
Clase 061 — CRISP-DM como framework metodológico

Parte: 1 — Machine Learning Clásico · Fuente: CRISP-DM 1.0 spec + Géron cap. 2.

Duración estimada: 50 min.

Clase 061 — CRISP-DM como framework metodológico

Parte: 1 — Machine Learning Clásico · Fuente: CRISP-DM 1.0 spec + Géron cap. 2. Duración estimada: 50 min.

Objetivo

Que el alumno entienda CRISP-DM (Cross-Industry Standard Process for Data Mining) como esqueleto metodológico para todo proyecto de ML/DS — sus 6 fases, su naturaleza iterativa (no en cascada), y cuándo conviene complementarlo o reemplazarlo por TDSP de Microsoft o por el "ML lifecycle moderno" con MLOps. La idea es dejar de improvisar el orden del trabajo y tener un mapa al que volver cuando un proyecto se trabe.

Resultados de aprendizaje

Al finalizar la clase, el alumno podrá:

1. Enumerar y explicar las 6 fases de CRISP-DM y los entregables típicos de cada una.
2. Identificar la fase actual de un proyecto real y la próxima transición esperada.
3. Definir business success criteria antes de tocar datos, distinguiéndolos de métricas técnicas (accuracy, RMSE).
4. Reconocer iteraciones legítimas (volver de Evaluation a Business Understanding) vs. retrabajo por mala planificación.
5. Comparar CRISP-DM con TDSP y el ML lifecycle moderno y elegir según contexto (PoC, equipo chico, producción seria, MLOps).

Temas

#	Tema	Por qué importa
1	Business Understanding	Sin objetivo de negocio claro, el modelo n
2	Data Understanding	EDA, calidad, volumen, disponibilidad — de
3	Data Preparation	El 60–80% del tiempo real; limpieza, featu
4	Modeling	Selección de algoritmos, tuning, validació
5	Evaluation	Compara con business success criteria, no
6	Deployment	Puesta en producción, monitoreo, retrainin
7	Iteración + comparación TDSP / ML lifecycl	CRISP-DM no es cascada; existen alternativ

Definiciones y características

Business Understanding

: Fase 1. Traduce el problema de negocio a un problema de DS. Entregables: objetivos de negocio, criterios de éxito del negocio (ej. "reducir churn 5pp en 6 meses"), inventario de recursos, plan de proyecto. Sin esto, todo lo demás es ejercicio académico.

Data Understanding

: Fase 2. Recolectar datos, describirlos, explorarlos (EDA), verificar calidad. Entregable: reporte de calidad y primeras hipótesis. Si los datos no alcanzan, se vuelve a Business Understanding a renegociar scope.

Data Preparation

: Fase 3. Limpieza, integración, formateo, feature engineering, splits (train/val/test). Suele consumir el 60–80% del proyecto. Entregable: dataset final modelable + script reproducible.

Modeling

: Fase 4. Selección de técnicas, diseño de test, entrenamiento, tuning. Entregable: modelos candidatos con métricas técnicas. Si el modelo requiere otro formato de datos, se vuelve a Data Preparation.

Evaluation

: Fase 5. ¿El modelo cumple los business success criteria definidos en fase 1? No solo accuracy — costo del falso positivo, latencia, fairness, interpretabilidad. Entregable: decisión go/no-go.

Deployment

: Fase 6. Producción, monitoreo de drift, plan de retraining, documentación, handoff. Termina con un sistema funcionando, no con un notebook.

Iteración

: CRISP-DM no es cascada — las flechas van en ambos sentidos entre fases adyacentes, y la flecha externa cierra el ciclo (Deployment → nueva ronda de Business Understanding). Iterar es la norma, no la excepción.

Business success criteria

: Métricas en términos del negocio (dinero, conversión, tiempo, NPS) acordadas con stakeholders antes de modelar. Distintas de métricas técnicas (accuracy, F1, RMSE). Sin ellas, no hay forma objetiva de cerrar la fase de Evaluation.

Dataset / recursos

No hay dataset. La clase es conceptual + un caso práctico para aplicar las 6 fases (sugerido: predicción de churn telco, o un caso del alumno).

