Unidad: Introducción a la Ciencia de Datos

Ciencia de Datos para Economía y Negocios

Nicolás Sidicaro
Marzo 2025

Nicolás Sidicaro

- Investigador en Fundar
- Docente FCE-UBA y UADE
- Licenciado en Economía
- Data Scientist
- Econométria y minería de datos

Contacto

- Email: nsidicaro.fce@gmail.com
- LinkedIn: linkedin.com/in/nicolassidicaro

- Hacia un dispositivo público de monitoreo del subte de Buenos Aires
  - Monitor de fallas del subte mediante API de Twitter
- Distribución de trabajadores en el AMBA
  - Georreferenciación
- Mapa productivo laboral de Argentina
  - Georreferenciación y visualización
- El precio de la ropa en Argentina
  - Relevamiento de precios

1. Ciencia de Datos: Introducción

¿Qué es la Ciencia de Datos?

- Campo interdisciplinario que utiliza métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento e información de datos estructurados y no estructurados
- Combina aspectos de:
  - Estadística
  - Matemáticas
  - Programación
  - Visualización
  - Conocimiento del dominio
- Se enfoca en descubrir patrones, tendencias y relaciones para generar información accionable

1. Ciencia de Datos: Introducción

¿Para qué se puede usar?

- Toma de decisiones basada en datos
- Optimización de procesos y recursos
- Detección de anomalías y fraudes
- Personalización de productos y servicios
- Predicción de tendencias y comportamientos
- Automatización de procesos

Ecosistema de Ciencia de Datos

Data Mining (Minería de Datos)

- Proceso de descubrir patrones y conocimientos interesantes a partir de grandes volúmenes de datos
- Se enfoca en la **extracción** de información no evidente
- Utiliza algoritmos para identificar relaciones, anomalías y tendencias

Data Analysis (Análisis de Datos)

- Proceso de inspección, limpieza, transformación y modelado de datos
- Se enfoca en el **examen** de datos para responder preguntas específicas
- Más orientado a la comprensión descriptiva y diagnóstica

Arquitectura de Datos

- Diseño de estructuras para recopilar, almacenar, procesar y consumir datos
- Definición de flujos de datos, bases de datos y sistemas de procesamiento
- Garantiza que los datos estén disponibles, seguros y sean de calidad

Ecosistema de Ciencia de Datos

Machine Learning (Aprendizaje Automático)

- Subcampo de la inteligencia artificial que permite a los sistemas aprender de datos
- Crea modelos para reconocer patrones y tomar decisiones con mínima intervención humana
- **Tipos principales**:
  - Supervisado: aprende de datos etiquetados
  - No supervisado: encuentra patrones en datos no etiquetados

Inteligencia Artificial (IA)

- Campo más amplio que busca crear sistemas que puedan percibir, razonar y actuar
- Machine Learning es un subconjunto de la IA
- Incluye procesamiento de lenguaje natural, visión por computadora, robótica, etc.
- Enfocada en crear soluciones que emulan aspectos de la inteligencia humana

Comparación de conceptos clave

Concepto	Definición	Enfoque	Aplicación en economía	Uso en negocios
Ciencia de datos	Campo interdisciplinario que extrae conocimiento y valor de los datos	Proceso completo desde preguntas hasta decisiones	Análisis integral de problemas económicos complejos	Transformación digital, innovación basada en datos, optimización de procesos empresariales
Análisis de datos	Examen de datos para sacar conclusiones sobre la información	Enfoque principalmente descriptivo y diagnóstico	Interpretación de indicadores económicos y tendencias	Informes de desempeño, dashboards de KPIs, análisis de competencia

Comparación de conceptos clave

Concepto	Definición	Enfoque	Aplicación en economía	Uso en negocios
Data Mining	Descubrimiento de patrones en grandes conjuntos de datos	Exploratorio, busca relaciones no evidentes	Segmentación de mercados, patrones de consumo	Análisis de canasta de mercado, sistemas de recomendación, segmentación de clientes
Machine Learning	Algoritmos que mejoran automáticamente con la experiencia	Predictivo y prescriptivo	Predicción de variables económicas, detección de anomalías	Predicción de demanda, detección de fraude, mantenimiento predictivo

Comparación de conceptos clave

Concepto	Definición	Enfoque	Aplicación en economía	Uso en negocios
Inteligencia Artificial	Sistemas que emulan comportamiento inteligente	Resolución autónoma de problemas complejos	Automatización de decisiones económicas complejas	Asistentes virtuales, automatización de servicio al cliente, optimización logística

Etapas del proceso de Ciencia de Datos

1. **Definición del problema**: Identificar objetivos y preguntas clave
2. **Recolección de datos**: Obtener información relevante de diversas fuentes
3. **Limpieza y preparación**: Transformar datos crudos en formato utilizable
4. **Exploración y análisis**: Identificar patrones y relaciones
5. **Modelado**: Crear modelos predictivos o descriptivos
6. **Evaluación e interpretación**: Validar resultados y extraer conclusiones
7. **Implementación y comunicación**: Aplicar hallazgos y presentarlos efectivamente

El 80% del tiempo de trabajo suele estar en las primeras 3 etapas. La parte **fancy** es marginal.

