Universitat Oberta de Catalunya

uoc.edu

TFG Área de Inteligencia Artificial

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

Autor: Andrés Espín Rearte
Tutor: David Isern Alarcón
Responsable de la asignatura: Susana Acedo Nadal

Contenido

1. **Contexto y justificación**
   a. Introducción: el *churn rate* en los gimnasios. Objetivos del proyecto.
   b. El análisis de supervivencia
   c. Planificación y metodología

2. **Desarrollo del proyecto**
   a. Origen de los datos y preprocesado
   b. Criterios para la selección de modelos
   c. Entrenamiento de los modelos
   d. Evaluación: C-Index, Brier, AUC

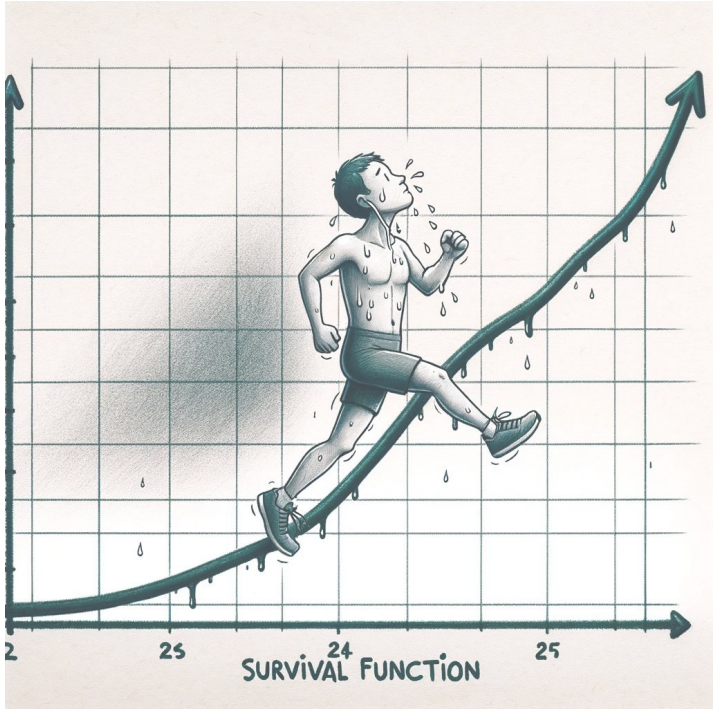
3. **Resultados obtenidos**
   a. Ejemplos de visualizaciones.
   b. Principales hallazgos y patrones observados
   c. Análisis de características

4. **Conclusiones del proyecto**
   a. Consecución de objetivos
   b. Líneas de futuro y aplicabilidad

Contexto y justificación

a. Introducción: El *churn-rate* en los gimnasios

- OMS: hasta 5 millones de muertes al año. Se estima que alrededor del 25% de los adultos y más del 80% de los adolescentes a nivel mundial **no alcanzan los niveles de actividad física**, lo que aumenta el riesgo de muerte en un 20% a 30%
- 4.000 clubs en España con 5,5 millones de socios según IHRSA
- Foco en la captación, **mejora pendiente en la retención**
- Tiempo de vida medio de un abonado: 7 meses
- "***churn rate***" y "***lifetime value***"