Ejercicios

1. Identificar la fase. Dadas 6 situaciones de proyecto (ej. "estoy probando XGBoost vs Random Forest", "estoy graficando histogramas para ver outliers"), asigná cada una a su fase CRISP-DM.
2. Business success criteria. Para un caso de detección de fraude bancario, escribí 3 criterios de éxito de negocio (no técnicos) y 3 técnicos. Distinguilos claramente.
3. Aplicar las 6 fases a un caso real. Elegí un problema (churn de Netflix, recomendación de productos, predicción de demanda en una panadería). Redactá un párrafo por cada fase con entregables concretos. Mínimo 1 iteración explícita (ej. "vuelvo de Modeling a Data Preparation porque...").
4. Comparar frameworks. Tabla de 3 columnas (CRISP-DM, TDSP, ML lifecycle moderno con MLOps) y 5 filas (origen, fases, foco, herramientas, cuándo usarlo).
5. Detectar iteración legítima vs retrabajo. Dados 4 escenarios de "estoy volviendo a una fase anterior", clasificalos como iteración esperada o como síntoma de mala planificación en la fase anterior.

Homework verificable

Documento Markdown (1–2 páginas) tomando un proyecto propio (real o sintético) y desarrollando las 6 fases de CRISP-DM con: (a) objetivo de negocio en una frase; (b) 2 business success criteria cuantificables; (c) descripción mínima de datos disponibles; (d) plan de data prep en bullets; (e) 2 algoritmos candidatos justificados; (f) cómo se evaluará contra los criterios de (b); (g) cómo se desplegaría y monitorearía.

Criterio de aceptación: Las 6 fases están presentes, los success criteria son cuantificables (con número y unidad), y la fase de Evaluation mapea explícitamente a los criterios de (b) — no solo a métricas técnicas.

Errores comunes

Síntoma	Causa y cómo arreglar
El modelo tiene gran accuracy pero el nego	Saltaste Business Understanding. Fix: volv
Tratar CRISP-DM como cascada (una fase des	Malentendido del framework. Fix: el spec o
Confundir métricas técnicas con success cr	Reportás F1=0.87 al gerente; el gerente no
Saltar Deployment ("ya tengo el notebook")	El proyecto muere en el laptop del DS. Fix
Hacer EDA infinito sin avanzar a Modeling	Parálisis en Data Understanding. Fix: time

Preguntas frecuentes

¿CRISP-DM o TDSP o ML lifecycle moderno?

Depende. CRISP-DM (1999, IBM/SPSS) es agnóstico de herramientas y sigue siendo el más enseñado — bueno como esqueleto mental y para proyectos chicos/PoC. TDSP (Microsoft, 2016) es más prescriptivo, define roles (data scientist, engineer, PM), templates de carpetas y artefactos Git — bueno para equipos. ML lifecycle moderno (con MLOps: feature stores, model registry, CI/CD de modelos, monitoring de drift) es lo que se usa cuando el modelo está en producción seria. No son excluyentes: CRISP-DM como mapa conceptual + TDSP como estructura de equipo + MLOps como tooling de producción.

¿Tengo que pasar por las 6 fases en orden estricto?

No. Las flechas entre fases adyacentes son bidireccionales y hay un loop externo. Lo que sí es no negociable: no podés saltarte Business Understanding ni Evaluation.

¿Cuánto tiempo lleva cada fase?

Regla aproximada: 10% Business, 10% Data Understanding, 50–70% Data Preparation, 10% Modeling, 10% Evaluation, variable Deployment. Si Modeling te consume el 50%, algo está mal — probablemente saltaste prep.

¿CRISP-DM aplica a deep learning?

Sí, conceptualmente. Las fases son las mismas. Cambia el detalle en Data Preparation (tensores, augmentation), Modeling (arquitecturas, GPU) y Deployment (serving de modelos pesados, ONNX). Pero el esqueleto se sostiene.

¿Y si el negocio no sabe qué quiere?

Ese es exactamente el trabajo de Business Understanding — ayudar al negocio a articularlo. Si después de varias sesiones no hay objetivo claro, el proyecto no está listo para empezar. Documentalo y volvé cuando lo

esté.

Referencias

- CRISP-DM 1.0 step-by-step data mining guide (Wirth & Hipp, 2000)
- Géron, cap. 2 End-to-End Machine Learning Project.
- Microsoft Team Data Science Process (TDSP)
- Google ML Lifecycle / MLOps levels

Siguiente clase

Clase 062 — Clasificación binaria con MNIST

Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
# Imports y configuración inicial
```

Archivos complementarios

- notebook.ipynb