



Taller de Programación¹

PROFESOR

María Noelia Romero

m.n.romero91@gmail.com

CLASES

Mier, 7:00 pm a 10:00 pm

HORAS DE CONSULTA²

Lun, 2:00 pm – 3:00 pm

OBJETIVOS

El objetivo de este curso es presentar las herramientas estadísticas, matemáticas, y computacionales más utilizadas para hacer predicciones y clasificaciones confiables. El curso presenta casos aplicados de cada herramienta en el ámbito investigación, políticas públicas y de negocios. Mediante el entendimiento teórico y práctico, los estudiantes desarrollan un pensamiento crítico de las ventajas y limitaciones de cada herramienta computacional y descubren distintas bases de datos disponibles. El curso desafía a los estudiantes a: (i) programar, manejar distintas bases de datos y crear sus propias bases, (ii) realizar una presentación oral breve de artículos de investigación, (iii) proponer una idea de investigación que utilice alguna de las herramientas computacionales vistas en clase. En resumen, estos objetivos del curso apuntan a que el alumno desarrolle gran versatilidad para comprender, utilizar y presentar datos e ideas según la demanda en su futuro desarrollo profesional (sea académico o no académico).

PERFIL

El curso tiene fuerte carácter técnico, computacional, y de pensamiento crítico. El estudiante está motivado por el uso de datos, ya sea con recopilación de datos primarios o el cuestionamiento de usos de datos secundarios. Además, el curso es ideal para estudiantes con inclinación por sintetizar ideas complejas de manera sencilla para el público en general.

REQUISITOS

Econometría (Lic. en Economía)

HABILIDADES COMPUTACIONALES

El curso se basa en Python, un lenguaje de programación estadístico potente y de amplio uso. No requiere conocimiento previo, pero sí ganas de aprender y experimentar.

¹ El presente programa está desarrollado e inspirado siguiendo el curso 2020 de Big Data por el Profesor Walter Sosa Escudero ([Website aqui](#)) dictado en la Universidad de San Andrés. Se actualizó con la bibliografía del curso *ACE 592 Big Data in Empirical Economics* (Fall 2022) por el Profesor [Peter Christensen](#) (University of Illinois Urbana-Champaign). El programa también con la experiencia de los cursos *E337 Big Data* en Primavera 2023 y *CC408 Ciencias de Datos* Otoño 2024 en la Universidad de San Andrés. Las clases tutoriales en Python fueron actualizadas con el dictado de *Big Data* para la Maestría en Economía en la Universidad Nacional de Tucumán.

² El horario de atención es SÓLO con cita previa. Los instructores tienen horario de oficina en estos días y horas particulares de la semana, pero hay que enviar por correo electrónico a cualquiera de ellos una pregunta (o preguntas) específicas para concertar una reunión con al menos 24 horas de antelación. No se recibirá a ningún estudiante sin cita previa, ni se dará cita sin una consulta explícita. Durante el periodo de exámenes se añadirán algunas horas de oficina adicionales.

MATERIAL

Todo el material del curso se encontrará disponible en el [aula virtual](#). Este programa y temario está sujeto a cambios si es necesario.

DINAMICA EL CURSO

En este curso no hay curva, no hay extensiones, no hay recuperatorio, ni tareas por crédito extra. La política descrita en este programa se aplica por igual a todos los estudiantes.

El curso incentiva la colaboración en grupo de **tres personas** (a determinar en la primer semana de clase). En este curso, vamos a trabajar en [aula virtual](#), para anuncios, post y recordatorios semanales (fechas claves), y coordinar las presentaciones. En dicha plataforma, se espera la activa participación de cada grupo (ver ítem **2.b.** para más detalle).

