

# Tarea: evaluación y comparación de modelos predictivos para Titanic

## 1. Objetivos

El objetivo de esta tarea es que los estudiantes:

- Implementen y evalúen distintos modelos predictivos sobre el conjunto de datos del Titanic.
- Utilicen validación cruzada (k-fold) con  $k = 10$  para obtener métricas de rendimiento consistentes.
- Experimenten con el preprocesamiento de datos, convirtiendo variables categóricas a numéricas de forma directa.
- Comparen estrategias de imputación: rellenar valores nulos con 0 y con la media (en aquellos casos en los que tenga sentido).
- Almacenen los resultados de la evaluación (por ejemplo, precisión, matrices de confusión, etc.) en una base de datos MySQL.
- Visualicen y comparen los resultados utilizando Grafana.

## 2. Planteamiento de la tarea

Se requiere desarrollar un pipeline de machine learning que abarque las siguientes etapas:

## 2.1. 1. Preprocesamiento de datos

- **Conversión de variables categóricas:**
  - Convertir directamente las variables categóricas a numéricas.
  - Utilizar `Label Encoding` o `OneHotEncoding`.
  - Utilizar la probabilidad de la clase objetivo.
- **Imputación de valores faltantes:**
  - Imputar los valores faltantes utilizando el valor 0.
  - Imputar los valores faltantes utilizando la media para variables numéricas.

## 2.2. 3. Selección de características

- Explorar diferentes métodos de selección de características:
  - Filtrado chi-cuadrado
  - Embebido RFE
  - Envolverte con regresión logística

## 2.3. 2. Evaluación de modelos con validación cruzada

Evaluar para la combinación de cada una de las estrategias de codificación de categorías, imputación y selección de características

- Implementar para dos modelos predictivos (por ejemplo, Random Forest y XGBoost).
- Utilizar validación cruzada de 10 *folds* para evaluar el rendimiento de cada modelo.
- Obtener métricas de desempeño (por ejemplo, precisión, exactitud, recall, F1-score)

## 2.4. 4. Almacenamiento de resultados en MySQL

- Diseñar y crear una base de datos MySQL que almacene los resultados de las evaluaciones. Se recomienda incluir tablas para:
  - Información del preprocesamiento (estrategia de imputación, codificación utilizada, etc.).
  - Métricas de desempeño por modelo y por *fold*.
  - Resumen general de la evaluación (por ejemplo, promedio y desviación estándar de las métricas).

## 2.5. 5. Visualización con Grafana

- Configurar Grafana para conectarse a la base de datos MySQL.
- Crear dashboards que permitan comparar los modelos en base a las métricas almacenadas.
- Incluir gráficos de barras, líneas y/o tablas que muestren la precisión, la matriz de confusión y otras métricas relevantes para cada modelo y estrategia.

## 3. Criterios a tener en cuenta

Los trabajos serán evaluados en función de:

- La correcta implementación de las técnicas de preprocesamiento, selección de características e imputación.
- La correcta aplicación de la validación cruzada de 10 *folds* y la obtención de métricas robustas.
- La calidad y claridad del código.
- La inclusión de celdas de documentación.
- La integración y almacenamiento correcto de los resultados en la base de datos MySQL.
- La calidad de la visualización en Grafana, y la capacidad para interpretar y comparar los resultados obtenidos.