

Extensiones de Machine Learning

Diseño de un proyecto de Ciencia de Datos

JUAN A. BOTÍA BLAYA

*CIENCIA DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL
DEPT. INGENIERÍA DE LA INFORMACIÓN Y LAS COMUNICACIONES*

Universidad de Murcia

Grado en Ciencia e Ingeniería de Datos

2025/2026



2025/2026

Extensiones de Machine Learning

¿Cómo diseñar mi proyecto de Ciencia de Datos?



2025/2026

Extensiones de Machine Learning

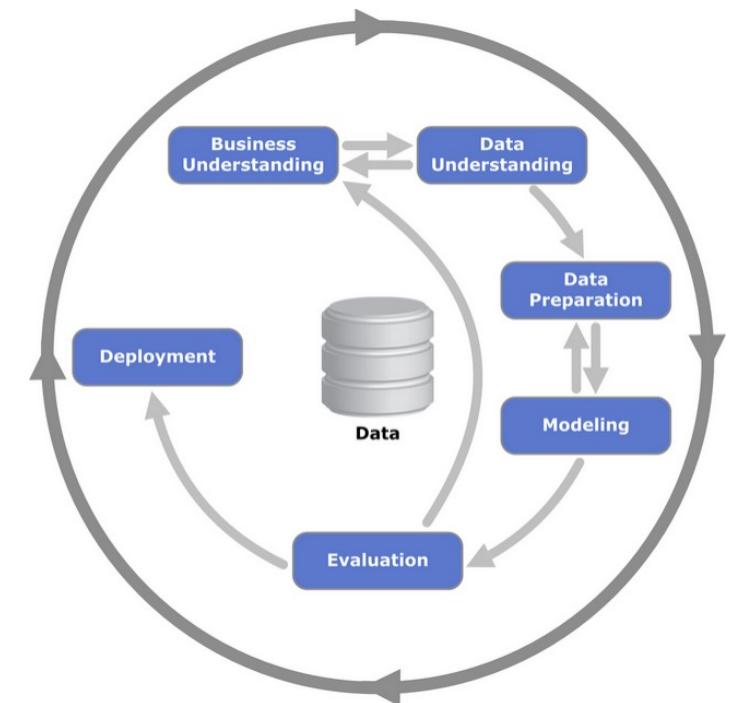
Tenemos una idea de análisis

- Queremos verificar una hipótesis
 - El nivel de lectura está negativamente asociado con la inversión en criptomonedas
 - Incrementar el precio de una suscripción premium no va a reducir los beneficios
 - La inflamación crónica está asociada con mayor riesgo de depresión
 - Queremos desarrollar un producto y necesitamos demostrar su viabilidad
 - Somos capaces de predecir la variable y, a partir de las características x, con una precisión mínima de z
 - Queremos pasárnoslo bien haciendo análisis de datos...
- ¿Cómo validamos la idea?



Pasos para diseñar un proyecto de Ciencia de datos

1. Definir el problema
2. Recolección de datos y caracterización
3. Limpieza y preparación de datos
4. Creación (i.e. modelado) y evaluación del modelo
5. Interpretación de resultados y conclusiones
6. Implementación de la idea
7. Documentación y preparación de la memoria



<https://www.datascience-pm.com/crisp-dm-2/>



1. Definir el problema

Pasos:

- **Identifica el objetivo**

- ¿Qué pregunta intentas responder? (Ej: *¿Podemos predecir la rotación de clientes?*)

- **Entiende el contexto**

- ¿Por qué es importante este problema? ¿Quién usará los resultados?

- **Establece criterios de éxito**

- ¿Qué define una buena solución? (precisión, interpretabilidad, impacto económico, etc.)

- **Define las restricciones**

- tiempo, disponibilidad de datos, recursos computacionales, etc.



2. Recolección de datos y caracterización

Pasos:

- Reúne los datos:
 - desde archivos CSV, bases de datos SQL, APIs, sensores, etc.
- Analiza las variables:
 - tipos de datos, distribución, valores ausentes.
- Realiza un Análisis Exploratorio de Datos (EDA):
 - Visualiza variables y relaciones.
 - Detecta valores atípicos y correlaciones.
 - Formula hipótesis preliminares.



3. Limpieza y preparación de datos

Pasos:

- Trata valores faltantes, duplicados y formatos inconsistentes.
- Codifica variables categóricas (por ejemplo, one-hot encoding).
- Escala o normaliza características si es necesario.
- Crea nuevas variables informativas (ingeniería de características).
- Divide los datos en entrenamiento, validación y prueba.



4. Modelado y evaluación

Pasos:

- Selecciona los algoritmos:
 - regresión, clasificación, clustering, etc.
- Entrena y ajusta:
 - usa validación cruzada, grid search o random search.
- Evalúa:
 - compara métricas (precisión, F1, ROC-AUC, RMSE, etc.).
- Interpreta:
 - analiza la importancia de las variables o usa métodos explicativos (SHAP, LIME).



5. Interpretación de resultados y conclusiones

Pasos:

- Resume los principales hallazgos del modelo.
- Presenta visualizaciones claras y comprensibles.
- Analiza sesgos, limitaciones y validez de los resultados.
- Conecta los resultados con la pregunta original y su impacto real.



7. Documentación y preparación de la memoria



Facultad
de Informática

**Normativa Interna para la oferta, asignación,
elaboración y defensa de los Trabajos Fin de
Grado y Fin de Máster de las titulaciones
impartidas en la Facultad de Informática de
la Universidad de Murcia**

(Aprobada por Junta de Centro el 19 de junio de 2025)

Estructura sugerida en la normativa de 2023

Portada y contraportada.

Autorización, si procede, de empresa/institución

Resumen

Extended Abstract (min 2000 palabras)

Índice

1 Introducción

2 Estado del Arte

3 Objetivos y metodología

4 Descripción del Diseño, trabajo realizado, pruebas, resultados,
etc.

5 Conclusiones y vías futuras

Bibliografía

Anexos, si procede.



2025/2026

Extensiones de Machine Learning