Universidad de SanAndrés

CC408. Ciencia de Datos[1]

Profesor

María Noelia Romero

mromero@udesa.edu.ar

Clases MAGISTRALES

Jueves, 15:40 a 18:50

H011 (Ed. Hirsch)

Horas de consulta[2]

Martes, 14:30 a 15:30

Tutor

Ignacio Spiousas

ispiousas@udesa.edu.ar

Clases Tutoriales

Martes 19:00 a 20:30

B109 (Ed. Juan y Florence Ball)

Martes 17:20 a 18:50

B109 (E. Juan y Florence Ball)

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

Programación y Estadística II (Lic. en Ciencias del Comportamiento)

HABILIDADES COMPUTACIONALES

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

Material

[1] El presente programa está desarrollado e inspirado siguiendo el curso 2020 de Big Data por el Profesor Walter Sosa Escudero (Website aqui) y 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). Ademas, el modo de evaluacion del curso esta está desarrollado e inspirado siguiendo el curso 2023 de *World Economic History* por el Profesor Tommy E. Murphy (syllabus original en su sitio web).

[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.

Todo el material obligatorio del curso se encontrará disponible en el campus virtual (http://campusvirtual.udesa.edu.ar/). Este programa y temario **está sujeto a cambios si es fuera necesario**.

EVALUACION EL CURSO

La aprobación del curso se basa en las siguientes evaluaciones grupales e individuales (que se explican en detalle más abajo):

- Una serie de **trabajos grupales** (N_{grupal}, 45% de la nota del curso)
- Una serie de **cuestionarios sorpresa** individuales (N_{quiz}, 15% de la nota del curso)
- Un **examen final** individual (N_{examen}, 40% de la nota del curso)

Tenga en cuenta que, aunque la evaluación se divide en varios componentes, **no** habrá **examen parcial**. La nota de este curso ($Nota_{final}$) será entonces:

$$Nota_{final} = 0.45N_{grupal} + 0.15N_{quiz} + 0.40N_{examen}$$

El redondeo se aplicará únicamente para calcular la nota final de la asignatura (0 a 10), no cualquier nota intermedia.

Para aprobar este curso es **necesario** obtener mas de **cuatro** (**4**) puntos en:

- En el examen final ()
- En la nota final del curso ()

Notar que **ambas** condiciones se deben cumplir para pasar el curso

Para los estudiantes que desaprueban el examen final ($N_{examen} < 4$), existe la posibilidad de un recuperatorio. El resultado de las dimensiones de cuestionarios sorpresas y trabajos grupales, crean una nota umbral ($Nota_{umbral}$) cuyo objetivo es evaluar el trabajo individual y grupal en el curso durante el semestre:

$$Nota_{umbral} = 1/2 \times N_{quiz}, \; + 1/2 \times N_{grupal}$$

Esto es un promedio no ponderado de notas, por lo que, elementos con pequeña ponderacion en la nota del curso (como los quizzes) en realidad son muy importantes para la nota umbral. Estudiantes con la **nota umbral** mayor o igual a seis ($Nota_{umbral} \geq 6$) podran acceder al **examen recuperatorio** en caso de desaprobar el examen final (ver mas detalles al en la seccion de **Recuperatorios**).

Trabajos Grupales (N_{grupal})

El curso incentiva la colaboración en grupo de **tres** personas (a determinar la primera semana de clase al tutor).[3] Todos los participantes del grupo tienen que pertenecer al mismo tutorial. En este curso, vamos a trabajar en el campus virtual, para anuncios, post y recordatorios semanales (fechas claves de entregas grupales), y coordinar las presentaciones. En dicha plataforma, se espera la activa participación de cada grupo

[3] Sólo se admitirá un numero *limitado* de grupos con **dos** personas dependiendo del numero de alumnos.

La nota de trabajo grupal consta (45% de la nota del curso) consiste en las siguientes partes e instancias:

- Cuatro **trabajos prácticos** (N_{tp}, 30% de la nota grupal);
- **Presentacion grupal** en clase ($N_{presentacion}$, 20% de la nota grupal);
- **Posteos semanales** a partir de la tercera semana de clases (N_{post}, 15% de la nota grupal)
- Un **borrador** de la propuesta de trabajo final en semana de parciales ($N_{borrador}$, 10% de la nota grupal);
- Una propuesta de **trabajo final** en la ultima semana de finales ($N_{trabajo}$, 25% de la nota grupal);

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.