La aprobación del curso se basa en las siguientes actividades:

- 1. Participación (30% de la nota):** Se esperan dos actividades grupales:
 - a. **Una presentación breves (15 minutos)** de un trabajo de investigación. El grupo debe elegir un paper de aquellos marcados con el símbolo ♦ (rombo) en lista de bibliografía más abajo. Cada grupo debe postear las diapositivas 24 horas antes de la presentación en el Foro academico del aula virtual designado a las presentaciones grupales. Las demoras en subir las slides seran penalizadas en la nota de presentacion (**20% de la nota**). No es necesario ser experto en el artículo, pero si se espera guiar la discusión con el resto de la clase.
 - b. Cada **semana** los grupos deben **postear** en el Foro academico del aula virtual designado, un enlace relevante (nota, discusión, video, conferencia, base de datos, etc.) relacionado con la temática de dicha semana en el curso y no mencionado en este programa. Se espera que el grupo realice un breve comentario en el post sobre la relevancia del enlace propuesto. En la clase tutorial, se discutirá con más detalles esta actividad. Está actividad pondera un **10% de la nota**.
- 2. Trabajos prácticos (30% de la nota):** Los trabajos prácticos usan datos de fuentes secundarias, requieren programación (entregar código de resolución de las consignas) y un reporte que interprete los resultados y discuta las limitaciones. Es requisito entregar y aprobar todos los trabajos prácticos. La entrega de trabajos practicos fuera de la fecha estipulada tiene la penalidad de 50% menos de la nota.
- 3. Propuesta de trabajo (40% de la nota):** puede ser una aplicación o un trabajo de investigación. En la primera parte del curso (ver cronograma), se debe entregar dos páginas con la idea preliminar (o varias alternativas) para la propuesta de investigación (**15% de la nota**). Las consignas de formato y expectativa de la propuesta se pueden encontrar en las Actividades del curso y en clases se discutirá cualquier duda respecto a la entrega. Al finalizar el curso (dos semanas despues de la ultima clase), se debe entregar la propuesta final (**25% de la nota**) con resultados preliminares.

Se espera un lenguaje profesional y/o académico en cada ítem, donde importa el contenido y la visualización de la información.

Asistencia y plagio: Se requiere asistir como mínimo al 75% de las clases. Velaremos por las cuestiones éticas en lo que se refiere a plagio y otras inconductas éticas.



TEMARIO TENTATIVO DEL CURSO (T indica por Teams, P indica presencial, Present. indica presentacion grupal)

	FECHA	TÓPICO DE CLASES	TÓPICO DE TUTORIALES	ENTREGAS GRUPALES
1	25/09 T	Motivación, Objetivos y Dinámica del curso. Introducción: Definiciones de Data mining, big data. Debate del Rol de Big Data. Aprendizaje supervisado y no supervisado. Visión desde Economía Aplicada.	T0 Introducción a Python (videos introductorios) T1 Introducción a Python y Anaconda Navegador T2 Introducción A Pandas y Matplotlib	
2	02/10 T	Métodos No Supervisados 1: Componentes principales y la Maldición de la Dimensionalidad.	T3 Web Scraping, Apis, Github T4 PCA (Python & R)	
3	09/10 T	Métodos No Supervisados 2: Clúster	T5 Clúster	Trabajo Práctico N 0 (Logica de Python) Viernes 11/10
4	16/10 T	Introducción a Métodos Supervisados: Métodos No paramétricos: Histogramas, Kernels y distribuciones	T6 Histogramas & Kernels	
5	23/10 T	Metodos Supervisados 1: Regresión. Modelos lineales. Error Cuadratico Medio.	T7 Regresión Lineal	Trabajo Práctico N 1 (APIs, descriptive & web scrapping) Viernes 25/10

6	30/10 T Present.	Metodos Supervisados 2: Introduccion al problema de Clasificación. Clasificador de Bayes. Regresión logística. Vecinos cercanos	T8 Clasificación. Bayes. Análisis Discriminante. Vecinos Cercanos.	
7	06/11 T Present.	Metodos Supervisados 3: Análisis discriminante. Aplicaciones. Análisis ROC. Comparacion de metodos y regression de Poisson	T9 LDA,QDA. Análisis ROC. Poisson	Trabajo Práctico N 2 (Limpieza, PCA & prediccion) Viernes 08/11

