

Programa de Curso Edición 2025

*Datos No Estructurados y Semiestructurados*

Especialización en Economía, opción Ciencia de Datos

**Docente/s y correo electrónico:** *Guillermo Lezama, guillermo.lezama@cienciassociales.edu.uy*

**Créditos:** 4

**Régimen de cursado:** *Online*

**Carga y distribución de horas estimada:** *20 horas presenciales y 40 horas de trabajo domiciliario.*

**Conocimientos previos recomendados:** Python básico, Jupyter Notebook.

**Contenidos y organización del curso:**

**Descripción:** En un mundo donde la mayoría de los datos ya no se almacenan en bases de datos tradicionales, este curso introduce herramientas y técnicas para procesar datos no estructurados y semiestructurados: texto, imágenes, audio, logs, y datos provenientes de APIs. El curso se enfoca en la conversión de estos datos a formatos estructurados, listos para análisis, visualización o modelado**.**

**Contenidos del Curso**

Módulo 1: Introducción a los Datos No Estructurados y Semiestructurados

* Definiciones, ejemplos, y desafíos principales.
* Formatos comunes: JSON, XML, YAML, HTML.
* Lectura, exploración y estructuración de JSON y XML en Python.
* Introducción a json, xml.etree.ElementTree y pandas.

Módulo 2: Procesamiento de Texto y NLP Básico (4 horas)

* Tokenización, eliminación de stopwords, stemming y lematización.
* Extracción de palabras frecuentes y vectores TF-IDF.
* Introducción a nltk, sklearn.feature\_extraction.text y TextBlob.
* Análisis exploratorio de textos (e.g., reseñas, discursos, publicaciones).

Módulo 3: Scraping Web, APIs y Análisis de Logs (4 horas)

* Introducción al scraping con requests y BeautifulSoup.
* Acceso y estructuración de datos desde APIs públicas (formato JSON).
* Análisis de logs con expresiones regulares y pandas.
* Introducción a re, json, requests y visualizaciones básicas.

Módulo 4: Imágenes y Sonido como Datos (5 horas)

* Carga y transformación de imágenes.
* Extracción de histogramas y bordes.
* Procesamiento de audio.
* Visualización y análisis comparativo de medios.

Módulo 5: Proyecto Final en Formato Hackatón (4 horas)

**Método de enseñanza:** Clases teóricas con aplicaciones prácticas. Ejercicios domiciliarios en cada clase. Trabajo final práctico en grupos.

**Sistema de evaluación:**

* Trabajos de clase (20%).
* Proyecto final (80%).

**Bibliografía:**

No se seguirá un libro específico, pero se utilizarán este libro para algunas referencias de código.

* "Python Data Science Handbook" by Jake VanderPlas.
* “An Introduction to Statistical Learning” by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor.