# Tarea Final (1 Semana)

Curso: Métodos Estadísticos / Machine Learning

Universidad del Desarrollo

**Dataset:** Online Shoppers Purchasing Intention Dataset (UCI / Kaggle)

# Propósito (Metodología TILT)

Al finalizar esta tarea, el estudiante será capaz de:

- Diagnosticar colinealidad mediante matrices de correlación y heatmaps triangulares.
- Seleccionar variables representativas para análisis inferencial y predictivo.
- Ajustar modelos supervisados lineales y probabilísticos (RegLog, LDA, KNN, Naive Bayes).
- Evaluar significancia estadística mediante regresión logística con `statsmodels`.
- Visualizar fronteras de decisión en dos dimensiones usando variables originales.
- Aplicar modelos no supervisados (K-means, GMM, clustering jerárquico, SOM).
- Integrar conclusiones para proponer estrategias accionables de negocio.

# Tareas

## 1. Exploración y Diagnóstico de Datos

Utilice el dataset Online Shoppers Purchasing Intention Dataset.

**1.1** Describa brevemente las variables del dataset y en especial la variable objetivo `Revenue`.

**1.2** Calcule la matriz de correlación y genere un heatmap triangular:

- Triángulo inferior: colores de correlación.
- Triángulo superior: valores numéricos en formato compacto.

**1.3** Identifique bloques de colinealidad e indique una variable representativa por cada bloque (ideal: 6–10 variables). Justifique su selección.

## 2. Modelos Supervisados

### 2.1 Regresión Logística (Enfoque Predictivo con sklearn)

Usando las variables seleccionadas:

**2.1.1** Construya un pipeline con `StandardScaler`.

**2.1.2** Entrene el modelo y evalúe:

- Accuracy
- Precision
- Recall
- F1
- AUC
- Matriz de confusión

**2.1.3** Interprete los resultados brevemente.

### 2.2 Regresión Logística (Enfoque Inferencial con statsmodels)

Usando las mismas variables seleccionadas:

**2.2.1** Ajuste un modelo `Logit` con `statsmodels`.

**2.2.2** Reporte coeficientes, errores estándar, estadístico z y p-valores.

**2.2.3** Comente: *¿Coinciden los resultados inferenciales con los resultados predictivos?*

### 2.3 Modelos Comparativos

Entrene los siguientes modelos usando las variables seleccionadas:

- LDA
- KNN (con GridSearchCV para $k$ y $weights$)
- Naive Bayes Gaussiano (GaussianNB)

Para cada modelo:

**2.3.1** Entrene usando train/test.

**2.3.2** Evalúe las métricas anteriores.

**2.3.3** Muestre la matriz de confusión.

**2.3.4** Discuta cuál modelo detecta mejor la clase positiva (`Revenue = 1`) y por qué.

## 3. Fronteras de Decisión (2 Variables Originales)

**3.1** Seleccione dos variables originales (no transformadas) que sean relevantes.

**3.2** Visualice:

- Frontera de decisión de la regresión logística.
- Áreas de clasificación del modelo Naive Bayes (usando GaussianNB).

**3.3** Comente diferencias geométricas entre modelos lineales y probabilísticos gaussianos.

## 4. Modelos No Supervisados

### 4.1 K-means

**4.1.1** Evalúe $k = 2, \ldots, 10$.

**4.1.2** Seleccione $k$ según el método del codo y el silhouette promedio.

**4.1.3** Grafique y describa los clusters encontrados.

### 4.2 Gaussian Mixture Models

**4.2.1** Evalúe 2–8 componentes.

**4.2.2** Seleccione el modelo usando BIC y AIC.

**4.2.3** Grafique los clusters probabilísticos.

### 4.3 Clustering Jerárquico

**4.3.1** Genere un dendrograma utilizando enlace Ward.

**4.3.2** Seleccione una cantidad razonable de clusters y descríbalos.

### 4.4 Self-Organizing Maps (SOM)

**4.4.1** Entrene un SOM utilizando MiniSom.

**4.4.2** Muestre:

- U-Matrix
- Mapa de hits
- Distribución de `Revenue` en el mapa

## 5. Integración Final

**5.1** Resuma qué variables resultaron más relevantes.

**5.2** Compare desempeño entre modelos lineales, LDA, KNN y Naive Bayes.

**5.3** Relacione los clusters con la variable `Revenue`.

**5.4** Proponga **tres acciones de negocio** basadas en sus hallazgos.

# Criterios de Evaluación (20 puntos)

| **Criterio** | **Pts** |
|---|---|
| Exploración + análisis de colinealidad | 3 |
| Regresión logística (sklearn + statsmodels) | 4 |
| Modelos comparativos (LDA, KNN, Naive Bayes) | 4 |
| Fronteras de decisión 2D | 2 |
| Modelos no supervisados (K-means, GMM, Jerárquico, SOM) | 5 |
| Integración final + acciones de negocio | 2 |
| **Total** | **20** |

**Entrega:** informe en PDF y notebook (.ipynb o .Rmd) con todo el desarrollo.