Una **presentación breve** debe ser breve de **15 minutos** sobre un trabajo de investigación con aprobación de los profesores. El grupo debe elegir un trabajo 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 un foro designado a las presentaciones grupales en el campus virtual. Cada hora de demora en no cumplir con esta instrucción resta un punto a la nota de la presentacion. No es necesario ser experto en el artículo, pero si se espera guiar la discusión con el resto de la clase.

Cada semana los grupos deben **postear en el foro de campus virtual** 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 la bibliografia de este curso. Se espera que el grupo realice un breve comentario en el post sobre la relevancia del enlace propuesto (ver en el campus virtual el template a seguir). En la clase tutorial, se discutirá con más detalles esta actividad.

La **propuesta de trabajo final** puede ser una aplicación o un trabajo de investigación. En la primera semana de parciales (ver cronograma), se debe entregar un borrador con de **3 (tres) páginas** con una o mas ideas preliminares para la propuesta de investigación. La **entrega final de la propuesta** sera en la ultima semana de parciales. Las consignas de formato y expectativa de la propuesta en cada instancia se pueden encontrar en el campus virtual y en las clases tutoriales y horas de consultas se discutirá cualquier duda respecto a la entrega.

En resumen, la nota grupal (N_{grupal}) se calcula:

$$N_{grupal} = 0.3N_{tp} + 0.2N_{presentacion} + 0.15N_{post} + 0.10N_{borrador} + 0.25\,N_{trabajo}$$

En todas las instancias, se espera un lenguaje profesional y/o académico en cada ítem, donde importa el contenido y la visualización de la información.

Cuestionarios sorpresas (Quiz, N_{quiz})

Los cuestionarios sorpresa (Pop-quiz) son pequeños examenes presenciales, y consistirán en 6 preguntas de opción múltiple o verdadero/falso sobre la **lectura obligatoria del día** (ver cronograma en la ultima pagina) y el **tema de la clase anterior**. Para cualquiera que haya realizado las lecturas y haya asistido a la clase anterior, responder las preguntas debería ser prácticamente trivial y una manera fácil de obtener puntos para la nota final. Serán un total de **6 (seis)** custionarios sorpresa y se realizarán ***sin previo*** aviso durante las clases magistrales, utilizando el Campus Virtual en tu dispositivo preferido (móvil, tableta, portátil), en cualquier momento del curso (excepto las dos primeras clases magistrales), al inicio de la clase, y tendrá una duración aproximada de 4 minutos.

Una vez indicado en el aula que el cuestionario sorpresa se realizará ese día, los estudiantes deberán:

1. Abrir la pestaña *Ciencia de Datos* en el Campus Virtual de su dispositivo.
2. Vaya a la sección Cuestionarios sorpresa.
3. Haga clic en el enlace "Pop Quiz X" (siendo X el número del cuestionario), donde aparece un cartel que dirá:

> *Intentos permitidos: 1*
>
> *Este cuestionario está abierto en [día], [mes] de 2024, [hora de inicio] Este cuestionario se cerrará el [día], [mes] de 2024, [hora de inicio + 5 min.]*
>
> *Límite de tiempo: 4 minutos*

4. Haga clic en "Previsualizar el cuestionario ahora" y luego en "Comenzar intento".
5. Responda las preguntas recordando que:

- ~~Debes leer las preguntas atentamente.~~
- Debes responder el cuestionario en 4 minutos.
- Las 6 preguntas del cuestionario valen lo mismo (=> se sugiere gastar la misma cantidad de tiempo en cada una)
- A menos que se indique lo contrario, sólo UNA respuesta es correcta.

