El Proceso KDD y la Planificación de Proyectos de Minería de Datos

Curso de Minería de Datos

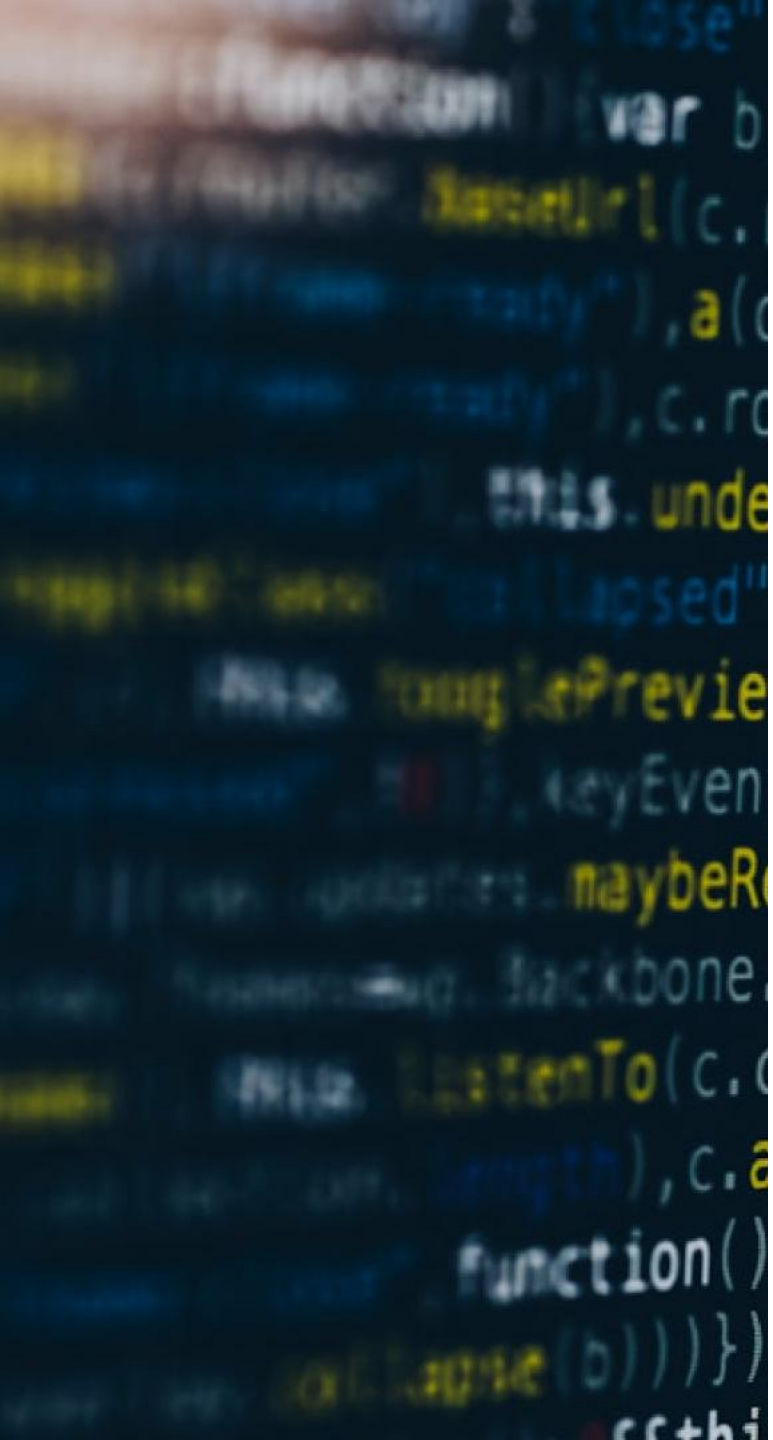
Maestría en Minería de Datos

Universidad Tecnológica Nacional - Rosario

Profesor Rodrigo Kataishi, Ph.D.

CONICET / UNTDF

rkataishi@untdf.edu.ar



Qué vamos a hacer?

Aproximación detallada al proceso de **Descubrimiento de Conocimiento en Bases de Datos (KDD)** como marco operativo fundamental para proyectos de minería de datos.

Detallaremos:

- Cada **fase** del proceso KDD.
- Sus **objetivos** específicos.
- **Técnicas** asociadas a cada etapa.
- **Decisiones clave** a tomar.
- **Productos esperados** en cada fase.

