

# Proyecto 3: *Métricas y validación cruzada*

---

Empleando la información del archivo:

- `cia_países.xlsx`

deben plantear tres modelos de aprendizaje automático supervisado (los que ustedes gusten) para realizar la clasificación de los países en:

- `ingreso-bajo`
- `ingreso-medio`
- `ingreso-alto`

conforme a su *Producto Interno Bruto* (en inglés *GDP - Gross Domestic Product*). Para ello deben resolver el planteamiento utilizando *validación cruzada de 10 partes* ( $k = 10$ ), a fin de determinar cuál es en realidad el mejor de los tres modelos planteados considerando las métricas de:

- Exactitud (*Accuracy*)
- Precisión
- Sensibilidad (*Recall*)
- $F_1$

**Observación:** ustedes deben fundamentar cuál es el mejor modelo de los planteados, dadas las métricas anteriores que la validación cruzada les reporte. Esto es, **hay un solo ganador**, no se trata de que señalen que conforme a la exactitud el mejor modelo es el  $m_1$ , con base a la precisión el mejor es  $m_2$ , con base al recall también es  $m_1$  y, según la métrica  $F_1$ , el modelo con mejor desempeño es  $m_3$ . Deben concluir cuál es el ganador y fundamentar el por qué.

---

## Evaluación

---

Para este proyecto serán evaluados dos aspectos (esto implica dos calificaciones de proyecto):

1. **Uso de la validación cruzada:** que la utilización y los resultados de la validación cruzada de 5 partes sean correctos.
2. **Determinación del modelo ganador:** fundamentar y concluir apropiadamente cuál es el modelo de aprendizaje supervisado *ganador*, de los tres que propusieron.

★ Fecha límite de entrega: **miércoles 03 de abril de 2024, 11:30 horas.**

## Información de los países del mundo

La información de cada país es:

#	Atributo	Definición
1	Name	Nombre corto del país
2	Continent	Continente en el que se encuentra el país
3	Area	Superficie en km <sup>2</sup>
4	Population	Número de habitantes
5	GDP	Producto Interno Bruto (Real GDP per capita)
6	Unemployment Rate	Tasa de desempleo (Unemployment rate)
7	Taxes	Tasa de impuestos (Taxes and other revenues)
8	External Debt	Deuda externa (Debt - external)
9	Exchange Rate	Tasa de cambio a dólares (Exchange rates in US Dollars)
10	Internet Users	Usuarios con acceso a Internet (Internet users - total)
11	Internet Users Percentage of Population	Porcentaje de la población con acceso a Internet (Internet users - percent of population)
12	Airports	Número de aeropuertos (Airports)
13	Roadways	Carreteras en servicio (km) # Roadways #
14	Militar Expenditures	Gasto militar en % del PIB (Militar expenditures: % of GDP)

**Nota.**- El atributo `gdp` debe transformarse a dato categórico (codificado en OHE) estableciendo la clasificación siguiente:

- Datos atípicos  $\Rightarrow GDP < 50$
- Ingreso bajo  $\Rightarrow 50 \leq GDP \leq 5,000$
- Ingreso medio  $\Rightarrow 5,000 < GDP \leq 25,000$
- Ingreso alto  $\Rightarrow GDP > 25,000$
- Los datos atípicos se deben eliminar para la realización de este proyecto

★ No olviden utilizar Pandas para leer el archivo `.xlsx` y realizar todo el *Data Wrangling* (incluyendo la limpieza de los datos). Además, usen la política de *imputing* de datos que mejor convenga.