Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

Temario del Curso

1. Fundamentos y Preparación de los Datos
2. Aprendizaje Supervisado
3. Aprendizaje No Supervisado
4. Introducción al Aprendizaje Automático Moderno

Unidad 1: Fundamentos y Preparación de los Datos

- Introducción a la Ciencia de Datos
- Flujo de un Proyecto de Ciencia de Datos
- Tipos de Datos y Estructuras Comunes
- Preprocesamiento de Datos
- Visualización Exploratoria de Datos
- Ética y Sesgos en el Análisis
- Herramientas Computacionales
- Actividad Práctica

Unidad 2: Aprendizaje Supervisado

- Clasificación óptima y regla de Bayes
- Regresión lineal y logística
- Discriminantes lineales: LDA, QDA, Fisher
- Vecinos más cercanos (`k-NN`)
- Árboles de decisión y bosques aleatorios
- Métricas de evaluación

Unidad 3: Aprendizaje No Supervisado

- Agrupamiento: `k`-medias, jerárquico, espectral, EM y mezclas gaussianas
- Reducción de dimensionalidad: PCA, SVD, NMF
- Proyecciones aleatorias y visualización

Unidad 4: Aprendizaje Automático “Moderno”

- Redes neuronales básicas
  - Perceptrón y redes multicapa
  - Funciones de activación
  - Concepto de aproximación universal
- Ejemplos y casos de uso actuales

Evaluación del Curso

- **30%** Tareas semanales
- **70%** Proyecto final
  - Entrega de código comentado
  - Informe escrito
  - Presentación oral

Ayudante del Curso

Gilberto Flores Vargas

- gilberto.flores@cimat.mx
- Oficina: D706
- **Horas de oficina:**
  - Miércoles: 12:30–13:30
  - Viernes: 11:00–12:00

Sobre este curso

Introducción a la Ciencia de Datos es un curso de tercer semestre de la **Maestría en Probabilidad y Estadística** del CIMAT.

Este curso se posiciona como una base sólida para el análisis moderno de datos desde una perspectiva estadística rigurosa, combinando teoría y práctica computacional.

Objetivos del curso:

- Entender las principales problemáticas de los datos modernos y su análisis.
- Dominar técnicas estadísticas de agrupamiento y clasificación de datos masivos.
- Entender y aplicar técnicas de reducción de dimensionalidad y regularización.
- Adquirir herramientas fundamentales para realizar análisis de datos en distintos contextos aplicados.

Unidad 1

Fundamentos y Preparación de los Datos

Unidad 1

1. Introducción a la Ciencia de Datos
2. Flujo de un Proyecto de Ciencia de Datos
3. Tipos de Datos y Estructuras Comunes
4. Preprocesamiento de Datos
5. Visualización Exploratoria de Datos
6. Ética y Sesgos en el Análisis
7. Herramientas Computacionales
8. Actividad Práctica

¿Qué NO es Ciencia de Datos?

- No es sólo **programar en Python/R**.
- No es únicamente **usar algoritmos de machine learning**.
- No se reduce a hacer **gráficas bonitas**.
- *Tampoco es simplemente* **estadística con un nombre moderno**.

Idea clave: Ciencia de datos es una disciplina integradora que requiere estadística, computación y conocimiento del dominio.