a. Introducción: Objetivo del proyecto

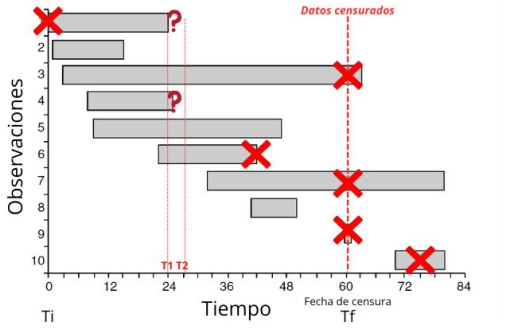
Objetivo General

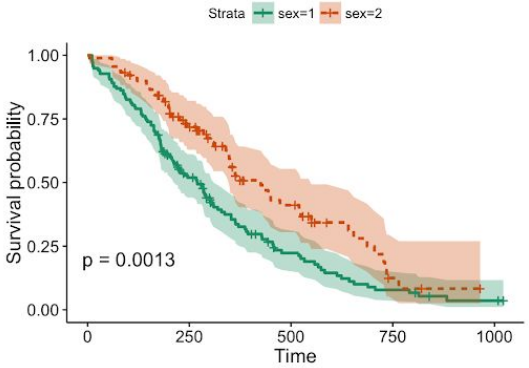
Desarrollar un sistema de análisis de datos basado en IA con el fin de predecir futuras bajas de abonados y determinar el tiempo de permanencia de los mismos.

b. El análisis de supervivencia

- **Origen:** se origina en la bioestadística para estudiar la **supervivencia de pacientes** con el modelo de Kaplan-Meier (1958).
- **Método:** técnicas estadísticas para manejar **datos censurados** y modelar el **tiempo hasta que ocurre un evento** de interés, utilizando diversas pruebas y modelos para la interpretación como el modelo de Cox.

  No-paramétricos vs paramétricos.
- **Áreas de Aplicación:** investigación médica, estudios clínicos, ciencias de la vida y en ingeniería de fiabilidad.





b. El análisis de supervivencia

- El **modelo de riesgos proporcionales de Cox**

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

Supuesto: el efecto relativo de una variable predictora sobre la tasa de riesgo debe ser constante a lo largo del tiempo.

- Algoritmos de **ML** adaptados para **trabajar con datos censurados**
  - **Random Survival Forest (RSF)**
  - **Support Vector Machines for Survival Analysis (SVM-SA)**
  - **Gradient Boosting Survival Analysis**
  - Survival Neural Networks (SNN)
  - Competing Risks Regression Models

c. Planificación y metodología

FASE 0

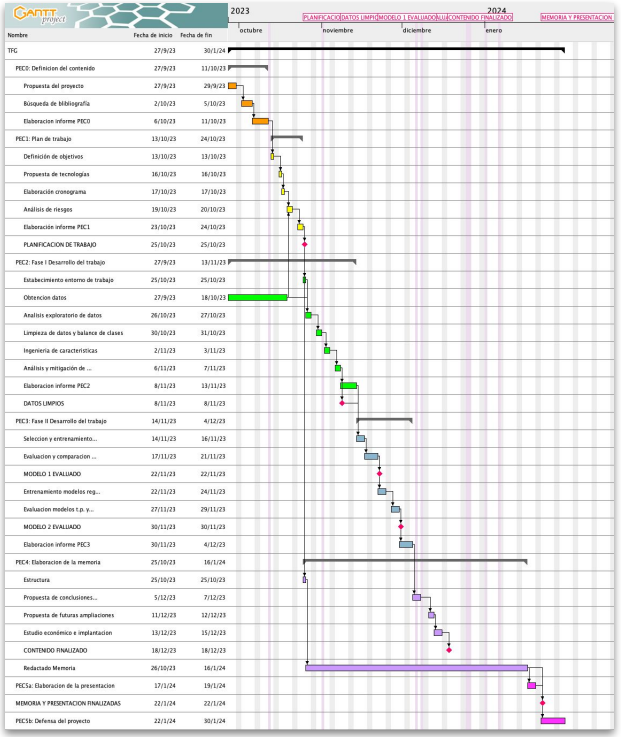
- Recolección de datos a partir de las BBDD del centro y estudio de estado del arte

FASE 1

- Anonimización y preparación datos:
- Análisis exploratorio y gráfico de los datos

FASE 2

- Selección y entrenamiento de algoritmos de aprendizaje automático adaptado al análisis de supervivencia
- Evaluación de la precisión de estos modelos
- Interpretación de resultados y análisis de características



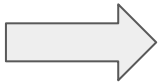
Desarrollo del proyecto

a. Origen de los datos y preprocesado

Origen

- BBDD del centro deportivo: 7 datasets

  'Clientes' '*Reservas*'
  '*Entradas*' '*Servicios Extra*'
  '*Última facturación*' '*Formas de pago*'
  '*Registro de altas y bajas*'

  `'Abonat'` y `'DataAlta'`

- Anonimización de datos
- Características propias del centro: variabilidad de cuotas, rango de edad, etc.