	FECHA	TÓPICO DE CLASES	TÓPICO DE TUTORIALES	ENTREGAS GRUPALES
8	13/11 T Present.	Métodos de Remuestreo. Problemas de Overfitting. Cross-validation. Bootstrap.	T9 K-Fold Cross validation	Trabajo Práctico N 3 (descriptivo & clasificacion) Viernes 15/11
9	20/11 T Present.	Regularización y elección de modelos. Lasso, ridge. Elastic net. Comparación de métodos, Aplicaciones y Mención de la Inferencia Causal	T10 Regularización & Funciones	Ideas borrador propuesta final Viernes 22/11
10	27/11 T Present.	Modelos no lineales: Polinomios, Splines, LOWESS. Modelos Aditivos Generales. Regresión Local por Kernels	T11 Regresiones No Paramétricas	Trabajo Práctico N 4 (regularizacion) Viernes 29/11
11	04/12 T Present.	Metodos de ensamble basados en árboles: CART: Árboles: árboles de regresión y clasificación (CART). Bagging, boosting. Random Forest y Casual Random Forest. Aplicaciones.	T12 Árboles. CART Métodos De Ensemble. Bagging. Random Forest. Boosting.	
12	12/12 P	Discussion de Economía de la Privacidad. Datos censurados y Análisis de Supervivencia.	T13 Survival Analysis	Trabajo Práctico N 5 (CART & Random Forest)

					Viernes 13/12
			Propuesta Final Viernes 20/12		



ESTRUCTURA DEL PROGRAMA Y LECTURAS

Los libros de referencia del curso son:

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer Nature. Descarga [gratis](#)

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires

* Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

La lista de lecturas del curso consta de capítulos de libro de referencia del curso y una serie de artículos académicos y no académicos, todos detallados a continuación. La lista es bastante completa, pero sólo una pequeña proporción de ella, los textos identificados con una “daga” (†) califican como **lecturas obligatorias**. Para las presentaciones grupales, los alumnos pueden elegir de los papers con rombo (♦). Los papers presentados por los grupos también son de **lectura obligatoria** para el examen final. Aquellas con un “asterisco” (*) califican como **lecturas recomendadas** más discutidas en clase.

Tras una breve introducción a la materia, clarificando de terminología comúnmente usada en Ciencia de Datos, debates generales y una introducción a análisis descriptivo y buena visualización de datos; el curso se organiza en cuatro grandes partes. Primero, revisamos con los métodos no supervisados que generalmente son usados en una etapa de análisis exploratorio de los datos. En la segunda parte, iniciamos con los métodos supervisados clásicos de regresión, clasificación, y métodos de remuestreo. En la tercera parte, cubrimos modelos de predicción lineales, métodos de regularización, métodos basados en árboles. En la parte final del curso, hacemos una pausa de métodos para discutir los problemas de la privacidad e introducir un tipo de datos especial (datos censurados y truncados), para luego ver como último método el análisis de supervivencia. En las últimas dos partes del curso (luego de la semana de aumento de complejidad en los conceptos con el siguiente icono:



semana de
aumento de

o. INTRODUCCION A CIENCIA DE DATOS Y VISUALIZACION

Tema 1: Big Data, terminología y debate en Economía. Visualización de datos.

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap. 1 & 2**

† Schwabish, J. A. (2014). An economist's guide to visualizing data. *Journal of Economic Perspectives*, 28(1), 209-234.

*Anderson, C. (2008). The end of theory. *Wired magazine*, 16(7), 16-07.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483- 485.

<http://science.sciencemag.org/content/sci/355/6324/483.full.pdf>

*Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1-24.