Ciencia de Datos vs. Estadística Clásica

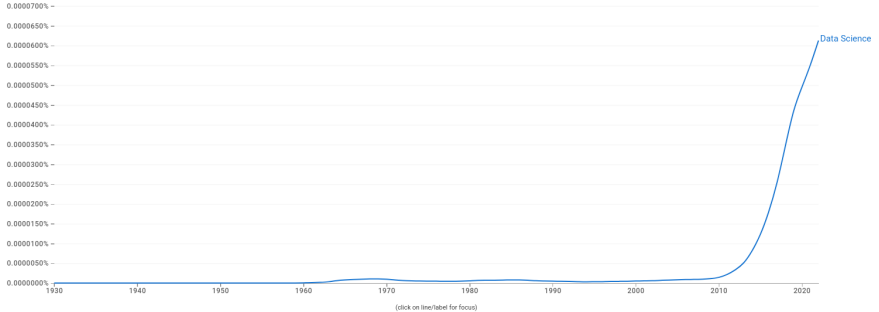
- **Estadística clásica:** centrada en inferencia, pruebas de hipótesis y modelos matemáticos.
- **Ciencia de datos:** aplicación práctica de técnicas estadísticas para resolver problemas concretos usando datos como medio.
- No se limita a la teoría: incluye recolección, procesamiento, modelado y comunicación de resultados.
- Integra **estadística, computación y contexto del problema** en un flujo orientado a objetivos.
- Énfasis en reproducibilidad, automatización y utilidad en la toma de decisiones.

Ambas disciplinas son complementarias: la estadística provee rigor; la ciencia de datos extiende la aplicación a contextos imperfectos.

¿Qué es Ciencia de Datos?

- Definición operativa: disciplina que combina **estadística, computación y conocimiento del dominio** para extraer información útil de los datos.
- Involucra el ciclo completo:
  - Recolección y gestión de datos
  - Análisis exploratorio
  - Modelado predictivo e inferencial
  - Evaluación y comunicación de resultados
  - Despliegue y retroalimentación
- Meta: transformar datos en conocimiento para apoyar decisiones.

Evolución del término “Data Science”



- Observamos el crecimiento del uso del término “Data Science” en libros impresos desde 1930 hasta 2022.
- La curva muestra un uso marginal hasta entrados los años 2000, con una aceleración clara en la última década.

`https://books.google.com/ngrams/graph?content=Data+Science&year_start=1930&year_end=2022&corpus=en&smoothing=3`

Primeros usos del término

- El término *Data Science* aparece en la literatura en los **años 1960s**, pero sin un significado definido.
- En **1974**, Peter Naur lo utiliza en su libro *Concise Survey of Computer Methods*, planteándolo como un campo alternativo a la informática tradicional.
- A finales de los 90s, se empieza a consolidar como un área distinta, con artículos que la definen como una disciplina emergente.

Fuente: Naur 1974

Desarrollo como disciplina

- En los **años 1990s–2000s**, varias revistas y conferencias comienzan a adoptar "Data Science" como título o tema central.
- Se resalta su rol como un **puente entre estadística y computación**, enfocado en la extracción de conocimiento útil de datos.
- El énfasis está en la **aplicación práctica y orientada a problemas**, no solo en el desarrollo teórico.

Fuente: Kelleher and Tierney 2018

¿Por qué un flujo de trabajo en Ciencia de Datos?

- La ciencia de datos es una **disciplina aplicada**: no basta con conocer las técnicas, es necesario integrarlas en un **proceso coherente**.
- Un flujo de trabajo permite:
  - Partir de un problema o pregunta inicial.
  - Seguir pasos organizados hacia una solución.
  - Asegurar reproducibilidad y comunicación clara.
- Sirve como guía para no perder de vista el **objetivo final**: transformar datos en conocimiento útil.

