Ciencia de Datos Aplicada para la administración pública y Ciencias Sociales

Docente: Guadalupe Gonzalez (Lic. en Ciencia Política)
Fecha del curso: 18/09 al 22/09

Descripción del curso:

Este curso intermedio de ciencias de datos está diseñado para aquellos que tienen experiencia previa en R y desean adquirir nuevas prácticas y metodologías para fortalecer la toma de decisión a partir de grandes conjuntos de datos.

El curso se enfoca en desarrollar un pensamiento computacional, y busca ayudar a los y las estudiantes a adquirir una intuición práctica en la programación y en la solución de problemas de política pública, pero también de investigación con datos. Este curso es intensivo y cubre una amplia variedad de temas, lo que permite a los y las estudiantes profundizar en su conocimiento y comprensión de la ciencia de datos.

A través de una combinación de teoría y práctica, se aprenderá cómo utilizar visualizaciones, web scraping, análisis de texto y machine learning para responder preguntas complejas.

Objetivos:

El objetivo del curso es dotar a los y las estudiantes de autonomía y confianza para usar R. Para lograr esto, se les enseñará a resolver errores y se les explicará la importancia del preprocesamiento y la limpieza de los datos para llevar a cabo un análisis adecuado. Además, se espera que los estudiantes comprendan cómo utilizar las visualizaciones de datos para identificar patrones relevantes. Es importante destacar que R no solo se limita al preprocesamiento y la visualización de datos, sino que ofrece muchas otras herramientas avanzadas que pueden ser útiles en la investigación, como el aprendizaje automático, el análisis de texto, la extracción de datos web, el análisis de redes, entre otros. Al final del curso, se espera que los estudiantes hayan adquirido las habilidades necesarias para utilizar estas herramientas y puedan aplicarlas en sus proyectos de investigación. En este sentido se espera qué al final del curso, los y las estudiantes puedan:

1. Utilizar R de manera autónoma para crear y ejecutar funciones eficientes, y resolver errores en el código.
2. Utilizar RMarkdown para crear trabajos reproducibles y documentados.
3. Aplicar técnicas de limpieza y manipulación de datos para transformar conjuntos de datos y prepararlos para su análisis. Se cubrirán herramientas específicas de R, como dplyr y tidyverse, para ordenar, filtrar, agrupar y unir tablas.

4. Crear visualizaciones efectivas y atractivas que permitan identificar patrones y tendencias en los datos. Los estudiantes aprenderán a utilizar ggplot2 para visualizaciones estáticas y plotly para visualizaciones interactivas.
5. Extraer datos de la web utilizando técnicas de scraping de datos en R para obtener información de tablas de Wikipedia, noticias de portales de medios, discursos de políticos, entre otros.
6. Realizar ejercicios de clasificación utilizando técnicas de aprendizaje automático para identificar patrones y relaciones en los datos. Los estudiantes aprenderán a aplicar algoritmos de clasificación como árboles de decisión y clustering.
7. Analizar grandes conjuntos de datos de texto utilizando técnicas de análisis de texto, como análisis de tópicos y sentimientos, para descubrir patrones y tendencias en los datos de texto.

Además, se espera que los estudiantes apliquen estas habilidades en situaciones del mundo real y que puedan comunicar sus resultados de manera efectiva a un público no técnico.

Metodología:

El curso se impartirá en un formato en línea de 15 horas, distribuidas en 5 días, a través de la plataforma Zoom. Durante cada sesión, la teoría y la aplicación práctica en R se intercalan de manera continua para proporcionar una experiencia de aprendizaje dinámica.

Los y las estudiantes tendrán acceso a todos los materiales utilizados en clase, así como a los scripts en R correspondientes, lo que les permitirá seguir y repasar el contenido del curso en su propio tiempo. Además, se proporcionarán ejercicios para el hogar que no son obligatorios para la aprobación del curso, pero que son una excelente manera de reforzar lo aprendido en clase y ayudar a consolidar la comprensión del material.

Cada día, se introducirá un nuevo tema, siempre relacionado con los conceptos y técnicas aprendidos en el día anterior. La participación activa se fomentará a través de discusiones en grupo, preguntas y respuestas, y ejercicios prácticos.

Evaluación:

La evaluación del curso se realizará a través de un reporte final en el que los y las estudiantes aplicarán las herramientas y técnicas aprendidas durante el curso para resolver un problema a elección de cada estudiante. El proyecto final deberá incluir al menos una de las siguientes técnicas: machine learning, text mining o web scraping, y se presentará en un informe detallado escrito en RMarkdown.