- ♦Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4), 171-98.
- ♦Henderson, J. V., A. Storeygard, and D. N. Weil. A Bright Idea for Measuring Economic Growth. *The American Economic Review* 101.3 (2011): 194-199.
- ♦ Lambert, K. J., & Fegley, T. (2023). Economic Calculation in Light of Advances in Big Data and Artificial Intelligence. *Journal of Economic Behavior and Organization*, 206, 243–250. <https://doi.org/10.1016/j.jebo.2022.12.009>
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, 343(6176), 1203-1205.
- Lazer, W. & Kennedy, R.. (2015). What We Can Learn From the Epic Failure of Google Flu Trends, *Wired*, 10.01.15.
- Nickerson, D., & Rogers, T. (2014). “Political Campaigns and Big Data”, *Journal of Economic Perspectives*, vol. 28(2), pp. 51-74.
- Sosa Escudero, W. (2014). Big data: otra vez arroz?, *Diario Clarin*, 6/4/2014.
- Sosa Escudero, W. (2016). Al infinito y más allá: Funes, Borges y big data, *Diario La Nacion*, 12/6/2016.
- Sosa Escudero, W., Anauati, V y Brau, W. (2022), Poverty and inequality studies with machine learning, en Matyas, L. y Chen, F., *Econometrics with Machine Learning*, Springer, New York

I. METODOS NO SUPERVISADOS: ANALISIS DE COMPONENTES PRINCIPALES & CLUSTER

Tema 2: Metodos no supervisados I: Análisis de Componentes Principales (PCA)

- † James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap. 6.3, 12.1, 12.2,**
- * Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics. **Chap 3.5.1, 10.2.3**
- Diamond, R. (2016). The determinants and welfare implications of US Workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3), 479–524.

Tema 3: Metodos no supervisados II: Cluster

- † James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 12.4**
- *Caruso, G., Sosa-Escudero, W., & Svarc, M. (2015). Deprivation and the dimensionality of welfare: a variable-selection cluster-analysis approach. *Review of Income and Wealth*, 61(4), 702-722.
- Levy-Yeyati, E. & Struzenegger F. (2023) Exchange Rate Regimes 20 years later: The prevalence of floats. *RedNIE Working Paper Series N182*
- Lopez, Juan Cruz *Caracterización socioeconómica de clusters electorales* (Tesis grado de Lic. En economía, 2024)

Tema 4: Introduccion a Método Supervisado: Histogramas y distribuciones de kernels

* Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics. **Chap. 6.2**

Chen, Yen-Chi (2024) "STAT 425: Introduction to Nonparametric Statistics"

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3), 1405-1454.

Jales, H. (2018). Estimating the effects of the minimum wage in a developing country: A density discontinuity design approach. *Journal of Applied Econometrics*, 33(1), 29-51.

John DiNardo, Nicole M. Fortin and Thomas Lemieux, 1996, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, Volume 64, Number 5 (September), pp. 1001- 1044.

II. METODOS SUPERVISADOS I: REGRESION, CLASIFICACION & TECNICAS DE REMUESTREO

Tema 5: Modelo de Regresion lineal

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 3**

♦ Anselin, L., & Williams, S. (2015). Digital neighborhoods. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 1-24.

♦ Cavallo, A. (2017) "Are Online and Offline Prices Similar? Evidence from Multi-Channel Retailers" *American Economic Review*. 107 (1). 283-303.

Tema 6: Calsificacion I: Introduccion a Clasificacion, Logit & Vecinos Cercanos (KNN)

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 4.1 a 4.3**

♦ Brinatti, A., Cavallo, A., Cravino, J., & Drenik, A. (2021). The international price of remote work (No. w29437). *National Bureau of Economic Research*.

Horowitz, Joel, L., and N. E. Savin. 2001. "Binary Response Models: Logits, Probits and Semiparametrics." *Journal of Economic Perspectives*, 15 (4): 43–56.

♦ Mougnot, B., Amaya, E., Mezones-Holguin, E., Rodriguez-Morales, A. J., & Cabieses, B. (2021). Immigration, perceived discrimination and mental health: evidence from Venezuelan population living in Peru. *Globalization and health*, 17, 1-9.

Tema 7: Calsificacion II: Analisis discriminante. Analisis ROC. Comparacion de metodos y regression de Poisson

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 4.4 a 4.6**

♦ Mullally, C., Rivas, M., & McArthur, T. (2021). Using Machine Learning to Estimate the Heterogeneous Effects of Livestock Transfers. *American Journal of Agricultural Economics*

*Baylé, Federico (2016) “Detección de villas y asentamientos informales en el partido de La Matanza mediante teledetección y sistemas de información geográfica” Tesis de Maestría.