Un flujo, no una receta

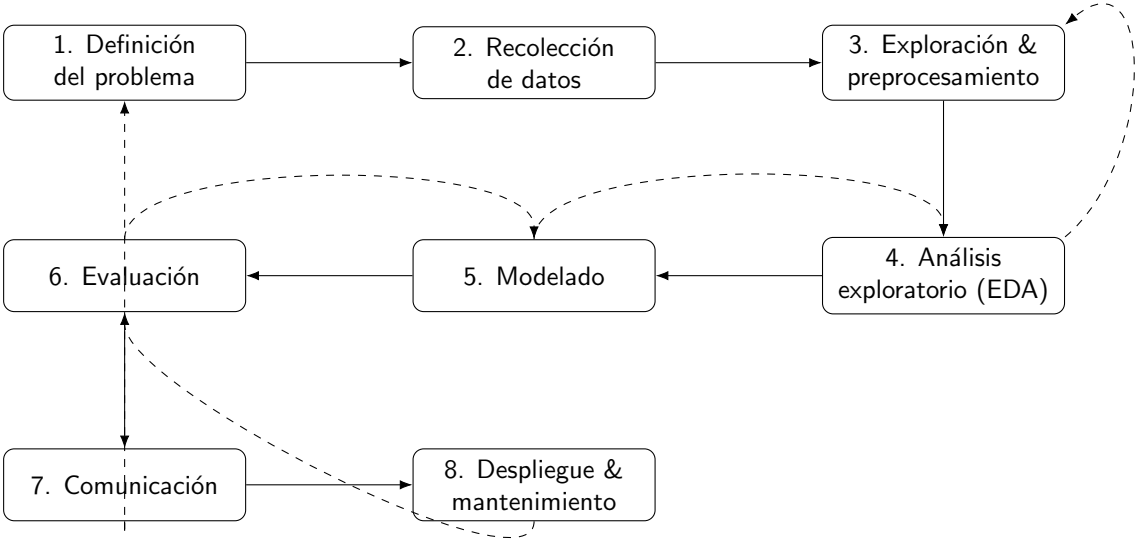
- Existen múltiples formas de organizar un proyecto de ciencia de datos.
- No todos los proyectos requieren exactamente los mismos pasos ni el mismo orden.
- Sin embargo, contar con un **esqueleto de referencia** ayuda a:
  - Estructurar el trabajo.
  - Evitar omitir fases críticas (limpieza, validación, comunicación).
  - Adaptar y modificar el proceso según las características del problema.

Propuesta de Flujo de un Proyecto de Ciencia de Datos

1. **Definición del problema:** plantear la pregunta analítica y los objetivos.
2. **Recolección de datos:** fuentes internas o externas.
3. **Exploración y preprocesamiento:** limpieza, codificación, imputación, normalización.
4. **Análisis exploratorio:** estadísticas descriptivas y visualizaciones.
5. **Modelado:** selección, entrenamiento y ajuste de modelos.
6. **Evaluación:** métricas, validación cruzada, comparación de enfoques.
7. **Comunicación:** resultados claros para públicos técnicos y no técnicos.
8. **Despliegue y mantenimiento:** implementación, documentación, monitoreo.

Nota: este es un esquema general; cada proyecto puede adaptarlo a sus necesidades.

Propuesta de Flujo de un Proyecto (Esquema iterativo)



De los Proyectos a los Datos

En cualquier proyecto de Ciencia de Datos,
el insumo fundamental son **los datos**.

De los Proyectos a los Datos

En cualquier proyecto de Ciencia de Datos,
el insumo fundamental son **los datos**.

Antes de analizarlos, debemos entender:

- ¿Qué tipo de datos tenemos?
- ¿Cómo se representan?
- ¿Qué operaciones son válidas sobre ellos?

De los Proyectos a los Datos

En cualquier proyecto de Ciencia de Datos,
el insumo fundamental son **los datos**.

Antes de analizarlos, debemos entender:

- ¿Qué tipo de datos tenemos?
- ¿Cómo se representan?
- ¿Qué operaciones son válidas sobre ellos?

Próximo paso: Tipos de datos y escalas de medición.

Tipos de Datos y Representación

- Los datos pueden clasificarse según:
  - **Escala de medición** (nivel de información que aportan).
  - **Estructura** (qué tan organizados están).
- Comprender estas categorías es fundamental para:
  - Selección de métodos de análisis.
  - Elección de visualizaciones adecuadas.
  - Definición de preprocesamiento.

