# Introducción a la Ciencia de los Datos

*Gonzalo Castañeda*, CIDE, primavera 2022

## Objetivos del curso:

En este curso, los alumnos conocerán herramientas de frontera para realizar análisis de fenómenos sociales a partir de la visualización de datos, la detección de patrones y el uso de métodos computacionales para realizar análisis predictivos y ejercicios experimentales. Conocerán los vínculos que existen entre la ciencia de datos, los modelos computacionales y el análisis de los sistemas complejos. En un primer paso se familiarizarán con la programación en Python, para luego poder realizar programas de cómputo con los que podrán entender e implementar algoritmos de aprendizaje automatizado, ciencia de redes, y modelos computacionales basados en agentes.

## Contenido:

**Parte I: Introducción a la Ciencia de los Datos**

I.1 Ciencias sociales computacionales

I.2 Ciencia de los datos y complejidad

I.3 Ciencia de los datos en la economía

I.4 Los datos masivos: detección de patrones o teorías sociales

**Parte II: Programación en Python**

II.1 Introducción e instalación del software

II.2 Estructura de datos, funciones y archivos

II.3 Numpy: arreglos y computación vectorizada

II.4 Introducción a la estructura de datos con Pandas

II.5 Manejo de archivos: cargar y guardar datos

**Parte III Procesamiento, Visualización y Análisis de Datos**

III.1 Limpieza y preparación de datos

III.2 Combinación y reordenamiento de datos

III.3 Graficas y visualización de datos

III.4 Agregación de datos y operaciones con grupos

III.5 Manejo de series de tiempo

III.6 Modelación con librerías de Python

**Parte IV Aprendizaje Automatizado**

IV.1 Introducción

IV.2 Aprendizaje con supervisión

IV.3 Aprendizaje sin supervisión

IV.4 Métodos de árboles de decisión

IV.5 Evaluación de modelos

IV.6 Reducción de dimensiones

IV.5 Algunas aplicaciones

**Parte V Redes complejas**

V.1 Introducción

V.2 Elementos básicos de redes

V.3 Coeficientes de aglutinamiento y motivos

V.4 Multigráficas

V.5 Comunidades

V.6 Modelación

**Parte VI: Investigación social en la era digital**

VI.1 Introducción

VI.2 Observar comportamientos

VI.3 Realizar preguntas

VI.4 Correr experimentos

VI.5 Crear colaboraciones en masa

**Parte VII: Modelos basados en agentes**

VII.1 Una visión de la economía desde la complejidad

VII.2 Complejidad y procesos de interacción

VII.3 Autómatas celulares y sociedades virtuales

VII.4 Algunos ejemplos de modelos computacionales

## Método de Evaluación:

El curso será calificado a partir de las siguientes actividades: (i) asistencia al laboratorio y entrega de tareas en los tiempos acordados y con la calidad requerida, 33.3%; (ii) entrega de un reporte gráfico sobre la economía mexicana con el código correspondiente y las bases de datos usadas. 33.3%; (iii) entregar el reporte de un proyecto en el que se aplique algún método de aprendizaje de máquina con el código correspondiente y las bases de datos usadas, 33%.

El reporte en el que se realizan visualizaciones sobre México habría que analizar la evolución del país a lo largo de los últimos 6 sexenios (Salinas de Gortari a la fecha) en algún tema en específico: (i) actividad económica (ingreso, consumo, inversión, sectores, etc.); (ii) variables relacionadas a la política social y pobreza; (iii) indicadores relacionados con el mercado laboral y el trabajo informal. (iv) También se puede realizar un análisis de impacto en distintos indicadores como consecuencia del entorno recesivo provocado por la pandemia y su incipiente recuperación. (v) Una opción más puede ser llevar a cabo un análisis comparado del desempeño de México en 2020-2021 en relación con otros países en términos de indicadores de salud y economía. El trabajo debe priorizar el aspecto visual de las gráficas y la calidad de la información que contiene. En el reporte, cada gráfica debe ir acompañada de las notas que sean necesarias para su comprensión, y un párrafo de interpretación de resultados. Presentar entre 6-10 gráficas, aunque se da mayor peso a calidad que a la cantidad.