6. Cuando termine recuerde asegurarse apretar el botón de "Enviar todo y terminar".

Si el estudiante no se conecta en el momento de realizar la prueba, **no podrá hacerlo en ningún otro momento y recibirá una calificación de 0 (cero)**. Si por **cualquier motivo** se omite una prueba, la puntuación será 0 (cero), **sin excepciones**. No habrá oportunidad de rehacer un quiz perdido. La nota de cada prueba estará entre 0 y 10, de la siguiente manera:

Preguntas correctas	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Nota del quiz	**0**	**1**	**2**	**4**	**6**	**8**	**10**

Su *'componente de calificación del cuestionario'* (N_{quiz}) se calculará tomando el promedio de los ***4 (cuatro)*** *mejores puntajes*. Esto es, se descartan los dos quizzes de peor desempeño dentro de los 6 que se tomen en el semestre.

Examen Final

El examen final será escrito, virtual y normalmente tendrá tres secciones:

- **Parte 1: una serie de verdadero o falso** con una **justificación** breve (enfocado en la parte teorica y conceptual de las clases magistrales)
- **Parte 2: Control de Lectura.** Se evalua el contro de lectura de dos trabajos academicos presentados por los grupos.
- **Parte 3: Aplicación en el mundo real**. Se presenta un caso de una consultora hipotética donde debe aplicar las herramientas vistas y discutidas en clase justificando las ventajas y desventajas para el caso.

La fecha del examen depende Alumnos, y se anunciara en tanto en clase magistral como en el campus virtual.

Recuperatorio

En caso de desaprobar el curso, existe la posibilidad de recuperatorio solo para los estudiantes que hayan demostrado un trabajo consistente a lo largo del curso, con una nota umbral igual o mayor a seis ($Nota_{umbral} \geq 6$). Este examen recuperatorio tendra las mismas partes del examen final. En esta instancia de recuperacion, la nota final del curso sera la nota del examen recuperatorio con un tope maximo de seis (6).

Notas del recuperatorio	**4,00 – 6,49**	**6,50 – 8,99**	**9,00 - 10**
Nota del curso	**4**	**5**	**6**

Cualquier nota menor a 4 (cuatro, por ejemplo, 3,99999) desaprueba el curso.

Asistencia: como es práctica de UdeSA, se requiere asistir como mínimo al 75% de las clases teóricas y tutoriales. Se toma asistencia el dia del cuestionario sorpresa.

PLAGIO Y DESHONESTIDAD INTELECTUAL

La Universidad de San Andrés exige un estricto apego a los cánones de honestidad intelectual. La existencia de plagio constituye un grave deshonor, impropio de la vida universitaria. Su configuración no sólo se produce con la existencia de copia literal en los exámenes presenciales, sino toda vez que se advierta un aprovechamiento abusivo del esfuerzo intelectual ajeno. El Código de Ética de la Universidad considera conducta punible la apropiación de la labor intelectual ajena, por lo que se recomienda apegarse a los formatos académicos generalmente aceptados (MLA, APA, Chicago, etc.) para las citas y referencias bibliografías (incluyendo los formatos online). La presunta violación a estas normas puede dar lugar a la conformación de un Tribunal de Ética que, en función de la gra- vedad de la falta, podrá recomendar sanciones disciplinarias que van desde el apercibimiento a la expulsión. En caso de duda consulte la guía que se encuentra disponible en el Centro de Escritura Universitaria.

Universidad de SanAndrés

TEMARIO GENERAL DEL CURSO, TUTORIALES Y ENTREGAS GRUPALES

Semana	*TÓPICO DE CLASES*	*TÓPICO DE TUTORIALES*	*ENTREGAS GRUPALES*
1	Motivación, Objetivos y Dinámica del curso. Introducción: Definiciones de Data mining, big data. Debate del Rol de Big Data. Aprendizaje supervisado y no supervisado. ***Analsis descriptivo I***: Visualizacion de datos.	Introducción a Python (videos introductorios)	
2	***Analisis descriptivo II***: Histogramas, Kernels y distribuciones. Aplicaciones.	Introducción a Python	
3	***Métodos No Supervisados I***: Componentes principales y la Maldición de la Dimensionalidad	Web Scraping, Apis, Github	
4	***Métodos No Supervisados II***: Clúster	Introducción A Pandas y Matplotlib	
5	Regresión. Modelos lineales, linealizables y no lineales.	Introducción A Numpy y Regresión Lineal	Trabajo Práctico N 1 Domingo 8 de Septiembre
6	***Clasificación I:*** Clasificador de Bayes. Regresión logística. Vecinos cercanos.	Componentes principales	
7	***Clasificación II***: Análisis discriminante. Aplicaciones. Análisis ROC. Comparacion de metodos y regression de Poisson	Cluster y Kernels	
8	Métodos de Remuestreo. Problemas de Overfiting. Cross-validation. Bootstrap.	Regresiones No Lineales y semi paramétricas	Trabajo Práctico N 2 Domingo 29 de Septiembre
	Semana Parciales		Borrador de ideas de propue Lunes 7 de Octubre