El Proceso KDD: Estructura General y Planificación

1. ¿Qué es el proceso KDD?

- **Definición:**
  KDD (Knowledge Discovery in Databases) es el proceso **completo e iterativo** de transformación de datos crudos en **conocimiento útil y accionable** mediante la aplicación integrada de técnicas:
  - Computacionales
  - Estadísticas
  - Analíticas

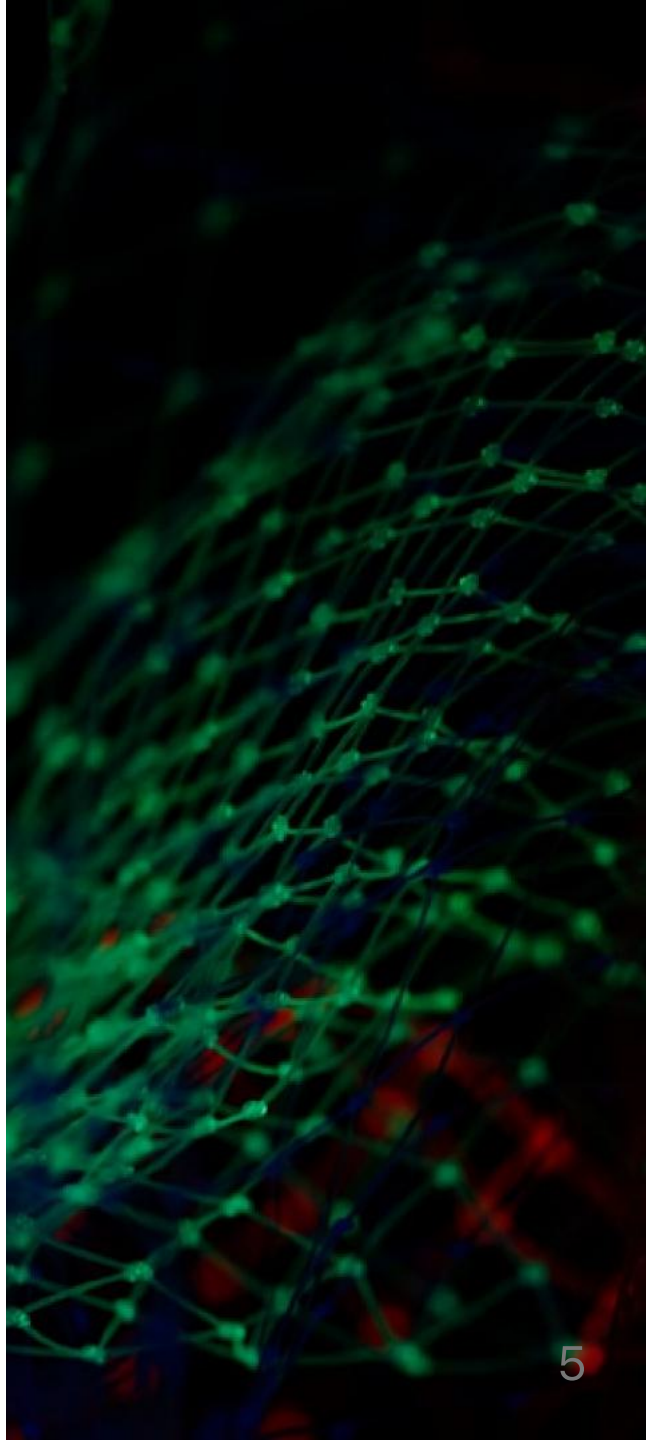
Fases y Objetivos Principales del KDD (Fayad et al., 1996)

El KDD estructura el proceso técnico a partir del acceso a los datos:

Fase	Objetivo principal
Selección	Extraer el subconjunto de datos relevante del universo disponible.
Preprocesamiento	Limpiar, corregir inconsistencias y homogenizar los datos seleccionados.
Transformación	Convertir los datos preprocesados a formatos adecuados para la minería.
Minería de datos (DM)	Aplicar algoritmos para descubrir patrones significativos.
Evaluación e Interpretación	Validar la relevancia y utilidad de los patrones y generar conocimiento accionable.

Características Clave y Planificación del KDD

- **Características del Proceso KDD:**
  - **No lineal e Iterativo:** Frecuentes ciclos de retroalimentación entre fases.
  - **Integración Multidisciplinar:** Requiere conocimientos de dominio, estadística y computación.
  - **Alcance Amplio:** Cubre desde la obtención de datos hasta la interpretación del conocimiento.
  - **Dependiente del Contexto:** Se adapta a los objetivos específicos y a la naturaleza de los datos.



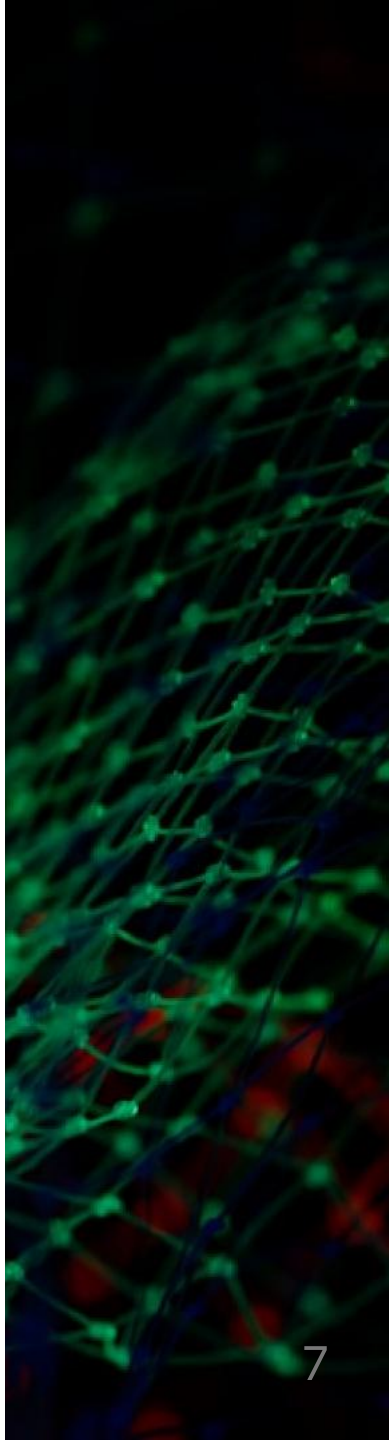
El KDD estructura el proceso técnico a partir del acceso a los datos:

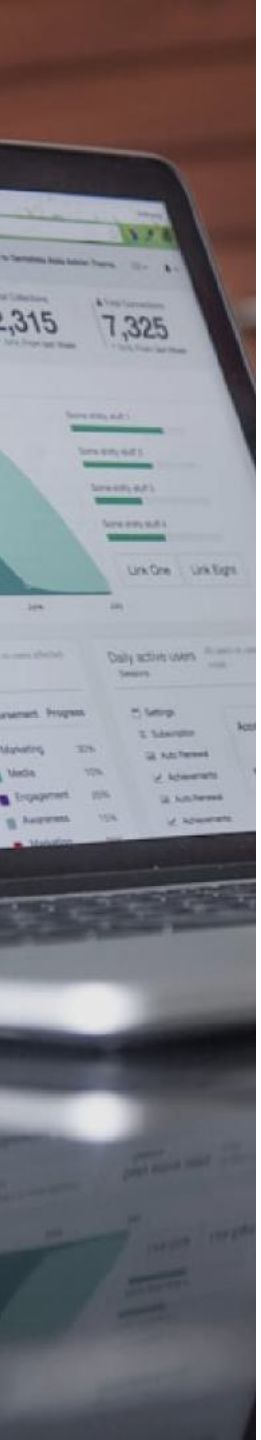
Fase	Objetivo principal
Selección (Y ANTES??)	Extraer el subconjunto de datos relevante del universo disponible.
Preprocesamiento	Limpiar, corregir inconsistencias y homogenizar los datos seleccionados.
Transformación	Convertir los datos preprocesados a formatos adecuados para la minería.
Minería de datos (DM)	Aplicar algoritmos para descubrir patrones significativos.
Evaluación e Interpretación	Validar la relevancia y utilidad de los patrones y generar conocimiento accionable.

Características Clave y Planificación del KDD

- **Planificación y Formulación del Problema:**
    - Una buena y clara **formulación del problema** es una *condición* crucial, pero no es una fase explícita del KDD original.
- **El Dilema:** Si el proceso *inicia* formalmente con la selección de datos... ¿Qué pasa si no hay una pregunta clara definida previamente?

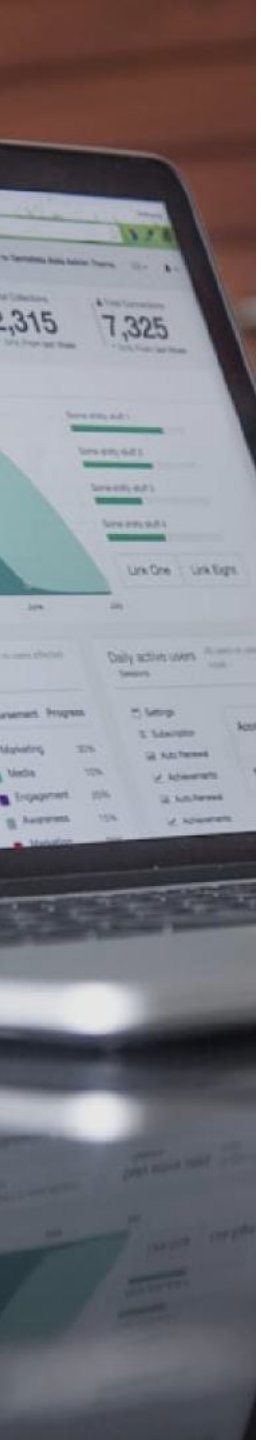
➡️ **Riesgo de "Naufragio en los Datos"**





No Todo es KDD, pero KDD es un Todo

- **El KDD** como **proceso técnico pionero** en la extracción de patrones datos. Se proyecta enfocado en las fases *operativas* una vez los datos están disponibles.
- **Lecciones de la Práctica:** La complejidad va más allá de los algoritmos. Fases externas al núcleo técnico son vitales:
  - **Formulación del Problema:** Definir qué se busca y por qué.
  - **Comprensión del Contexto:** Entender el dominio y las necesidades del negocio.
  - **Comunicación e Implementación:** Traducir hallazgos en acciones.



No Todo es KDD, pero KDD es un Todo

- **Luego de la implementación generalizada y el uso de KDD, se avanzó en la identificación de algunos límites:** surgieron otros modelos que abordan dimensiones clave ausentes en KDD:
  - **SEMMA (SAS, 1998):** Enfoque técnico exploratorio (Sample, Explore, Modify, Model, Assess).
  - **CRISP-DM (2000):** Foco explícito en objetivos de negocio, iteración y despliegue.
  - **Data Science Pipeline (ca. 2012–):** Visión modular, flexible y pragmática.
- **KDD Reubicado:** Estos modelos sitúan al KDD como el **núcleo técnico** dentro de un proceso más amplio, enfatizando la estrategia y la aplicación.

Macro-etapa	Fase / componente	Descripción técnica	KDD (1996)	SEMMA (1998)	CRISP-DM (2000)	Data Science (2012 –)
I. Formulación	Definición del problema	Identificar qué pregunta se intenta responder, para qué y con qué criterios de éxito			×	×
	Comprensión del contexto / dominio	Relevar actores, restricciones, objetivos y entorno de uso del análisis			×	×
II. Exploración inicial	Comprensión exploratoria de los datos	Evaluar calidad, tipo, estructura, fuentes y limitaciones del dataset		× (Explore)	×	×
	Acceso y selección de datos	Obtener el subconjunto más relevante para el problema	×	× (Sample)	×	×
III. Preparación	Preprocesamiento / limpieza	Corregir errores, tratar valores faltantes, unificar formatos	×	× (Modify)	×	×
	Transformación de variables	Estandarizar, codificar, crear variables derivadas	×	× (Modify)	×	×
	Reducción / selección de variables	Filtrar, comprimir o priorizar información relevante	×	× (Modify)	×	×
IV. Modelado	Aplicación de algoritmos	Ejecutar técnicas de clasificación, regresión, segmentación, asociación, etc.	×	× (Model)	×	×
	Evaluación técnica del modelo	Validar desempeño según métricas estadísticas (accuracy, AUC, etc.)	×	× (Assess)	×	×
V. Interpretación	Análisis de patrones y comprensión del modelo	Interpretar significados, reglas, límites y riesgos del conocimiento extraído	×		×	×
	Evaluación contextual y utilidad	Evaluar si el resultado es relevante, accionable y comprensible para el usuario	×		×	×
VI. Despliegue	Implementación / comunicación	Incorporar modelos al flujo de trabajo, generar reportes, facilitar decisiones			×	×
	Iteración y mejora	Realimentar el proceso con nuevos datos o ajustes según desempeño observado	× (implícita)	×	×	×

Volvamos a KDD: El Inicio en la Selección

Revisando las fases centrales del KDD:

Fase	Objetivo principal
Selección	Extraer el subconjunto de datos relevante
Preprocesamiento	Limpiar, corregir y homogenizar
Transformación	Convertir los datos a formatos adecuados
Minería de datos	Aplicar algoritmos para descubrir patrones
Evaluación e interpretación	Validar los resultados y generar conocimiento útil

El Dilema: Si el proceso *inicia* formalmente con la selección de datos... ¿Qué pasa si no hay una pregunta clara definida previamente?

Riesgo de "Naufragio en los Datos"

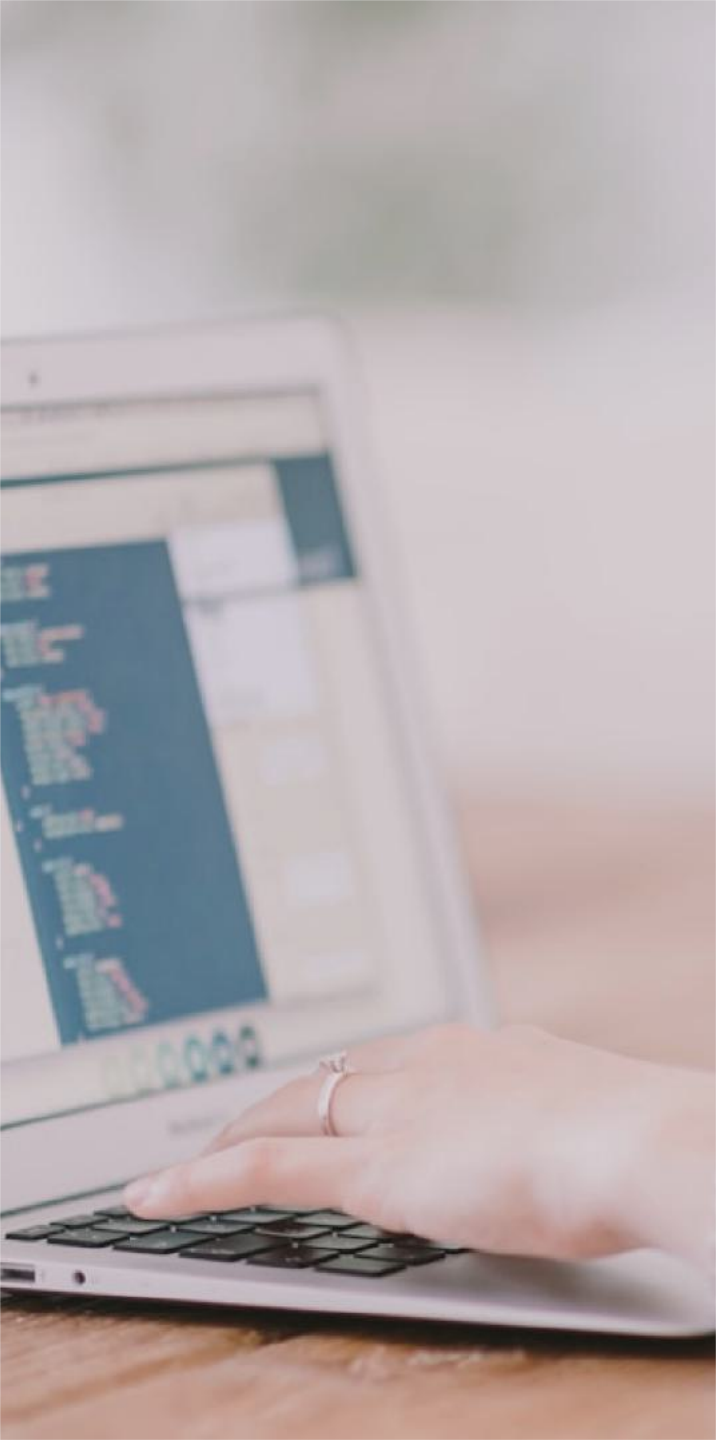
Exploración Agnóstica (sin supuestos): ¿Cuándo se Justifica?

Explorar los datos sin hipótesis previas fuertes (enfoque agnóstico) tiene sentido cuando:

- **Alta Incertidumbre / Dominio Nuevo:** Se carece de conocimiento previo sólido sobre el fenómeno o el conjunto de datos.
- **Búsqueda de Descubrimiento Estructural:** El objetivo principal es entender la estructura latente (e.g., ¿existen grupos naturales?) más que validar una idea preexistente.
- **Detección de lo Inesperado:** Se desea encontrar patrones, relaciones o anomalías no anticipadas (útil en clustering, reglas de asociación, detección de outliers).
- **Generación de Hipótesis:** La exploración inicial sirve como base para formular hipótesis que luego serán probadas rigurosamente.

Es habitual en:

- Fenómenos poco conocidos o complejos.
- Análisis preliminares para construir modelos teóricos.
- Monitorización de sistemas dinámicos (finanzas, redes sociales, bioinformática).



Entonces: Definición Previa del Problema?

Ante la posibilidad de "naufragio en los datos":

- **Práctica Común:** Se adopta una **fase explícita de definición del problema** *antes* de iniciar el ciclo KDD técnico.
    - **Incluye:** Entrevistas con expertos, identificación de metas, mapeo de acciones posibles, delimitación de restricciones.
- **KDD como Núcleo Operativo:** El proceso KDD (selección, preproc., etc.) se mantiene como el motor técnico, pero **precedido y guiado** por esta fase estratégica (inspirada en CRISP-DM, etc.).
- **Tensión Metodológica:** Esto introduce una dinámica clave:
    - **Descubrimiento Guiado por Objetivos:** Enfocado en responder preguntas específicas.
    - **Exploración Agnóstica Orientada por Datos:** Abierta a patrones emergentes no previstos.

Tensión Epistemológica: Guiado vs. Agnóstico

¿Es contradictorio tener objetivos definidos y buscar descubrimiento? No necesariamente, pero es una **tensión real** entre dos lógicas:

Enfoque	Lógica dominante	Supuesto epistémico	Riesgo principal
Exploración agnóstica	*Data-driven* (inducción)	El patrón emerge desde los datos	Hallazgos espurios, triviales; fallos en identificar info clave (errores Tipo I/II)
Exploración guiada	*Problem-driven* (abducción)	El patrón debe responder una pregunta	Ceguera ante lo no anticipado; omisión de fenómenos clave por sesgo confirmatorio

- Ambos enfoques pueden convivir, pero es crucial **declarar cuál domina en cada fase**.
- **Validación Crucial:** La exploración agnóstica **requiere validación rigurosa posterior** para evitar **patrones espurios** (regularidades aparentes o circunstanciales).

Cuidado: Tipologías Comunes de Patrones Espurios

Son regularidades aparentes que no reflejan estructuras reales ni tienen valor predictivo o interpretativo.

Tipo de error	Descripción	Ejemplo concreto
Correlaciones Espúreas	Fuerte asociación estadística entre variables sin relación causal o funcional real.	*Consumo de margarina correlacionado con divorcios (Vigen, 2015).*
Artefactos del Muestreo	Patrones inducidos por cómo se seleccionaron, organizaron o segmentaron los datos.	*Comparar grupos con orígenes distintos sin controlar composición (sesgo de selección).*
Coincidencias Temporales / Agregación	Alineamientos cronológicos o agregaciones que no implican relación estructural (o la ocultan/invierten - Paradoja de Simpson).	*Aumento de búsquedas "helado" y "ruido urbano" en verano (causa común: calor).*
Sobreajuste (Overfitting)	Modelo que captura ruido o particularidades del set de entrenamiento, sin generalizar.	*Árbol de decisión complejo que clasifica 100% en entreno por artefacto de codificación.*
Regularidades Inducidas Artificialmente	Patrones que surgen por decisiones técnicas (imputación, codificación, discretización).	*Valor "-99" (imputado para NAs) se vuelve predictor clave sin significado real.*

Profundizando: Artefactos del Muestreo

Patrones que no reflejan el fenómeno, sino **consecuencias del diseño o sesgos del dataset**:

Tipo de artefacto	Descripción técnica	Ejemplo específico
Muestreo no representativo	La muestra no refleja la población objetivo, generando patrones no generalizables.	*Modelo de abandono escolar entrenado solo en escuelas privadas urbanas falla en rurales/públicas.*
Comparación de grupos con composición distinta (Sesgo de Selección)	Se comparan subconjuntos sin controlar diferencias estructurales preexistentes.	*Programa A parece mejor que B, pero A selecciona estudiantes con notas más altas al ingreso.*
Corte temporal / espacial artificial	Segmentación (año, región) crea patrones ligados a estacionalidad, admin., o factores no controlados.	*"Productividad" baja en enero sin considerar cierres por vacaciones/mantenimiento.*
Errores por censura o truncamiento	Dataset solo incluye casos observados bajo ciertas condiciones, omitiendo otros sistemáticamente.	*Análisis de actividad en plataforma online solo con usuarios registrados, excluyendo abandonos iniciales.*
Agrupamiento oculto (Variable Latente /	Diferencias observadas se deben a una variable de agrupamiento no incluida en el análisis.	*Variación regional de consumo explicada realmente por política fiscal local (variable*

Validación Rigurosa del Descubrimiento Agnóstico

Un patrón descubierto solo se consolida como **conocimiento** si supera múltiples filtros:

Dimensión de validación	Estrategia operativa	Herramientas técnicas / Enfoques
Estadística	Evaluar significancia, robustez y generalización del patrón.	Hold-out, K-fold CV, Bootstrapping, Tests de hipótesis (p-values), Métricas (AUC, F1), Baselines.
Empírica / Predictiva	Comprobar desempeño en datos nuevos o contexto real.	Test externo (out-of-sample), Backtesting (series temporales), Pruebas A/B, Simulación.
Contextual / Semántica	Evaluar plausibilidad, coherencia con el dominio y interpretabilidad.	Validación por expertos, Revisión de literatura, Consistencia teórica, Interpretabilidad (SHAP, LIME).
Instrumental / Práctica	Medir utilidad para la toma de decisiones o acción, y valor agregado.	Impacto en KPIs, Análisis Costo-Beneficio, Comparación con alternativas, Feedback de usuarios.

Criterio Final (Fayyad et al., 1996): *Valid, Novel, Useful, Understandable*

Entonces, ¿trabajar con Objetivos limita el Descubrimiento?

No, al contrario: lo **enmarcan para que tenga sentido y valor.**

- **Acotan, No Anulan:** La formulación del problema no impone *qué* encontrar, pero sí **orienta la búsqueda** hacia dominios relevantes y evita la dispersión en ruido.
- **Dirigen la Exploración:** Permiten enfocar recursos y técnicas, sin impedir que **emerjan hallazgos inesperados** *dentro* de ese marco relevante.
- **Establecen Criterios de Valor:** Definen qué se considera un "buen" resultado. ¿Cómo saber si un patrón es interesante si no sabemos qué problema intentamos resolver?
- **Conectan Hallazgo y Acción:** Permiten evaluar si lo descubierto es **accionable** y útil en el contexto del problema.

La exploración sin marco puede ser fértil, pero CUIDADO, porque ante el desconocimiento total del fenómeno a analizar a menudo puede ser **ineficiente**.

No hay descubrimiento valioso si no hay un marco que permita reconocerlo, validarlo y actuar sobre él.

Perspectiva de Modelos Modernos (CRISP-DM, DS Pipeline)

Estos enfoques integran la exploración, pero de forma estratégica:

- **Fase Específica:** La exploración agnóstica (EDA) es una **etapa fundamental pero delimitada** (e.g., "Data Understanding" en CRISP-DM), realizada *tras* entender el problema.
- **Subordinación al Contexto:** Los hallazgos exploratorios se interpretan **en función** de los objetivos y el problema definido.
- **Integración con Validación:** Se combinan con técnicas robustas para **mitigar el riesgo de sobreajuste interpretativo** y patrones espurios.

Resultado: El descubrimiento se vuelve **más estratégico, menos aleatorio.** Se reconoce que los datos pueden "hablar", pero **no lo harán útilmente si no sabemos qué preguntar, qué escuchar y cómo interpretar.**

