

Universidad Rafael Landívar  
Inteligencia Artificial  
Primer Semestre 2025

## Laboratorio 4

### Probabilidad y Estadística con Titanic

#### Descripción

En este laboratorio, trabajarán con el famoso dataset del Titanic, uno de los conjuntos de datos más utilizados para el aprendizaje de ciencia de datos y análisis estadístico. El objetivo es aplicar conceptos de probabilidad y estadística para explorar las características de los pasajeros y analizar las probabilidades de supervivencia utilizando métricas descriptivas y condicionales.

Como primer paso, el estudiante debe realizar el preprocesamiento de los datos, lo que incluye:

- Limpieza (manejo de valores nulos, duplicados, entre otros).
- Transformación (creación de nuevas columnas, conversión de tipos de datos, normalización si es necesario).
- Selección y filtrado de columnas relevantes.

#### Objetivo General

Aplicar técnicas de probabilidad y estadística para analizar y extraer conclusiones del dataset del Titanic, implementando las funciones requeridas de forma modular en un archivo externo (**tarea.py**), y validando el correcto funcionamiento mediante un archivo de pruebas automatizadas (**test\_tarea.py**).

#### Objetivos Específicos

- Aplicar conceptos de espacio muestral, eventos y probabilidad condicional.
- Calcular medidas de tendencia central y dispersión.
- Realizar análisis de probabilidad de supervivencia en función de diferentes variables (sexo, clase, edad).
- Preparar código limpio, modular y documentado.
- Familiarizarse con un flujo de trabajo basado en notebooks, archivos externos y pruebas automatizadas (GitHub Classroom).

#### Requerimientos

- Implementar las funciones solicitadas en el archivo **tarea.py**.
- Efectuar preprocesamiento (tratamiento de valores nulos, limpieza de datos, etc.) en el notebook.
- Ejecutar las funciones del archivo **tarea.py** en el notebook **enunciado.ipynb** para explicar y demostrar el uso.
- Validar las funciones mediante el archivo **test\_tarea.py**, que incluye 5 pruebas unitarias para evaluar aspectos clave.



- Al menos 3 gráficos (Puedes usar Seaborn y/o Matplotlib):
  - a. Gráfico de barras: Probabilidad de supervivencia por Género. Mostrar la probabilidad de supervivencia en función de la variable categórica 'Sex'.
  - b. Gráfico de barras agrupadas: Supervivencia por clase y género. Mostrar la interacción entre clase, sexo y supervivencia.
  - c. Histograma: Distribución de edades. Visualizar la distribución de la edad de los pasajeros y cómo se relaciona con la supervivencia.
- Documentar el desarrollo con Markdown en el notebook, explicando cada análisis y resultado.
- Utilizar el dataset de Titanic (`tarea.cargar_datos()`) También disponible en Kaggle.

### Funciones a implementar (y explicar) en el notebook:

1. Cálculo de probabilidades básicas.
2. Probabilidad condicional (supervivencia por clase o sexo).
3. Cálculo de medidas de tendencia central (media, mediana, moda).
4. Cálculo de medidas de dispersión (desviación estándar, varianza).
5. Análisis de distribución por categorías (sexo, clase, edad).

### Flujo de Trabajo

1. Preparar el entorno (Python, Jupyter Notebook, bibliotecas).
2. Descargar el dataset de Titanic.
3. Implementar las funciones en `tarea.py`.
4. Realizar los pasos listados anteriormente en el notebook `enunciado.ipynb` usando las funciones de `tarea.py`.
5. Ejecutar las pruebas unitarias en `test_tarea.py`.
6. Subir el trabajo a GitHub Classroom, donde el autograding validará la implementación.

### Entregables

- `enunciado.ipynb`: Notebook con desarrollo, explicación y resultados.
- `tarea.py`: Archivo con funciones implementadas.
- Generar un PDF del Notebook (solucionado) e incluirlo (el PDF) en el repositorio.
- Asegurarse de que todas las pruebas unitarias pasen correctamente.
- Subir el código al repositorio (`commit/push`) de GitHub Classroom antes de la fecha límite (indicada en el repositorio, la zona horaria es UTC).

### Criterios de Evaluación

CRITERIO	PONDERACIÓN
Implementación correcta de funciones en <code>tarea.py</code> (Ejecución exitosa de pruebas)	50%
Compleitud y claridad del notebook <code>enunciado.ipynb</code> (Markdown)	30%
Documentación y análisis estadístico (Subir PDF al repositorio)	20%