Data Science para políticas públicas

Miércoles 18:00 – 20:45

Pablo Aguirre Hörmann - https://github.com/pjaguirreh

Objetivos: El fin de este curso es hacer de los estudiantes mejores productores y consumidores de conceptos/herramientas relacionadas a lo que conoce como *"data science",* y/o "*machine learning*" en el ámbito de problemas vinculados a las políticas públicas. Para lograr esto, se requiere entender la teoría detrás de estos conceptos: qué son, cómo funcionan, qué los diferencian de las herramientas que normalmente usamos en el contexto de las políticas públicas, y cómo podemos implementarlas. Pero, por otro lado, también se requiere de un entendimiento sobre qué problemas relacionados a las políticas públicas son adecuados para ser abordados por estas herramientas y cuáles no, así como que desafíos se presentan a la hora de implementar estas.

Metodología: Clases expositivas y demostraciones prácticas. Se intercalarán presentaciones sobre conceptos/teoría relacionada al *"data science"* y/o *"machine learning"* con clases sobre técnicas de captura/transformación/manejo de datos. Todos los conceptos serán aplicados a través de ejemplos usando el lenguaje de programación R.

Prerrequisitos: Los estudiantes deberán tener conocimientos previos sobre probabilidad, estadística, y evaluación de programas además de experiencia utilizando el lenguaje de programación R.

Evaluaciones: Se realizarán (3) tareas que involucrarán responder preguntas conceptuales sobre la teoría detrás de los temas tratados en clases, así como ejercicios prácticos utilizando R. Sumado a lo anterior, se realizará un trabajo empírico durante el trimestre el cual estará compuesto por una entrega preliminar, un informe final, y una presentación. El resto de la nota se calculará a partir de la participación en clases de cada estudiante.

- 3 tareas: 30% (10% cada una)
- Trabajo: 60%
    - Informe preliminar: 10%
    - Informe final: 25%
    - Presentación: 25%
- Participación: 10%

Software: Se deberá tener instalado en sus computadores personales, antes del inicio del curso, tanto R (https://cran.r-project.org/) como RStudio (www.rstudio.com).

Bibliografía (disponibles online):

- Gareth James, Daniela Witten, Travor Hastie y Robert Tibshirani (2013). ***An Introduction to Statistical Learning with Applications in R (ISL)***. Disponible en: https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
- Galit Shmueli (2010). ***To Explain or to Predict?***. Disponible en: https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf
- Garret Grolemund y Hadley Wickham (2016). ***R for Data Science (R4DS)***. Disponible en: *https://r4ds.had.co.nz/* (Versión en desarrollo en español disponible en: https://es.r4ds.hadley.nz/)
- Francisco Urdinez y Andrés Cruz Labrín (2019). ***AnalizaR Datos Políticos (ADP)***. Disponible en: https://arcruz0.github.io/libroadp/
- R Development Core Team (2000). ***Introducción a R (IaR)***. Disponible en: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf

Contenidos

Semana	Fecha	Contenidos	Lectura previa	Evaluación
1	05/08	Descripción del curso Introducción al uso de datos para políticas públicas Introducción a la visualización de datos	**ISL**: 2.1 **IaR:** 2 **ADP**: 2.1 y 2.2 **R4DS:** 3	
2	12/08	Visualización de datos	**R4DS:** 3	
3	19/08	Ordenar y transformar datos	**R4DS**: 5 y 12	Tarea 1 Idea de trabajo
4	26/08	Ordenar y transformar datos II	**R4DS**: 5, 12, 13, y 14	
5	02/09	Regresión vs Clasificación Regresión Lineal Regresión logística	**ISL**: 3.1 a 3.3; 4.1 a 4.3 **ADP**: 6.1 a 6.4; 7.1 y 7.2 **R4DS: 23 y 24**	Tarea 2
6	09/09	Predicción vs Inferencia Dilema sesgo-varianza Validación cruzada	**Shmueli, 2010** **ISL**: 2.2; 5.1 y 5.2	
7	23/09	Regularización de modelos lineales Regresión "*stepwise*" Regresión de componentes principales	**ISL**: 6.1 a 6.4 y 10.2	Informe preliminar
8	30/09	Árboles de decisión Clustering	**ISL**: 6.1 a 6.4 ; 10.1 y 10.2 **ADP**: 10	Tarea 3
9	07/10	Web scraping y otros		
10	14/10	Presentación trabajos		Presentación Informe final