E337. Big Data y Aprendizaje Automático: Herramientas e Ideas para Economistas[[1]](#footnote-1)

|  |  |
| --- | --- |
| Profesor María Noelia Romero  [mromero@udesa.edu.ar](mailto:mromero@udesa.edu.ar) Clases MAGISTRALES Lun, 5:20 pm – 6:50 pm  Mie, 7:00 pm – 8:30 pm  Aula M007  (Ed. Otero Monsegur) Horas de consulta[[2]](#footnote-2) Martes 2:30 pm – 3:30 pm Tutor Victoria Oubiña  [v.oubina@gmail.com](mailto:v.oubina@gmail.com) Clases Tutoriales *Online*  Mie, 5:20 pm - 6:50pm | 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).[[3]](#footnote-3) 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. |

# Material

Todo el material del curso se encontrará disponible en el campus 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 descripta en este programa se aplica por igual a todos los estudiantes**.

El curso incentiva la colaboración en grupo de tres personas (a determinar el primer día de clase). En este curso, vamos a trabajar en el campus 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 (10% de la nota):** Se esperan dos actividades grupales:
   1. Una **presentación breve (15 minutos)** de un trabajo de investigación con aprobación de los profesores. 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 un foro designado a las presentaciones grupales en el campus virtual. No es necesario ser experto en el artículo, pero si se espera guiar la discusión con el resto de la clase.
   2. 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 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.
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.
3. **Propuesta de trabajo (20% de la nota):** puede ser una aplicación o un trabajo de investigación. En la primera semana de parciales (ver cronograma), se debe entregar una página con la idea preliminar para la propuesta de investigación. Las consignas de formato y expectativa de la propuesta se pueden encontrar en el campus virtual y en las clases tutoriales se discutirá cualquier duda respecto a la entrega.
4. **Examen final (40% de la nota):** evaluación integral e individual de todo el contenido del curso, incluyendo lecturas y habilidades computacionales. Importante: es **condición** **necesaria** **aprobar** el examen final.

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

*Asistencia y plagio*: como es práctica de UdeSA, se requiere asistir como mínimo al 75% de las clases teóricas y tutoriales. Velaremos por las cuestiones éticas en lo que se refiere a plagio y otras inconductas éticas.

# Temario tentativo del curso

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FECHA | | TÓPICO DE CLASES | TÓPICO DE TUTORIALES | ENTREGAS GRUPALES |
| 1 | | 26/2 | 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. | Introducción a Python (videos introductorios) |  |
| 2 | | 28/2  4/3 | *Métodos No Supervisados 1*: Componentes principales y la Maldición de la Dimensionalidad | 6/3. Introducción a Python |  |
| 3 | | 6/3 11/3 | *Métodos No Supervisados 2*: Clúster | 13/3. Web Scraping, Apis, Github |  |
| 4 | | 13/3  18/3 | *Introducción a Métodos Supervisados*. *Métodos Semiparamétricos*: Histogramas, Kernels y otros | 20/3. Introducción A Pandas y Matplotlib |  |
| 5 | | 20/3  25/3 | Regresión. Modelos lineales, linealizables y no lineales. | 27/3. Introducción A Numpy y Regresión Lineal | Trabajo Práctico N 1  Domingo 24 de marzo |
| 6 | | 27/3  1/4 | Clasificación 1: Clasificador de Bayes. Regresión logística. Vecinos cercanos. | 3/4. Componentes principales |  |
| 7 | | 3/4**,** 8/4  10/4 | Clasificación 2: Análisis discriminante. Aplicaciones. Análisis ROC | 10/4. Cluster y Kernels |  |
| 8 | | 15 /4  17/4 | Métodos de Remuestreo. Problemas de Overfiting. Cross-validation. Bootstrap. | 17/4. Regresiones No Lineales y semi paramétricas | Trabajo Práctico N 2  Domingo 21 de abril |
| 9 | | 22/4  24/4 | Regularización y elección de modelos. Lasso, ridge. Elastic net | 8/5. Clasificación. Bayes. Análisis Discriminante. Vecinos Cercanos. Análisis ROC |  |
|  | |  | *Semana Parciales: 22/4 a 3/5* |  | Propuesta Preliminar  Miércoles 24 de abril |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | FECHA | TÓPICO DE CLASES | TÓPICO DE TUTORIALES | ENTREGAS GRUPALES |
| 10 | 6/5  8/5 | Regularización y elección de modelos II: Comparación de métodos, Aplicaciones y Mención de la Inferencia Causal | 15/5. K-Fold Cross validation |  |
| 11 | 13/5  15/5 | Modelos no lineales: Polinomios, Splines, LOWESS. Modelos Aditivos Generales. Regresión Local por Kernels | 22/5. Regularización | Trabajo Práctico N 3  Domingo 26 de mayo |
| 12 | 20/5  22/5 | CART: Árboles: árboles de regresión y clasificación. Bagging, boosting. | 29/5. Funciones |  |
| 13 | 27/5  29/5 | Random Forest y Casual Random Forest. Aplicaciones. | 5/6. Árboles. CART |  |
| 14 | 3/6**,** 5/6  10/6 | Datos censurados y Análisis de Supervivencia. Introducción a Economía de la Privacidad. | 12/6. Métodos De Ensemble. Bagging. Random Forest. Boosting. |  |
| 15 | 12/6  19/6 | Herramienta Estadística para Economía Aplicada: Multiple Test Hypothesis | 19/6. Survival Analysis | Trabajo Práctico N 4  Viernes 21 de junio |
|  |  | Semana de Exámenes Finales: 24/6 al 5/7  Exámen Final: Miércoles 26 de junio |  | Propuesta Final  Lunes 1 de julio |

**†** 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.

**†** 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

**†** 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.

**†** Sosa Escudero, W. (2017). Big data y aprendizaje automatico: Ideas y desafios para economistas, mimeo.

1. El presente programa está desarrollado e inspirado siguiendo el curso 2020 de Big Data por el Profesor Walter Sosa Escudero ([Website aqui](https://bigdataudesa.weebly.com/)) y se actualizó con la bibliografía del curso *ACE 592 Big Data in Empirical Economics* (Fall 2022) por el Profesor [Peter Christensen](https://www.uiuc-bdeep.org/christensen) (University of Illinois Urbana-Champaign) [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
3. Para el alumno de la Licenciatura que luego planea hacer la Maestría en Economía en nuestro Departamento de Economía, el objetivo particular es que en el presente curso tenga un conocimiento general de las herramientas de Big Data y Aprendizaje Automático, entendiendo las aplicaciones en Python. Luego, en el curso de Big Data dictado por el Profesor Walter Sosa Escudero, se esperará desarrolle una idea de investigación a lo largo del cursado, aprendiendo las propiedades teóricas formales y el estado de la literatura en Machine Learning. Por lo que, sería posible, con permiso del Departamento, hacer la secuencia de Big Data nivel Licenciatura y Maestría, con una lógica similar a cursar Econometría y Econometría Avanzada, respectivamente. [↑](#footnote-ref-3)