El reporte final se centrará en un problema real de análisis de datos y deberá incluir una descripción detallada del problema a resolver, así como una discusión sobre la metodología utilizada para abordar el problema. Los y las estudiantes deberán mostrar su capacidad para aplicar las técnicas de análisis de datos adecuadas y proporcionar una interpretación significativa de los resultados obtenidos.

Requisitos:

IMPORTANTE: Se va a necesitar que todo/as traigan ya instalados el programa R y R Studio. Eso nos va a ahorrar mucho tiempo al comienzo.
Instalar R es fácil, independiente de si usas Windows, Mac o Linux. Basta con ingresar a https://cran.r-project.org/ y seguir las instrucciones de descarga e instalación.
Para instalar RStudio, es necesario ya haber instalado R. La descarga e instalación es accesible en Windows, Mac y Linux. El link es https://www.rstudio.com/products/rstudio/download/#download

Contenidos y Cronograma

Clase 1 (Limpieza y manipulación de datos):
Qué es RMarkdown. RMarkdown para investigaciones reproducibles en R. Repaso sobre qué es una librería. Introducción al universo Tidyverse. Pipe in R. Limpieza y transformación de datos en R. Primary Key, Foreign Key. Cómo unir dos tablas (joins).

Clase 2 (Funciones y visualización de datos):
Qué es una función. Cómo definir una función. Funciones y estructuras de control (for, while loop).
Que es una visualización. Leyes de Percepción de Gestalt. El impacto de las visualizaciones en política pública. Paquete ggplot2: barplot, histograma, lineplot. Trabajo en capas y customización. Visualizaciones interactivas con plotly.

Clase 3 (Herramientas para el análisis I: Clasificación):
Machine learning. Predicciones en ciencias sociales. Tipos de aprendizaje. Clasificación y regresión. Tradeoff / Compensación sesgo-varianza. Sets de entrenamiento y test. Árboles de decisión. Validación Cruzada. Evaluación de resultados (precisión, recall, accuracy, F1). Curva ROC. Aprendizaje no supervisado. Clustering. Tipos de clustering. Tipos de distancia. K-means.

Clase 4 (Herramientas para el análisis II: Web Scraping):

Introducción a web scraping. Introducción a HTML y CSS (estructura de un sitio web). Qué es Selector Gadget. Rvest package in R. Scrapeando Página 12.

Clase 5 (Herramientas para el análisis III: Text Mining):

Introducción al análisis de texto en R. Trabajando con data no estructurada. Preprocesamiento: stopwords. Corpus, Lexicon, Tokenización, Bag of Words. Matriz término-documento. Stemming. Lematización. Regex. Análisis cuantitativo de texto: Frecuencias, TF-IDF. Análisis de tópicos y sentimientos en Discursos presidenciales.

Día y horario	Tema
Lunes 18/09 de 18 a 21 por la plataforma Zoom	RMarkdown, limpieza y manipulación de base de datos con Tidyverse
Martes 19/09 de 18 a 21 por la plataforma Zoom	Funciones en R y Visualización de datos
Miércoles 20/09 de 18 a 21 por la plataforma Zoom	Introducción a Machine Learning. Clasificación.
Jueves 21/09 de 18 a 21 por la plataforma Zoom	Introducción a Web Scraping
Viernes 22/09 de 18 a 21 por la plataforma Zoom	Introducción a Text Mining (Análisis de sentimiento y tópicos)

Bibliografía

Alvarez. (2016). Computational social science : discovery and prediction (Alvarez, Ed.). Cambridge University Press.

Aydin. (2018). R Web Scraping Quick Start Guide: Techniques and tools to crawl and scrape data from websites. Packt Publishing.

James, Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed. 2021). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Urdinez, F. and Cruz, A. (Eds.) (2021). AnalizaR Datos Políticos. https://arcruz0.github.io/libroadp/index.html

Rhys. (2020). Machine Learning with R, the tidyverse, and mlr (1st edition). Manning.

Vazquez Brust, A. (2021). Ciencia de Datos para Gente Sociable: Una introducción a la exploración, análisis y visualización de datos. https://bitsandbricks.github.io/ciencia_de_datos_gente_sociable/

Bibliografía complementaria:

Hovy. (2022). Text analysis in Python for social scientists : prediction and classification. Cambridge University Press.

Mayaffre, Misuraca, M., & Iezzi, D. F. (2020). Text analytics : advances and challenges (Mayaffre, M. Misuraca, & D. F. Iezzi, Eds.; 1st ed. 2020.). Springer. https://doi.org/10.1007/978-3-030-52680-1

Wilkinson, & Wills, G. (2005). The grammar of graphics (2nd ed.). Springer.