



EX-2023-06012575- -UBA-DME#SAHDU\_FSOC

## ANEXO I

### Programa de actualización en

### Estrategias computacionales para la investigación social

2024

**Coordinadora: Paula Miguel**

#### Fundamentación

Las ciencias sociales se encuentran ante una expansión sin precedentes de sus fronteras ante los procesos de digitalización de la vida cotidiana. Cada vez más prácticas y relaciones sociales comprenden un comportamiento digital que genera datos y deja huellas. El volumen, la complejidad y la disponibilidad de datos ha crecido como nunca antes. En ese sentido, las tecnologías digitales han tenido y tienen un fuerte impacto cultural, social, epistémico e inauguran áreas de estudio interdisciplinario como las llamadas ciencias sociales computacionales y las humanidades digitales. El mundo digital no solamente ofrece nuevas fuentes y formatos de datos, sino también nuevas formas de vínculo social, que merecen ser abordados con herramientas adecuadas y metodologías pertinentes a la investigación desde la teoría social. Así, crece la investigación que explora los impactos culturales y sociales de las tecnologías emergentes, especialmente en torno a grandes volúmenes de datos y desarrollos relacionados con la inteligencia artificial; metodologías de investigación digital; investigación basada en archivos digitales, entre otras posibilidades. Esos aportes pueden ser aplicados a la toma de decisiones y el desarrollo de nuevos procesos en la gestión pública y productiva, en la elaboración de estudios de mercado y de opinión pública, en la gestión y análisis de la comunicación digital, la sistematización de información y la investigación académica, entre otras. Al mismo tiempo, se requiere de análisis apropiados desde lo técnico que puedan valerse de los problemas y preguntas más clásicas de la teoría social para generar respuestas y soluciones relevantes.

Este programa invita a que las y los estudiantes puedan valerse de los aportes de la ciencia de datos y acercarse a las herramientas analíticas, métodos y estrategias computacionales para trabajar de manera dinámica en la recolección, procesamiento y análisis de datos complejos. Para eso, propone aplicar esas herramientas y técnicas sobre diferentes fuentes, tipos y volúmenes de datos digitales. Esto implica una reflexión, no solamente en cuanto a los aportes de estas metodologías para el análisis de lo social, sino también en relación con su lugar teórico-epistemológico, recuperando



los alcances y límites de las técnicas específicas y evaluando de manera crítica sus fortalezas, sesgos y tensiones éticas.

El objetivo del programa es que los estudiantes adquieran herramientas y estrategias que les permitan profundizar en el análisis de diferentes fuentes y volúmenes de datos complejos. Esto implica incorporar el flujo de trabajo con datos: comprender y plantear un problema, obtener y entender la información, minar, refinar, construir un modelo interpretativo sobre los datos y poder presentar los resultados. Para eso, la modalidad de trabajo prevé el desarrollo de ejercicios y aplicaciones prácticas abordando casos específicos, utilizando bases de datos públicos y/o datos recolectados a tal efecto. Esto permitirá avanzar sobre aspectos epistemológicos y metodológicos, a la vez que el acercamiento a las estrategias computacionales desde la práctica permitirá sopesar su aporte para las ciencias sociales, observando los contextos de uso y aplicación, sus alcances y sus límites.

## **Objetivos**

### *Objetivo general:*

El objetivo del programa es que los participantes adquieran herramientas y estrategias que les permitan profundizar en el análisis de diferentes fuentes y volúmenes de datos complejos y avanzar sobre los aspectos epistemológicos y metodológicos de las estrategias computacionales para las ciencias sociales. Esto implica evaluar de manera crítica el contexto de uso y aplicación, alcances y límites de las técnicas específicas para poder tomar decisiones metodológicas justificadas en función de los intereses teórico-conceptuales.

### *Objetivos específicos:*

- Conocer e incorporar distintas herramientas disponibles y su marco de referencia teórico-conceptual, evaluando sus fortalezas y límites.
- Incorporar fundamentos básicos de programación y manejo de bases de datos para la extracción, análisis y visualización de análisis de datos complejos.
- Familiarizarse con la aplicación e implementación de algoritmos y modelización de datos.
- Integrar los conocimientos específicos en diseños de investigación coherentes, consistentes y factibles.

## **Destinatarios**

Los contenidos de este programa de actualización están orientados especialmente a profesionales de las ciencias sociales que trabajan en sectores que requieren del análisis de datos y la sistematización de información para la administración pública; la elaboración de estudios de mercado y de opinión pública; la gestión productiva; la gestión y análisis de la comunicación digital; la investigación académica, entre otras



posibilidades. También está abierto a profesionales de otras áreas y al público en general interesado en la temática.

Se esperan postulantes que cuenten con título de carreras de al menos 4 años, provenientes de universidades nacionales o extranjeras o formación profesional equivalente y cumplan con los requisitos que determine el programa.