Preprocesado

- Ingeniería de variables:
  - Respuesta: `'dias_totales'` y `'BaixaDefinitiva'`
  - Predictoras: `'freq'`, `'mes_de_alta'`, `'cost_mensual'`, `'EdatAlta'` y fechas
- Interpretación de valores faltantes
- One Hot Encoding
- Selección de variables correlacionadas y ventana de tiempo

b. Criterios para la selección de modelos

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

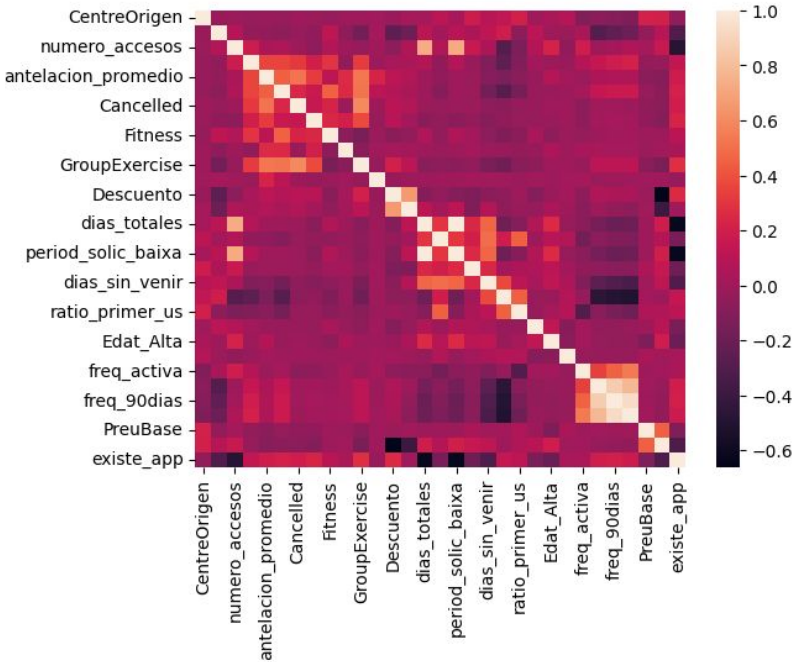
c. Entrenamiento de los modelos

- Análisis de correlación de variables para evitar la no convergencia.
  ej: `'period_solic_baixa'`

- División en conjunto de entrenamiento y de test al 20%

```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)
```

- Validación cruzada con K-Fold para el cálculo del C-Índex



d. Métricas de evaluación

C-index: proporción de pares de eventos ordenados suceden en el mismo orden que la predicción realizada

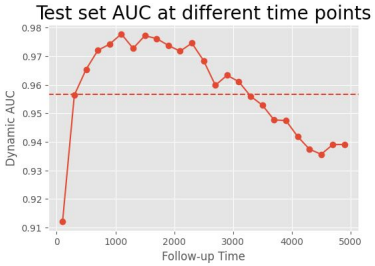
```
El c-index de 'CoxPHSurvivalAnalysis' es: 0.8749683570692541
El c-index de 'CoxnetSurvivalAnalysis' es: 0.8799080046935773
El c-index de RandomSurvivalForest' es: 0.8719596727599783
El c-index de 'FastKernelSurvivalSVM' es: 0.7605369098035011
El c-index de 'FastSurvivalSVM' es: 0.9108577691212646
El c-index de 'GradientBoostingSurvivalAnalysis' es: 0.9547241778794
```

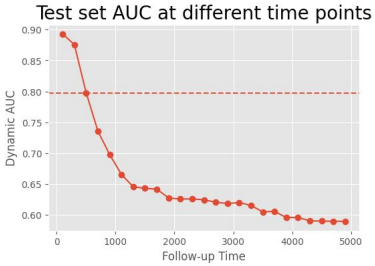
Brier Score: precisión de las predicciones en todo el rango de tiempos

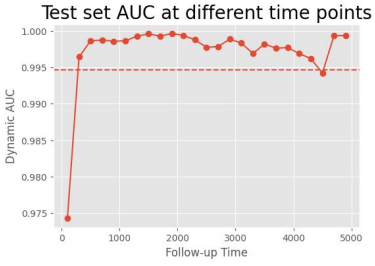
```
IBS for CoxPHSurvivalAnalysis: 0.16034807337139584
IBS for CoxnetSurvivalAnalysis: 0.14479654969393985
IBS for RandomSurvivalForest: 0.12252872121701221
IBS for GradientBoostingSurvivalAnalysis: 0.12548748825038
```

AUC-ROC dinámica mide la capacidad a lo largo de varios *t* y permite evaluar cómo cambia la precisión predictiva del modelo

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







Resultados obtenidos

a. Resultados. Ejemplos de visualizaciones

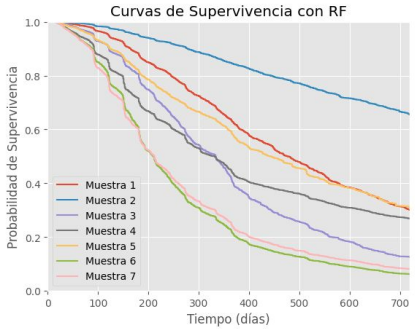
- Tiempo: la biblioteca scikit survival ofrece métodos para la visualización de:
  - Función de supervivencia

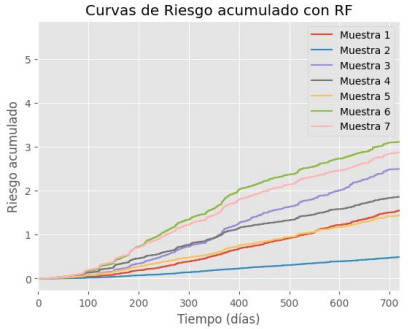
$$S(t) = P(T > t)$$

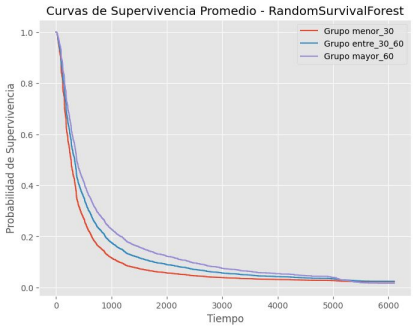
  - Función de riesgo acumulado

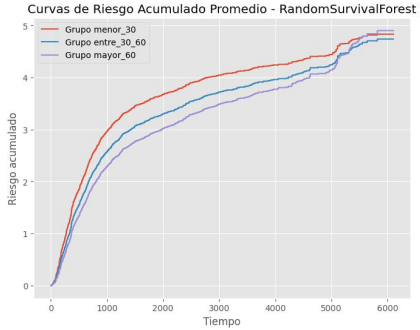
$$H(t) = \int_0^t h(u)\,du = -log(S(t))$$

- Evento: la biblioteca scikit survival ofrece el método para obtener el índice acumulado de riesgo para cada instancia *.predict(x)*