2. Herramientas fundamentales: GitHub

¿Qué es GitHub?

- Plataforma basada en Git para control de versiones y colaboración
- Permite a múltiples personas trabajar en los mismos archivos sin conflictos
- Funciona como un "repositorio" central para código y documentación

Funcionalidades principales

- **Repositorios**: Almacenamiento de proyectos con historial completo
- **Branches (Ramas)**: Versiones paralelas para desarrollo simultáneo
- **Pull Requests**: Mecanismo para revisar y aprobar cambios
- **Issues**: Seguimiento de tareas, errores y funcionalidades
- **Actions**: Automatización de flujos de trabajo

2. Herramientas fundamentales

¿Qué es Google Colaboratory (Colab)?

- Entorno de notebook basado en Jupyter en la nube
- Permite escribir y ejecutar código Python directamente en el navegador
- No requiere instalación ni configuración local

Características principales

- **Integración con Google Drive**: Almacenamiento y acceso a datos
- **GPU/TPU gratuitas**: Aceleración para modelos complejos
- **Entorno preconfigurado**: Bibliotecas populares ya instaladas
- **Interfaz interactiva**: Combina código, texto narrativo y visualizaciones
- **Fácil compartición**: Colaboración en tiempo real

2. Herramientas fundamentales: R

¿Qué es R?

- Lenguaje de programación especializado en computación estadística y gráficos
- Software libre y de código abierto
- Creado por Ross Ihaka y Robert Gentleman en 1993
- Ampliamente utilizado en investigación estadística, ciencia de datos y machine learning

Características principales

- **Orientado al análisis estadístico**: Diseñado específicamente para esta tarea
- **Extensible**: Más de 18,000 paquetes adicionales en CRAN
- **Capacidades gráficas avanzadas**: Excelente para visualización de datos
- **Comunidad activa**: Amplio soporte y recursos disponibles
- **Reproducibilidad**: Facilita documentar y compartir análisis completos

Paquetes esenciales que utilizaremos

tidyverse: **dplyr**; **ggplot2**; **tidyr**; **plotly**, **data.table**, **lubridate**, **caret**; **rvest**,**RSelenium**,**haven**

2. Herramientas fundamentales: RStudio

¿Qué es RStudio?

- Entorno de desarrollo integrado (IDE) para R
- Interfaz gráfica que facilita el uso de R
- Disponible en versión de escritorio y servidor
- Desarrollado por Posit (anteriormente RStudio, Inc.)

Componentes principales

- **Editor de código**: Con resaltado de sintaxis y autocompletado
- **Consola**: Para ejecutar comandos de R
- **Entorno y variables**: Visualización de datos y objetos en memoria
- **Historial**: Registro de comandos ejecutados
- **Gráficos**: Visualización de resultados
- **Ayuda y documentación**: Acceso rápido a información sobre funciones

2. Herramientas fundamentales: Python

¿Qué es Python?

- Lenguaje de programación de alto nivel, interpretado y de propósito general
- Diseño centrado en la legibilidad del código
- Multiparadigma: soporta programación orientada a objetos, imperativa y funcional
- Uno de los lenguajes más populares para ciencia de datos y machine learning

Características principales para ciencia de datos

- **Sintaxis clara y legible**: Facilita el aprendizaje y mantenimiento
- **Ecosistema científico robusto**: NumPy, pandas, SciPy, scikit-learn, etc.
- **Visualización potente**: Matplotlib, Seaborn, Plotly
- **Versatilidad**: Se integra fácilmente con otros lenguajes y sistemas
- **Machine learning y deep learning**: Bibliotecas como scikit-learn, TensorFlow, PyTorch

Aplicaciones en nuestro curso

- Implementación de algoritmos de machine learning

2. Herramientas fundamentales:

¿Qué es Anaconda?

- Distribución de Python y R para computación científica
- Incluye más de 1500 paquetes pre-instalados
- Gestión sencilla de entornos y paquetes con Conda
- Simplifica enormemente la configuración del entorno de trabajo

Componentes principales

- **Navigator**: Interfaz gráfica para lanzar aplicaciones y gestionar paquetes
- **Conda**: Gestor de paquetes y entornos virtuales
- **Jupyter Notebook/Lab**: Entorno interactivo basado en web
- **Spyder**: IDE científico específico para Python

2. Herramientas fundamentales:

Ventajas de Spyder para ciencia de datos

- Diseñado específicamente para análisis científico
- Explorador de variables similar a RStudio
- Editor de código con IPython integrado
- Depurador avanzado
- Panel de ayuda y documentación
- Perfiles de ejecución para análisis de rendimiento

2. Herramientas fundamentales: SQL

¿Qué es SQL?