Semana	TÓPICO DE CLASES	TÓPICO DE TUTORIALES	ENTREGAS GRUPALES
9	Regularización y elección de modelos: Lasso, ridge. Elastic net. Comparación de métodos. Regularizacion e Inferencia Causal. Aplicaciones	Clasificación. Bayes. Análisis Discriminante. Vecinos Cercanos. Análisis ROC	
10	Modelos lineales que reducen la dimensionalidad: Regresion de Componentes principales y Regresion parcial local (PLS). Modelos no lineales: Polinomios, Splines, Regresión Local por KNN & Kernels	K-Fold Cross validation	Trabajo Práctico N 3 Domingo 27 de Octubre
11	Metodos de Ensamble I: Árboles: árboles de regresión y clasificación. Bagging, boosting.	Regularización	
12	Metodos de Ensamble II: Random Forest. Casual Random Forest. Aplicaciones.	Árboles. CART	
13	Introducción a Economía de la Privacidad & Datos censurados Análisis de Supervivencia I: Funcion de supervivencia	Funciones	
14	Analisis de Supervivencia II: Funciones de riesgo y Metodo de proporcional de Cox	Métodos De Ensemble. Bagging. Random Forest. Boosting. Survival Analysis	Trabajo Práctico N 4 Domingo 24 de Noviembre
15	Repaso General y Practica de Ejericio Aplicado (Parte 3 del examen final)		
	Semana de Finales: Fecha de Examen a definir		Propuesta Final Sabado 7 de Diciembre

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 academicos, 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 tambien son de **lectura obligatoria** para el examen final. Aquellas con un "asterísco" (*) califican como **lecturas recomendadas** más discutidas en clase.

Tras una breve introducción a la materia, clarificando de terminologia comunmente usada en Ciencia de Datos, debates generales y una introduccion a analisis descriptivo y buena visualizacion de datos; el curso se organiza en cuatro grandes partes. Primero, revisamos con los metodos no supervisados que generalmente son usados en una etapa de analisis exploratorio de los datos. En la segunda parte, iniciamos con los metodos supervisados clasicos de regresion, clasificacion, y metodos de remuestreos. En el tercer parte, cubrimos modelos de prediccion lineales, metodos de regularizacion, metodos basados en arboles. En la parte final del curso, hacemos una pausa de metodos para discutir los problemas de la privacidad e introducir un tipo de datos especial (datos censurados y truncados), para luego ver el último metodo el analisis de supervivencia. En las ultimas dos partes del curso (luego de la semana de parciales), el *nivel de dificultad* empieza a aumentar. En dichas secciones, indicamos el  aumento de complejidad en los conceptos con el siguiente icono:

0. Introduccion A CIENCIA DE DATOS Y VISUALIZACION

Tema 1: Ciencia de Datos, terminologia y debate de Big Data. Analisis descriptivo I: Visualizacion 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**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires **Cap 1, pag 23 a 33**

*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.

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.

♦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.

Nickerson, D., & Rogers, T. (2014). "Political Campaigns and Big Data", *Journal of Economic Perspectives*, vol. 28(2), pp. 51-74.

♦Ohayon, M. M., & Milesi, C. (2016). Artificial outdoor nighttime lights associate with altered sleep behavior in the American general population. *Sleep*, 39(6), 1311-1320.

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

Tema 2: Analsis descriptivo II: 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. 6.1 (**pag. 190 a 194**)**

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.