### **Modalidad de cursada, estrategia didáctica y evaluación:**

#### *Modalidad de cursada*

El programa se dicta en modalidad a distancia. Se prevé que la duración de la cursada que abarque un cuatrimestre (32 clases), con una carga horaria total de 128 horas. Cada clase consiste en un encuentro sincrónico de dos horas, que se complementa con dos horas de trabajo asincrónico.

#### *Estrategia didáctica*

En los encuentros sincrónicos se dictarán clases teórico-prácticas, tipo taller, para abordar los contenidos del programa. Se espera una comprometida dedicación y activa participación de los estudiantes que desarrollarán prácticas, ejercicios y análisis sobre los temas propuestos. Se apuntará a que se construyan preguntas y planteos de problemas desde los distintos ejes de contenidos, pensando en su aplicación a casos concretos de investigación. Las actividades indicadas en el programa serán de carácter obligatorio.

#### *Recursos y fuentes*

Los módulos temáticos del programa se complementan con recursos para obtener bases de datos públicos, fuentes documentales, etc. que presenten casos de interés para realizar prácticas y análisis a lo largo de la cursada (entorno Python, Google Colab, Classroom, etc.).

#### *Evaluación*

Se tendrá en cuenta la asistencia a los encuentros sincrónicos y la resolución de las actividades y ejercicios propuestos a lo largo de la cursada de los módulos.

El programa prevé la elaboración de un Trabajo Integrador Final (TIF) que podrá ser diseñado y resuelto en pequeños grupos. Se espera que en este trabajo se incorporen los temas, estrategias metodológicas y analíticas en un caso concreto de investigación e implementación.

Se propone la realización de un último encuentro sincrónico para la presentación de los TIF terminados.



### Seminarios y contenidos mínimos

Módulo	Carga Horaria
I. Introducción a las estrategias computacionales	16
II. Extracción, exploración y divulgación de datos	32
III. Modelos de regresión y clasificación	16
IV. Fundamentos del aprendizaje automático	16
V. Análisis de textos y procesamiento del lenguaje natural (NLP)	24
VI. Implementación y buenas prácticas de desarrollo	24
<b>Carga Horaria Total</b>	<b>128 Hs</b>

#### *Contenidos mínimos*

##### *Módulo I. Introducción a las estrategias computacionales*

Ciencia de datos y computación aplicada a las ciencias sociales. Actualización, debates, aportes, mirada crítica. Las posibilidades, límites, alcances de las estrategias de abordaje. Distintos tipos de datos. Principios de programación. Lógica y estructuración de código. Algoritmos y secuencialidad. Sintaxis de un programa. Buenas prácticas. Cuestiones éticas.

##### *Módulo II. Extracción, exploración y divulgación de datos*

Datos estructurados y no estructurados. Fuentes de información estadística. Acceso, recolección y extracción de datos (APIs, scrapping, pedidos). Estructura de los datos. Manejo de base de datos. Consultas. Preparación de datos numéricos. Estandarización, normalización, trabajo con outliers y datos faltantes. Fundamentos del análisis estadístico descriptivo. Medidas de tendencia central y de dispersión. Básicos de probabilidad. Estimación, probabilidad y distribución. Outputs y visualización de datos (tablas y gráficos básicos). Tipos de información y gráficos. Mapas.

##### *Módulo III. Modelos de regresión y clasificación*

Modelos de regresión lineal y logística, modelos de clasificación y clustering. Regresión lineal, cuadrados mínimos. Regresión logística. Overfitting. Regresión y clasificación. Interpretación y validación. Problemas y errores típicos. Regularización. Entrenamiento y validación (Test set-train set, Cross-validation). Métricas para la medición de performance en regresión y clasificación.



#### *Módulo IV · Fundamentos del aprendizaje automático*

Modelos lineales y de clasificación. Modelos de aprendizaje supervisado y no supervisado. Modelo de árboles de decisiones y clusterización. Árboles de decisión. Ensamblados aleatorios (random forest). Clustering de datos y reducción de la dimensionalidad.

#### *Módulo V. Análisis de textos y procesamiento del lenguaje natural (NLP)*

Extracción de datos textuales (APIs, OCR, scrapping). Preprocesado y limpieza de datos textuales. Expresiones regulares. Categorías gramaticales (POS). Reconocimiento de entidades (NER). Matriz de frecuencias, TF-IDF, nubes de palabras, análisis de emotividad. Clasificación de documentos. Detección de tópicos. Grandes modelos de lenguaje (LLM, GPT).

#### *Módulo VI. Implementación y buenas prácticas de desarrollo*

(Módulo transversal a la cursada. Acompaña la preparación del TIF)

Diseños de investigación integrales. Modularización de código. Python autoejecutable. Entornos: virtual-environment y conda. Buenas prácticas de desarrollo. Variables de entorno. Conceptos básicos de cloud computing. Introducción a APIs. Estandarización, pruebas y depuración de código. Documentación del desarrollo. Presentación de resultados, repositorios (Git). Parámetros de calidad.