Askatas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107-120.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073-1076.

Tema 8: Metodos de Remuestreo: Cross-validation & Bootstrap

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 5**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires. **Cap 5**

* Hastie, T., Tibshirani, T. Y Freedman, J. (2013) *The Elements of Statistical Learning*. **Chap 7.2, 7.10**

*Sosa Escudero, W., & Gasparini, L. (2000). A note on the statistical significance of changes in inequality. *Económica*, 46.

III. METODOS SUPERVISADOS II: REGULARIZACION, NO LINEALES & BASADOS EN ARBOLES



Tema 9: Regularizacion: LASSO & RIDGE. Elastic Net, comparaciones & discusion de causalidad usando LASSO

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 6**

Belloni, V. Chernozhukov, C. Hansen: “High-Dimensional Methods and Inference on Structural and Treatment Effects,” *Journal of Economic Perspectives*, 28 (2), Spring 2014, 29-50.

*Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *The American Economic Review*, 105(5), 481-485.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India (No. w24678). National Bureau of Economic Research.

*Danton S. Char, Michael D. Abràmoff & Chris Feudtner (2020) Identifying Ethical Considerations for Machine Learning Healthcare Applications, *The American Journal of Bioethics*, 20:11, 7-17,

*Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491-495.

♦Lusk, J. L. (2017). Consumer research with big data: applications from the food demand survey (FoodS). *American Journal of Agricultural Economics*, 99(2), 303-320.

Zou, H. y Hastie, T., 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society*, 67, 2, 301-320.

† Wüthrich, K., & Zhu, Y. (2023). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics*, 105(4), 982-997.

Tema 10: Modelos no lineales de reducción de la dimensionalidad y no linealidad

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. 7.1 a 7.6

* Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics. **Chap. 3.6.2, 5.1, 5.2**

Tema 11: Métodos de ensamble basados en Árboles: Árboles (CART). Boosting, Bagging, Random Forest

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. **Chap 8**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires **Cap 3, pag 85 a 94.**

Breiman, L. (2003). Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1), 81-82.

*Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

♦ Keely, L. C., & Tan, C. M. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, 92(5), 944-961.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). Party: A laboratory for recursive partytioning.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2), 3-27.

* Wager, S. and Athey, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523), pp.1228- 1242.

IV. MÉTODOS SUPERVISADOS III: PRIVACIDAD, TIPO DE DATOS & ANÁLISIS DE SUPERVIVENCIA



Tema 12: Discusión de los problemas de la privacidad, disponibilidad de datos & Análisis de supervivencia

† Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of economic Literature*, 54(2), 442-492.

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Chap.11.1 a 11.5

* Bauer, Sebastian and Hnilo, Florencia, Scars of the Gestapo: Remembrance and Privacy Concerns (June 05, 2024). Available at SSRN: <https://ssrn.com/abstract=4328676> or <http://dx.doi.org/10.2139/ssrn.4328676>

Borgschulte, M., Cho, H., & Lubotsky, D. (2022). Partisanship and survey refusal. *Journal of Economic Behavior & Organization*, 200, 332-357.

Gilchrist, D.S. & Sands, E. G. (2016). "Something to Talk About: Social Spillovers in Movie Consumption", *Journal of Political Economy*. vol. 24(105), pp. 1339-1382.

† Kleemans, M., & Thornton, R. L. (2021). Who Belongs? The Determinants of Selective Membership into the National Bureau of Economic Research. *AEA Papers and Proceedings*, 111, 117-122.

◆ Kleemans, M., & Thornton, R. (2023). Fully Promoted: The Distribution and Determinants of Full Professorship in the Economics Profession. *AEA Papers and Proceedings*. 113: 467-472

◆ Leak, A. & Lansley, G. (2018). “Geotemporal Twitter Demographics”, *Consumer Data Research*, capítulo 11, UCL Press.

* Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press. **Chap. 16.1 & 17.1**