I. Metodos no Supervisados: ANALISIS DE COMPONENTES PRINCIPALES & CLÚSTER

Tema 3: 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 4: Metodos no supervisados II: Clúster

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* **Chap 12.4**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires **Cap 3, pag 69 a 76**

*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.

*Gerlach, M., Farb, B., Revelle, W., & Nunes Amaral, L. A. (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature human behaviour*, 2(10), 735-742.

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 clústers electorales* (Tesis grado de Lic. En economia, 2024)

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**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires **Cap 3, pag 76 a 85**

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

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

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

♦Matz, S. C. (2021). Personal echo chambers: Openness-to-experience is linked to higher levels of psychological interest diversity in large-scale behavioral data. *Journal of Personality and Social Psychology*, 121(6), 1284.

Tema 6: Calsificacion I: Introduccion, 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**

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

Mougenot, 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. Análisis 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.

Askitas, 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. Pag 121 a 130**

* 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**

*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.

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.

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

Tema 10: Modelos no lineales de reduccion 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: Metodos basados en Arboles I: Arboles. Boosting & Bagging

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* **Chap 8.1, 8.2.1, 8.2.3**

† 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.

Tema 12: Metodos basados en Arboles II: Random Forest & Aplicaciones

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* **Chap 8.2.2., 8.2.5**

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. Metodos Supervisados III: Privacidad, tipo de Datos & Analisis de supervivencia

Tema 13: Discusion de los problemas de la privacidad y disponibilidad de datos

* 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 & 11.2**

† Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires **Cap 6, pag 139 a 148**

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

* 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.

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

* Walsh, A. E., Naughton, G., Sharpe, T., Zajkowska, Z., Malys, M., van Heerden, A., & Mondelli, V. (2024). A collaborative realist review of remote measurement technologies for depression in young people. Nature Human Behaviour, 8(3), 480-492.

Tema 14: Analisis de supervivencia, funcion, riesgo proporcional & estimacion de Cox

† James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* Chap.**11.3** a **11.5**

† 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

Cronograma de lecturas obligatorias

		Jueves
7-ago	1	**Intro.Big Data & Analsis descriptivo I: Visualizacion de Datos** † Schwabish, J. A. (2014)
15-ago	2	**Analisis descriptivo II: Histogramas, Kernels**
22-ago	3	**Metodos No Supervisados I: PCA (Quizz no sorpresa)** † James, et al (2023) Chap 6.3
29-ago	4	**Metodos No Supervisados II: Cluster** † Sosa Escudero, W. (2021) Cap 3, pag 69 a 76
5-sep	5	**Modelo Lineal. Regresion** † Sosa Escudero, W. (2021) Cap 3, pag 76 a 85
12-sep	6	**Clasificacion I: Bayes, Logit. KNN neighbor** † James, et al (2023) Chap 4.1 a 4.3
19-sep	7	**Clasificacion II: LDA QDA Aplicaciones** † James, et al (2023) Chap 4.4 a 4.6
26-sep	8	**Remuestreo. Overfitting. CV. Bootstrapping** † Sosa Escudero, W. (2021) Cap 5. Pag 121 a 130
3-oct		RECESO PARCIALES
10-oct		RECESO PARCIALES
17-oct	9	**Regularizacion. LASSO. Ridge. Elastic Net.PCR. Comparacion. Aplicaciones**

		† James, et al (2023) Chap 6.2
24-oct	10	**Modelos no lineales: stepwise function, PLS, splines, etc** † James, et al (2023) Chap 7.4
31-oct	11	**Ensamble I: CART. Bagging, Boosting** † Sosa Escudero, W. (2021) Cap 3, pag 85 a 94
7-nov	12	**Ensamble II: Random Forest. Aplicaciones** † James, et al (2023) Chap 8.2.2., 8.2.5
14-nov	13	**Debate sobre la Privacidad y Tipo de datos: censurados/truncados. Introduccion a Analisis de supervivencia I: funcion de supervivencia** † Sosa Escudero, W. (2021) Cap 6, pag 139 a 148
21-nov	14	**Analisis de supervivencia II: Riesgo proporcional & Metodo de Cox** † James, et al (2023) Chap 11.3 a 11.5
28-nov	15	**Repaso General & Practica de Examen**