Programa de Estudio

Carrera:	Doctorado en Economía
Asignatura:	Ciencia de datos
Carga Horaria:	36 hs (24 teóricas, 12 prácticas)
Curso:	Miércoles de 18:30-21:30, del 29 de junio al 19 de septiembre
Profesor Titular:	Maximiliano Scocozza Meza
Profesores Ayudantes:	-

1. Fundamentación de la materia

Este curso consta de tres partes. La primera trata de los métodos de obtención y tratamiento de los datos antes de la utilización de los mismos en el análisis. En esta sección se revisan, con base en metodologías estadísticas y a través del uso del lenguaje de programación R, el proceso de "cosecha de datos" a través de robots en internet y la imputación de valores perdidos entre otras metodologías. En la segunda parte se exponen una serie de problemas que pueden resolverse con distintos algoritmos de aprendizaje automático dentro de los cuales se presentarán ejemplos de (1) Clusterización de casos, (2) Desarrollo de pronósticos, (3) Análisis de lenguaje y el discurso, (4) Análisis de redes/grafos. Se explica la diferencia entre la explotación de datos supervisada y no supervisada, así como también la importancia de las metodologías de la validación cruzada. La exploración de grandes cantidades de datos, que en muchos casos son multidimensionales e incluso pueden presentarse de manera no estructurada, implica que su visualización presente una serie de nuevos desafíos. Cada uno de los algoritmos anteriores traen aparejada sus particulares complejidades para comunicar sus resultados de manera integral. La tercera parte de la materia profundiza en distintas metodologías de visualización de datos presentando las librerías de ggplot, ggmap de uso en R y la introducción del programa gephi para el análisis y visualización de redes relacionales.

2. Objetivos

2.1 Generales:

Iniciar a los alumnos en los conceptos y herramientas estadísticas e informáticas básicas que les permita disponer y hacer uso de grandes volúmenes de datos para analizar, interpretar y comunicar patrones en la información.

2.2. Específicos:

Que los alumnos entiendan cómo utilizar las distintas librerías disponibles en el lenguaje R para extraer datos de manera sistemática de internet, ordenar y limpiar grandes bases de datos, aplicar algoritmos de aprendizaje automático y luego visualizar los resultados.

3. Contenidos Mínimos

Captura, estructuración y limpieza de los datos a partir de la librería Tidyverse en el lenguaje R. Aplicación de algoritmos de clusterización, desarrollo de pronósticos, análisis de lenguaje y el discurso, análisis de redes/grafos. Desarrollo de gráficos de manera automática con la librería ggplot. Mapeo de datos georreferenciados con ggmap. Visualización de redes con la herramienta gephy.

4. Unidades de desarrollo de los contenidos:

Unidad	Temas
1 01/07/2020	Qué entendemos por Ciencia de datos? Teoría de la Información. De la Inferencia al análisis cuasi poblacional. Economía y grandes datos. Aplicación práctica de la ciencia de datos.
2 08/07/2020	Captura de datos de manera automática. Acceso a API. Scrapeo de la Web.
3 15/07/2020 22/07/2020	Exploración de la estructura y alistamiento de los datos. Data Wrangling. Tratamiento de datos perdidos. Combinación de bases de datos. Análisis y tratamiento de texto.
4 29/07/2020 05/08/2020	Predicciones con regresión lineal y logística. Modelos de clusterización automática. Árboles de decisión. Random forest. Reducción de dimensiones. Detección de anomalías.
5 12/08/2020 19/08/2020	Aprendizaje automático. Distinción entre modelos supervisados y no supervisados. Trade off entre sesgo y varianza. Decisiones en el diseño del aprendizaje automático. Función de costo. Descenso de grandiente. Validación cruzada. Bootstrap. Overfiting y Regularización.
6 26/08/2020	Teoría de redes. Análisis de redes de poder. Indicadores estructurales. Tipología de roles. Detección automática de grupos. exploración de grafos en gephy.
7 02/09/2020	Visualización de resultados. Criterios de selección de gráficos. Customización y automatización de múltiples resultados y reportes. Visualizaciones interactivas.
8 09/09/2020	Datos georreferenciados. Librería ggmap. Rasterización y Heatmaps, Geocodificación, Visualización de polígonos y vectores georreferenciados.

5. Bibliografía

5.1. Bibliografía Obligatoria

Los textos de referencia para el curso son:

Skiena Steven S , *The Data Science Design Manual,* Springer International Publishing, 2017. https://link.springer.com/book/10.1007%2F978-3-319-55444-0

Nilsson Nils J., *Introduction to Machine Learning*, Robotics Laboratory Department of Computer Science, Stanford University, Stanford, CA 94305, 1998. https://ai.stanford.edu/~nilsson/MLBOOK.pdf

Bishop Christopher M., *Pattern Recognition and Machine Learning*, Microsoft Research Ltd, Cambridge, U.K, 2006) https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Grolemund Garrett, Wickham Hadley, *R for Data Science*, O'Reilly. https://r4ds.had.co.nz/

Blei, D. M., & Lafferty, J. D. (2009). Topic Models. Text Mining: Classification, Clustering, and Applications, 71–89. https://doi.org/10.1145/1143844.1143859

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide, 1–42. https://doi.org/10.1016/j.physrep.2016.09.002

Husson, F., Josse, J., Le, S., & Mazet, J. (2017). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. Retrieved from https://cran.r-project.org/package=FactoMineR

Steinert-Threlkeld, Z. (2017, forthcoming) Twitter as Data. Cambridge University Press.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. ICWSM, 14, 505-514.

6. Modalidad de Enseñanza

En una parte importante del curso los alumnos deberán utilizar la computadora para ejecutar el código en R y así replicar el ejercicio de ejemplo. A lo largo del curso se solicitará a los alumnos que apliquen los conocimientos desarrollados en clase para ir conformando el trabajo final para lo cual se reservará el final de cada clase para avanzar en el diseño de los trabajos y resolver consultas sobre los mismos.

7. Material Didáctico:

Pizarrón, diapositivas de resumen y código en lenguaje R.

8. Modalidad de Evaluación y requisitos de promoción:

La calificación de cada alumno se determinará con la nota obtenida en un trabajo final.