

Data Science para políticas públicas

Miércoles 18:00 – 20:45

Pablo Aguirre Hörmann

Objetivos: El fin de este curso es hacer de los estudiantes mejores productores y consumidores de herramientas relacionadas a lo que conoce como “*big data*”, “*data science*”, y “*machine learning*” en el ámbito de problemas vinculados a las políticas públicas. Para esto, se requiere entender los conceptos básicos detrás de estos conceptos: qué son, cómo funcionan, qué los diferencian de las herramientas estadísticas/econométricas que tendemos a usar 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 cuales no así como que desafíos se presentan a la hora de implementar estas.

Metodología: Clases expositivas y demostraciones prácticas. La primera parte del curso se concentrará en enseñar a utilizar el lenguaje y ambiente de programación R mientras que el resto del curso se revisarán distintas herramientas relacionadas a “*big data*”, “*data science*”, y/o “*machine learning*” y cómo implementar estas a través de R.

Pre-requisitos: Los estudiantes deberán tener conocimientos previos sobre estadística y evaluación de programas además de exposición a algún tipo de software/paquete estadístico (**no se necesita de experiencia previa con R**).

Evaluaciones: Se realizarán (5) tareas que involucrarán responder preguntas conceptuales sobre la teoría detrás de los conceptos relacionados al curso así como ejercicios prácticos utilizando R. Sumado a lo anterior, se realizará un trabajo durante el trimestre 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.

- 5 tareas: 50% (10% cada una)
- Trabajo: 45%
 - o Informe preliminar: 10%
 - o Informe final: 20%
 - o Presentación: 15%
- Participación: 5%

Software: Se requiere el uso de computador personal en la sala de clases. Se deberá descargar, antes del inicio del curso, tanto R (<https://cran.r-project.org/>) como RStudio (www.rstudio.com).

Libros (disponibles online):

- Gareth James, Daniela Witten, Trevor 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 Grolmund 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
Parte I: Introducción a la programación / Manejo de datos				
1	30/10	Descripción del curso e introducción al uso de datos para políticas públicas Introducción a R: R y RStudio	Shmueli, 2010 ISL: 2.1 laR: 2 ADP: 2.1 y 2.2	
2	06/11	Introducción a R: Tipo de datos y sintaxis	laR: 2	
3	13/11	Manejo de datos 1	R4DS: 12	Tarea 1
4	20/11	Manejo de datos 2	R4DS: 5	
5	27/11	Visualización de datos	laR: 12 R4DS: 3	Tarea 2
Parte II: Modelos/Machine Learning				
6	04/12	Regresión Lineal y logística	ISL: 3.1 a 3.3; 4.1 a 4.3 ADP: 6.1 a 6.4; 7.1 y 7.2	Tarea 3
7	11/12	Dilema varianza sesgo y Métodos de remuestreo	ISL: 2.2; 5.1 y 5.2	Informe preliminar
8	18/12	Regularización de modelos lineales Regresión “ <i>stepwise</i> ” Análisis de componentes principales	ISL: 6.1 a 6.4 y 10.2 ADP: 10	Tarea 4
9	08/01	Árboles de decision, Bagging, y Random Forest	ISL: 8.1 y 8.2	Tarea 5
10	13/01	Presentación trabajo		Informe Final Presentación