En el reporte sobre el proyecto de aprendizaje de máquina, el estudiante tiene la libertad para elegir el tema, la base de datos y el método usado. Además de incluir una breve exposición sobre la naturaleza del tema a abordar y la hipótesis a analizar, el reporte debe presentar de la manera más ilustrativa los resultados obtenidos y su interpretación. La extensión no debe exceder a las10 páginas.

## Bibliografía:


* Athey, Susan. 2018. “The Impact of Machine Learning on Economics”, *The Economics of Artificial Intelligence: An Agenda*. Chicago: Chicago University Press.

* Athey, Susan y GuidoW. Imbens. 2019. “Machine Learning Methods that Economists Should Know About”, *Annual Review of Economics*, 11, pp 685–725.

* Bowles, Michael. 2015. “*Machine Learning in Python. Essential Techniques for Predictive Analysis*”, Indiana USA: John Wiley and Sons.

*Caldarelli, Guido y Alessandro Chessa. 2016. “*Data Science and Complex Networks*”, Oxford UK: Oxford University Press.

* Carbonea, Anna, Meiko Jensen y Aki-Hiro Sato. 2016. “Challenges in Data Science: A Complex Systems Perspective”, *Chaos, Solitons and Fractals*, doi.org/10.1016/j.chaos.2016.04.020

* Castañeda, Gonzalo. 2021. "*The Paradigm of Social Complexity. Volume I: An Alternative Way of Understanding Societies and their Economies*", CDMX: CEEY

* Castañeda, Gonzalo. 2021. "*The Paradigm of Social Complexity. Volume II: Computational Models, Validation, and Applications*", CDMX: CEEY

* Chen, Shu-Heng. 2018. "*Big Data in Computational Social Science and Humanities*", Suiza: Springer.

*Embarak, Ossama. 2018 "*Data Analysis and Visualization Using Python*", New York NY: Apress.

* Gallic, Ewen. 2019. "*Python for Economists*", Manuscrito, Francia: École d'Economie d'Aix-Marseille

* Géron, Aurélien. 2017. "*Hands-On Machine Learning with Scikit-Learn, Keras, and Tensor Flow*", 2nd Edition, California USA: O'Reilly Media Inc.

* Hofman, Jake M., et al. 2021. "Integrating Explanation and Prediction in Computational Social Science", Nature, 595, pp 181-188.

* Johansson, Robert. 2016. "*Introduction to Scientific Computing in Python*", manuscrito

* McKinney, Wes. 2018. "*Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython*", 2a edición, California USA: O'Reilly Media, Inc.,

* Mazzocchi, Fulvio.2015 . "Could Big Data be the End of Theory in Science? A Few Remarks on the Epistemology of Data-driven Science", *Science & Society, EMBO reports*, 16(10).

* Müller Andreas C., y Sarah Guido. 2017. "*Introduction to Machine Learning with Python. A Guide for Data Scientist*", California USA: O'Reilly Media Inc.

*Raschka, Sebastian and Vahid Mirjalili. 2019. "*Python Machine Learning*". 3rd Edition. Birmingham, UK: Packt Publishing.

* Radford, Jason y Kenneth Joseph. 2020. "Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science", *Frontiers in Big Data*. 3 (18).

*Salganick, Metthew J. 2018. "*Bit by Bit. Social Research in the Digital Age*", Princeton NY: Princeton University Press

*Sarkar, Dipanjan, Raghav Bali y Tushar Sharma. 2018." *Practical Machine Learning with Python. A Problem-Solver's Guide to Building Real-World Intelligent Systems*", New York NY: Apress.

* Succi, Sauro y Peter V. Coveney. 2019. "Big Data: the End of the Scientific Method?", *Phil. Trans. R. Soc. A* 377: 20180145

* Severance, Charles R. 2016. "*Python for Everybody. Exploring Data Using Python 3*", manuscrito de libre acceso.

* Vander Plas, Jake. 2016. *"Python Data Science Handbook*". California USA: O'Reilly Media

* Zhang, Jun, Wei Wang, Feng Xia, Yu-Ru Lin, y Hanghang Tong. 2020. "Data-driven Computational Social Science: A Survey", arXiv:2008.12372v1

*Zuhair Al-Taie, Mohammed. 2017. "*Python for Graph and Network Analysis*". Suiza: Springer