Escalas de Medición

- **Nominal:** categorías sin orden. Ejemplo: colores, géneros musicales.
- **Ordinal:** categorías con orden, pero sin magnitud definida. Ejemplo: nivel educativo (primaria, secundaria, universidad).
- **Intervalar:** diferencias son significativas, pero no existe cero absoluto. Ejemplo: temperatura en °C.
- **Razón:** poseen cero absoluto y permiten razones. Ejemplo: peso, altura, ingresos.

Ejemplo de Datos Nominales

ID	Color favorito
1	Azul
2	Verde
3	Rojo
4	Amarillo

Table: Datos nominales: categorías sin orden.

Ejemplo de Datos Ordinales

ID	Nivel educativo
1	Primaria
2	Secundaria
3	Licenciatura
4	Maestría

Table: Datos ordinales: categorías con orden implícito.

Ejemplo de Datos Intervalares

ID	Temperatura (°C)
1	18
2	21
3	25
4	30

Table: Datos intervalares: tambien conocidos como datos de intervalo.

Ejemplo de Datos de Razón

ID	Peso (kg)
1	55
2	68
3	72
4	80

Table: Datos de razón: poseen cero absoluto, permiten proporciones.

Estructura de los Datos

- **Estructurados:** organizados en tablas (ej. CSV, bases de datos relacionales).
- **Semiestructurados:** tienen organización flexible (ej. JSON, XML).
- **No estructurados:** sin un formato fijo (ej. texto libre, imágenes, audio, video).

Ejemplo

Un proyecto puede incluir datos tabulares (ventas), texto (reseñas de clientes) e imágenes (fotos de productos).

Caso de estudio: Datos de estudiantes

Supongamos que tenemos información de un grupo de estudiantes de un curso:

- Nombre del estudiante
- Género
- Edad
- Semestre
- Calificación final

Caso de estudio: Datos de estudiantes

Supongamos que tenemos información de un grupo de estudiantes de un curso:

- Nombre del estudiante
- Género
- Edad
- Semestre
- Calificación final

¿Cómo clasificaríamos cada uno de estos datos?

Datos de estudiantes

Nombre	Género	Edad	Semestre	Calificación
Ana	F	20	3	85
Luis	M	22	5	92
María	F	21	4	78
Carlos	M	23	7	88

Clasificación de los datos

- **Nombre**: Nominal (categoría sin orden)
- **Género**: Nominal (F/M)
- **Edad**: Razón (cero tiene sentido, operaciones válidas)
- **Semestre**: Ordinal (tiene orden, diferencias no necesariamente iguales)
- **Calificación**: Intervalo/Razón (según escala)

Reflexión final

¿Por qué es importante distinguir el tipo de dato?

Reflexión final

¿Por qué es importante distinguir el tipo de dato?

- Determina qué operaciones matemáticas son válidas.
- Define qué gráficos o resúmenes estadísticos usar.
- Afecta los algoritmos de modelado.

Problemas comunes en datos crudos

Antes de analizar o modelar, los datos casi nunca vienen “perfectos”...

Problemas comunes en datos crudos

Antes de analizar o modelar, los datos casi nunca vienen "perfectos"...

Algunos problemas frecuentes:

- **Valores faltantes** (*¿qué hacemos con ellos?*)
- **Outliers y ruido** (*¿error o fenómeno real?*)
- **Variables irrelevantes o redundantes**
- **Formatos inconsistentes** (fechas, categorías, escalas)

Problemas comunes en datos crudos

Antes de analizar o modelar, los datos casi nunca vienen "perfectos"...

Algunos problemas frecuentes:

- **Valores faltantes** (*¿qué hacemos con ellos?*)
- **Outliers y ruido** (*¿error o fenómeno real?*)
- **Variables irrelevantes o redundantes**
- **Formatos inconsistentes** (fechas, categorías, escalas)

Estos retos abren la puerta al **preprocesamiento de datos**, tema central en nuestra siguiente sesión.