspago

Contiene información sobre el código, los meses y el descuento incluido en un a determinada forma de pago.

Id	Nombre	Descripción	Formato	Tipo	Relacionada
7.1	ModalitatPagament	Forma de pago de la cuota	Objeto	Cat Nominal	1.8 - 2.7
7.2	MesesModalidadPago	Meses totales que incluye	Entero	Num Discreta	-
7.3	Descuento	Descuento que se aplica sobre el precio mensual	Real	Num Continua	-

ANEXO II. Conjunto de datos final

Dataset data_reducido

El dataset resultante a partir del cual se entrenan los modelos.

Id	Nombre	Descripción	Formato
dr.1	CentreOrigen	Centro de origen	Numérica
dr.2	BaixaDefinitiva	Indicador de Baja	Numérica
dr.3	Sexe	Género	Numérica
dr.4	Nacionalitat	Nacionalidad	Numérica
dr.5	numero_accesos	Número de accesos	Numérica
dr.6	dia_habitual	Día habitual de visita	Categórica
dr.7	franja_habitual	Franja horaria habitual	Categórica
dr.8	ServiciosExtra	Consumo de servicios extra	Numérica
dr.9	numero_reservas	Número de reservas	Numérica
dr.10	antelacion_promedio	Antelación promedio de reserva	Numérica
dr.11	Attended	% Asistencias confirmadas	Numérica
dr.12	Cancelled	% Reservas canceladas	Numérica
dr.13	NotAttended	% No asistencias	Numérica
dr.14	Fitness	% Uso de área de fitness	Numérica
dr.15	PersTraining	% Entrenamiento personal	Numérica
dr.16	GroupExercise	% Uso de clases colectivas	Numérica
dr.17	Nutrition	% Uso del servicio de nutrición	Numérica
dr.18	act_preferida	Actividad preferida	Categórica
dr.19	CategoriaModalitat	Categoría de modalidad	Categórica
dr.20	Descuento	Descuento aplicado	Numérica
dr.21	DiasModalitatPago	Días modalidad de pago	Numérica
dr.22	Mes_de_Alta	Número de mes de alta	Numérica
dr.23	dias_totales	Días totales desde alta	Numérica
dr.24	period_primer_acces	Días hasta primer acceso	Numérica
dr.25	period_solic_baixa	Días hasta solicitud de baja	Numérica

Id	Nombre	Descripción	Formato
dr.26	percent_consumit	Porcentaje de abono consumido	Numérica
dr.27	dias_sin_venir	Días sin visitar el centro	Numérica
dr.28	ratio_inactivo	Ratio de inactividad	Numérica
dr.29	ratio_primer_us	Ratio hasta el uso inicial	Numérica
dr.30	num_inscripcions	Número de inscripciones anteriores	Numérica
dr.31	Edat_Alta	Edad en el momento de alta	Numérica
dr.32	Distancia	Distancia al centro	Numérica
dr.33	freq_activa	Frecuencia semanal de visita	Numérica
dr.34	freq_30dias	Frecuencia en los últimos 30 días	Numérica
dr.35	freq_90dias	Frecuencia en los últimos 90 días	Numérica
dr.36	freq_180dias	Frecuencia en los últimos 180 días	Numérica
dr.37	PreuBase	Precio base de la cuota	Numérica
dr.38	CostMensual	Costo mensual final	Numérica
dr.39	existe_app	% convivencia con la aplicación	Numérica