- Lenguaje de consulta estructurado (Structured Query Language)
- Estándar para gestionar y consultar bases de datos relacionales
- Pilar fundamental para el trabajo con datos estructurados
- Elemento indispensable en el kit de herramientas de un científico de datos

Componentes principales de SQL

DDL (Data Definition Language): CREATE, ALTER, DROP; **DML (Data Manipulation Language)**: SELECT, INSERT, UPDATE, DELETE; **DCL (Data Control Language)**: GRANT, REVOKE; **TCL (Transaction Control Language)**: COMMIT, ROLLBACK

Aplicaciones en ciencia de datos

- Extracción y filtrado de datos
- Agregación y resumen de información
- Combinación de múltiples fuentes de datos (JOINS)
- Preparación de datos para análisis posterior
- Creación de vistas y reportes

3. Estructura del curso

Modalidad de dictado

- **Clases presenciales** (martes): enfoque conceptual y teórico
- **Clases virtuales** (viernes): enfoque práctico
  - Algunas sincrónicas, otras asincrónicas según cronograma
- Complemento con lecturas de bibliografía y ejercicios prácticos

Unidades temáticas

1. Introducción a la estadística para Ciencia de Datos
2. Trabajo con bases de datos
3. Modelado de datos
4. Web Scraping
5. Visualización de datos
6. Machine Learning
7. Tópicos de ciencia de datos

3. Estructura del curso: Evaluación

Sistema de evaluación

- **Evaluación parcial (30%)**
  - Trabajo domiciliario
  - Se realiza durante el horario de una clase virtual (±1 hora)
  - Entrega obligatoria en la fecha establecida
  - Posibilidad de correcciones menores
- **Trabajo Práctico Integrador (70%)**
  - Informe final sobre temas planteados un mes antes de la entrega
  - Integración de conceptos y herramientas del curso
  - Análisis de datos reales aplicados a economía y negocios
- **Nota final**: Ponderación de ambas evaluaciones
- **Sin examen final**

3. Estructura del curso: Materiales

Herramientas principales

- **R**: Principal lenguaje de programación del curso
- **SQL**: Para trabajo con bases de datos
- **Python**: Para machine learning

4. Recursos

- **Material de trabajo**:
  - Slides
  - Scripts compartidos en GitHub (desde semana próxima)
  - Datasets de práctica en Github
  - Recursos online (tutoriales, documentación)
  - Bibliografía complementaria

4. Aplicaciones en Economía y Negocios

Economía

- Análisis de tendencias macroeconómicas
- Predicción de indicadores económicos
- Evaluación de impacto de políticas públicas
- Estudios de comportamiento del consumidor
- Análisis de mercados laborales (con EPH)
- Comercio internacional y análisis de exportaciones

Negocios

- Segmentación de clientes
- Análisis predictivo de ventas
- Optimización de precios
- Detección de fraude
- Análisis de sentimiento en redes sociales
- Optimización de cadenas de suministro
- Sistemas de recomendación

4. Aplicaciones en Economía y Negocios

Casos de estudio que abordaremos

- **Análisis de datos de la EPH** (Encuesta Permanente de Hogares)
- **Análisis de datos de comercio exterior**
- **Web scraping de precios**
- **Segmentación de clientes por comportamiento de compra**
- **Modelos predictivos para variables económicas**

Limitaciones y consideraciones éticas

- Calidad y representatividad de los datos
- Interpretación causal vs. correlacional
- Privacidad y protección de datos personales
- Sesgos algorítmicos y su impacto en la toma de decisiones
- Transparencia y explicabilidad de modelos

Desafíos y oportunidades

Desafíos actuales

- Volumen y velocidad creciente de datos
- Integración de fuentes heterogéneas
- Necesidad de infraestructura adecuada
- Escasez de profesionales calificados
- Interpretación y comunicación efectiva de resultados
- Consideraciones éticas y regulatorias

Oportunidades profesionales

- **Roles**: Analista de datos, Científico de datos, Ingeniero de datos, Especialista en visualización
- **Sectores**: Finanzas, Consultoría, Sector público, E-commerce, Salud, Educación
- **Competencias valoradas**: Programación, Estadística, Comunicación, Conocimiento del dominio

Próxima clase: Introducción a la estadística para Ciencia de Datos

Contacto: nsidicaro.fce@gmail.com