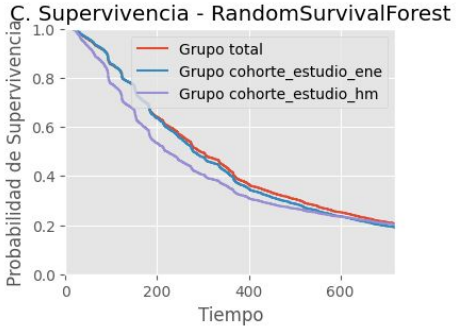


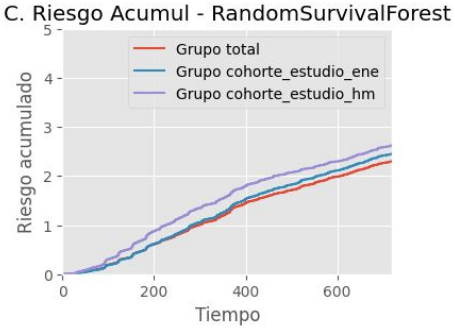




a. Resultados: Un caso de uso

Estudio comparativo de riego de la cohorte formada por ***hombres entre 40 y 50 años de edad que se inscribieron en enero*** contra el total de observaciones.





Id	Prob_Supv_180_dias
833	0.399167
1090	0.448450
6700	0.529572
4103	0.552564
6785	0.628360

- No hay una diferencia significativa por **mes** sobre el total, pero si hay en ese grupo de **edad y género** - > valdrá la pena estrategias sobre todo ese colectivo

	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**

b. Análisis de características

- El método .***coef*** de los modelos de COX -> coeficientes de regresión

```
ratio_primer_us    4.173390    64.935230    CatModalitat_partnerVIP       -2.998147    0.049879
ratio_inactivo     3.403440    30.067355    franja_habitual_No_acceso     -1.503310    0.222393
```

- El método .***coef*** de los modelos SSVM -> contribución al hiperplano
- El metodo .***feature_importances*** en GBS -> cuánto contribuye una característica a mejorar la capacidad predictiva del modelo.
- El método .***permutation_importance*** *en RSF* -> es equivalente a *.feature_importances*

```
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

b. Principales hallazgos y patrones observados

- Cinco modelos ofrecen muy buenos resultados en las **métricas, por encima del 90%**, siendo el mejor GBS. FastKernelSVM no converge por colinealidad.
- **RSF capta mejor la complejidad del modelo** y el único que individualiza la función de riesgo, baja rendimiento al muy largo plazo (no interés).
- Las características de '***días sin venir***', '***frecuencia de uso activo***', '***dias primer uso***' '***ratio inactivo***' son las que mejor definen en todos los modelos la probabilidad de baja. '***número de accesos***' también pero tiene alta correlación.

  INSIGHT: Sobre la asistencia al centro deberían centrarse las estrategias de retención.
- Estas características transversales **opacan otras demográficas o de precio** (seguramente debido a su complejidad).

Conclusiones del proyecto

a. Consecución de objetivos

- **Objetivo General:** se ha conseguido un modelo con RandomSurvivalForest, 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.

- **Conclusiones y lecciones aprendidas:**
  - Adaptabilidad del modelo según características del centro. Aplicable a centros de la misma cadena pero necesidad de reentreno en otros casos.
  - Necesidad de la flexibilidad y adaptabilidad en la gestión de proyectos.
  - Priorización estratégica de objetivos y tareas.
  - Necesidad de respaldo en un modelo teórico sólido.

b. Líneas de futuro y aplicabilidad

- **Aplicabilidad**
  - Realización de estudios de caso y pruebas piloto.
  - Desarrollo de herramientas de visualización interactiva.
  - Evaluación de la viabilidad técnica y económica.
- **Mejora del modelo**
  - Integración de nuevas fuentes de datos y corrección de variables.
  - Aplicación de técnicas de Deep Learning.
  - Contraste del cumplimiento de hipótesis del modelo estadístico.
  - Consideración de incorporar datos sintéticos.

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversitat
UOCuniversitat