#### **Bibliografía general:**

Bengfort, Benjamin; Bilbro, Rebecca y Ojeda, Tony (2018). *Applied Text Analysis with Python*. Sebastopol, CA: O'Reilly Media.

Cady, Field (2017). *The data science handbook*. Hoboken, NJ: John Wiley & Sons.

Dipanjan, Sarkar (2016). *Text Analytics with Python. A Practical Real-World Approach to Gaining Actionable Insights from your Data*. Apress.

Downey, A., Elkner, J., Meyers, C. (2002). *Aprenda a pensar como un programador con Python*. Wellesley, MA: Green Tea Press.

Cramer, D. and Howitt, D. (2004). *The SAGE Dictionary of Statistics. A practical resource for students in the social sciences*. Londres: SAGE.

Foster, I., Ghani, R., Jarmin, R. et al. (eds.) (2017). *Big Data and Social Science. A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press.

Illiinsky, N. and Steele, J. (2011). *Designing Data Visualizations*. Wellesley, MA: Green Tea Press.

Lynch, Scott M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Nueva York, NY: Springer.



Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.

Salganik, M. (2018). *Bit by Bit. Social research in the digital age*. Princeton: Princeton University Press.

vanden Broucke, S. and Baesens, B. (2018). *Practical Web Scraping for Data Science. Best Practices and Examples with Python*. Nueva York, NY: Apress.

VanderPlas, Jake. (2017). *Python data science handbook. Essential tools for working with data*. Sebastopol, CA: O'Reilly.

### **Plantel Docente**

Los contenidos de cada módulo se encuentran a cargo de profesionales de las ciencias sociales con experiencia en programación y en la aplicación de estas estrategias en el campo profesional y académico. También se contará con la participación de especialistas en temáticas específicas.

#### *Zacarías Abuchanab*

Es Licenciado en Sociología (UBA). Maestrando en Metodología de la investigación social (UNTREF). Posgrado en Estadística Matemática - sin finalizar - (UBA). Consultor permanente de arquitectura de datos para la División de Estadísticas de la CEPAL (UN). Profesor de la materia "Fundamentos de Programación para Ciencias Sociales" (Universidad Austral) y del seminario de "Procesamiento de Lenguaje Natural" en la Diplomatura en Ciencias Sociales Computacionales (EIDAES-UNSAM).

#### *Betsabé Cohen*

Es Licenciada en Sociología (UBA) y Especialista en Gestión y Políticas Culturales (UNSAM). Se encuentra finalizando su Maestría en Generación y Análisis de la Información Estadística (UNTREF). Es parte de la comunidad de programadoras R-Ladies Buenos Aires. Ha liderado investigaciones en opinión pública e investigación de mercado. Se especializó en estadísticas de la industria editorial. Actualmente, se desempeña en el equipo de analistas de la Dirección de Planificación y Seguimiento de Gestión del Ministerio de Cultura.

#### *Hernán Escudero*

Es Licenciado en Sociología (UBA), periodista y Machine Learning Engineer (GCP Certified). Se especializó en analítica avanzada para la toma de decisiones y resolución de problemas. Como Lead Data Scientist, ha participado en desarrollos para varias industrias (banca, telecomunicaciones, fintech, cripto, gobierno, I+D). Es fundador y ML Engineer en deployr, una start-up dedicada a soluciones productivas de arquitecturas de datos y modelos de machine learning.



**.UBA40**<sup>∞</sup>  
AÑOS DE  
DEMOCRACIA

**.UBA SOCIALES 35**  
Facultad de Ciencias Sociales

*Andrea Gómez Vargas*

Licenciada en sociología (UBA). Actualmente se desempeña como Analista de Estadísticas Poblacionales en el Instituto Nacional de Estadística y Censos (INDEC), también ejerce la docencia universitaria en ciencia de datos en la UNSAM y UNAB. Es instructora certificada The Carpentries y forma parte de la comunidad de práctica R-Ladies Global, que promueve la participación y diversidad de género(s) en la comunidad de usuarios de R.

*Pedro Orden*

Es Licenciado en Sociología (UBA) y se ha especializado en técnicas de análisis computacional. Realizó un Posgrado en Big Data e Inteligencia Territorial (FLACSO). Entre 2015 y 2023 fue Presidente del Colegio de Sociólogos de la Provincia de Buenos Aires, donde aun coordina el Núcleo de Innovación Social. Tiene experiencia como consultor freelance en generación y análisis masivo de datos sociales para instituciones públicas y privadas. Actualmente, se desempeña como Científico de Datos en el área de Audiencias y Data del multimedio Infobae.