Reto Kaggle – Titanic classification

El objetivo de este proyecto es resolver el problema Titanic - Machine Learning from Disaster de Kaggle Competition (https://www.kaggle.com/competitions/titanic/overview) utilizando algoritmos de clasificación. Para ello, se seguirán las siguientes etapas:

1. Exploración y preprocesamiento de los datos:
    a. Distribuciones:
        i. Verificar si los datos están balanceados en las clases en las cuales se deben separar los datos.
        ii. Analizar la distribución de los datos categóricos y su relación con la clase "survived". Comienza a identificar características relevantes para una clasificación precisa.
        i. Comprender la distribución de los datos numéricos y determinar si es necesario aplicar procesos de normalización o estandarización.
    b. Datos faltantes:
        i. Identificar y visualizar los datos faltantes.
        ii. Decidir qué características requieren imputaciones y cuáles no. Justificar la decisión y eliminar las columnas no seleccionadas.
        iii. Aplicar técnicas de imputación para los datos faltantes. Seleccionar la mejor técnica y justificar la elección.
    c. Análisis de correlación:
        i. Realizar un análisis de correlación para decidir qué características deben mantenerse y cuáles descartarse.
    d. Transformación de datos:
        i. Convertir los datos categóricos en numéricos. Explorar diferentes métodos y seleccionar el más adecuado. Justificar la elección.

2. Clasificación:
    a. Selección de clasificadores:
        i. Elegir tres algoritmos de clasificación que se utilizarán en el proyecto. Justificar la selección de cada algoritmo.
    b. Train-test-validate split:
        i. Utilizar k-cross validation para realizar la clasificación. Seleccionar el valor de "k" y justificar la elección.
    c. Métricas de evaluación:
        i. Calcular la exactitud, precisión, matriz de confusión, curva ROC y AUC. Explicar cada una de estas métricas.
        ii. Con base en estas métricas, determinar el mejor clasificador y justificar la elección.

Los resultados de la predicción deberán subirse a Kaggle (utilizar test set y formato indicado en kaggle) para poder obtener el rendimiento de su algoritmo.

Rubrica de evaluación

Requisito	Cumple	Cumple parcialmente	No cumple
Verifica si los datos están balanceados e identifica la importancia de tener dicha información.	12.5	6.25	0
Verifica si los datos numéricos requieren procesos de normalización o estandarización, justifican la decisión e identifican la importancia de dicho proceso.	12.5	6.25	0
Identifica las características con datos faltantes, utiliza técnicas de imputación correctamente justificadas y descarta características que no sean candidatas a imputación.	15	7.5	0
Analiza la correlación entre variables y, en conjunto con el análisis de distribuciones, decide que atributos son importantes y cuales se descartan.	15	7.5	0
De los atributos restantes, realiza la conversión de datos categóricos a numéricos mediante técnicas de transformación. Justifica la selección de dichas técnicas e identifica la importancia de este proceso.	15	7.5	0
Se realiza un proceso de selección de clasificadores y se justifica la selección de cada uno de ellos.	15	7.5	0
Se compara el rendimiento de cada clasificador utilizando las métricas de desempeño solicitadas y k-cross-validation. Se decide cuál es el mejor clasificador para solucionar el problema y se justifica la selección.	15	7.5	0

Finalmente, deberán preparar una presentación que contenga toda la información necesaria para la comprensión de su proyecto. Deberá contener una descripción elaborada sobre los puntos descritos anteriormente. Esto servirá como evidencia y se subirá a e-Lumen.

Fechas relevantes:

Presentación final: miércoles 11 de septiembre a las 16:00 con una duración máxima de 10 minutos por equipo.

Examen final: jueves 12 de septiembre a las 16:00

Entrega en CANVAS de documentos y links: 11 de septiembre a las 23:59