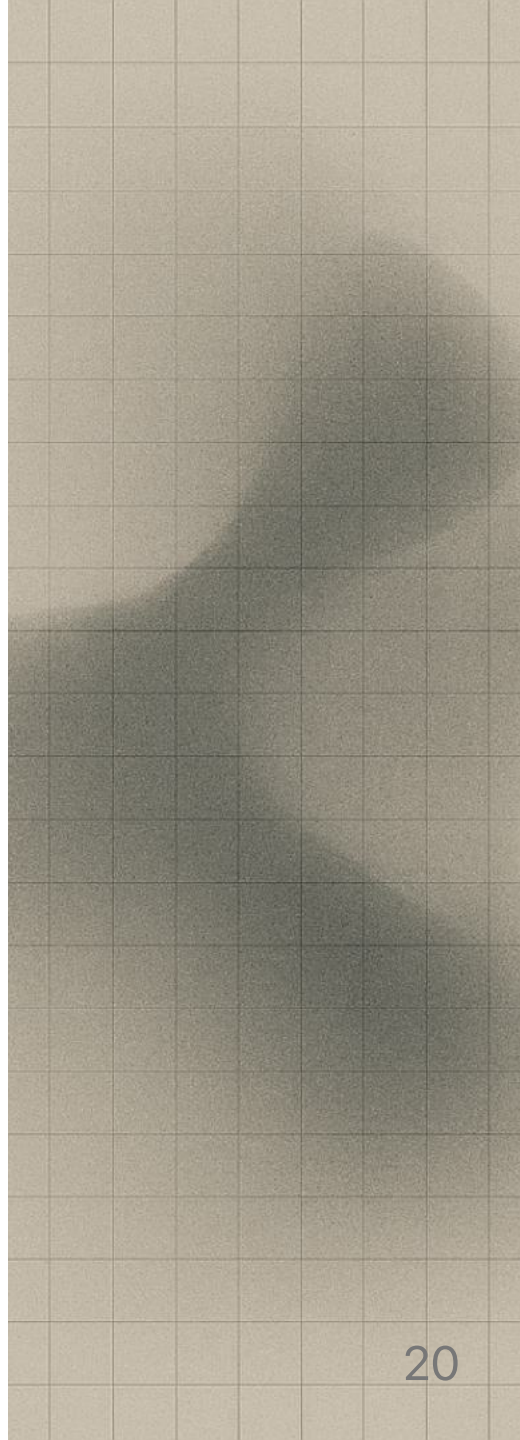
Punto Clave: Del Problema al Proyecto

Aquí la teoría del descubrimiento se encuentra con la **planificación práctica**:

- **Buscamos** lograr una **articulación clara entre la definición del problema (marco lógico) y su traducción operativa (marco técnico)**.
- **Formulación como Habilitante:** Ver la definición del problema no como una limitación, sino como una **condición necesaria para el descubrimiento estratégico y valioso.**

Funciones Clave de la Formulación del Problema:

1. **Acota** el espacio de búsqueda relevante.
2. **Define** los criterios de evaluación del éxito.
3. **Establece** restricciones operativas (tiempo, recursos, ética).
4. **Vincula** la técnica con la decisión y la acción.



Decisiones Clave en la Fase de Formulación del Proyecto

Antes de iniciar las fases técnicas, decisiones estratégicas guiarán el proyecto, definiendo una solución técnica viable y apropiada:

Decisión clave	Ejemplo de implicancia técnica / Preguntas a responder
Nivel de Automatización vs. Intervención Humana	¿Se requiere un modelo 100% autónomo o con validación humana? ¿Quién interpreta/usa los resultados?
Requisitos de Interpretabilidad	¿Es aceptable un modelo "caja negra" si es preciso? ¿O se necesitan reglas/factores comprensibles para justificar decisiones (regulación, confianza)?
Restricciones (Tiempo, Cómputo, Datos)	¿Plazo? ¿Capacidad computacional? ¿Datos estáticos o streaming? ¿Limitaciones legales de acceso o uso?
Criterios de Éxito del Proyecto	¿Maximizar precisión? ¿Descubrir algo novedoso? ¿Facilidad de implementación? ¿Robustez? ¿Impacto medible en KPIs (Key Performance Indicators)?

Producto Esperado: Plan de Trabajo Inicial

Culminada la toma de decisiones relacionadas con la formulación estructural, se construye un **Plan de Trabajo Inicial** que actúa como **puente entre el problema y la ejecución técnica**.

Contenido Esencial:

- **Objetivo Analítico Claro:** Qué se busca y por qué es relevante. Justificado.
- **Tipo de Tarea de Minería:** Identificación preliminar (Clasificación, Regresión, Clustering, Asociación, etc.).
- **Entradas Esperadas:** Descripción de datos disponibles (fuentes, volumen, estructura, calidad inicial percibida).
- **Métodos Candidatos:** Listado inicial de algoritmos/enfoques considerados apropiados.
- **Criterios de Validación:** Cómo se evaluará el desempeño (métricas estadísticas) y la utilidad (criterios contextuales/negocio).

Importante: Este plan **no es rígido**. Es una hoja de ruta inicial que **se revisa y actualiza** iterativamente. Su función clave: **transformar la pregunta en un proyecto estructurado.**

Resumen: Planificación Estratégica, exploración y objetivos

Para ganar eficiencia, y evitar potenciales "naufragios en los datos", antes de las fases operativas del KDD es recomenda avanzar en:

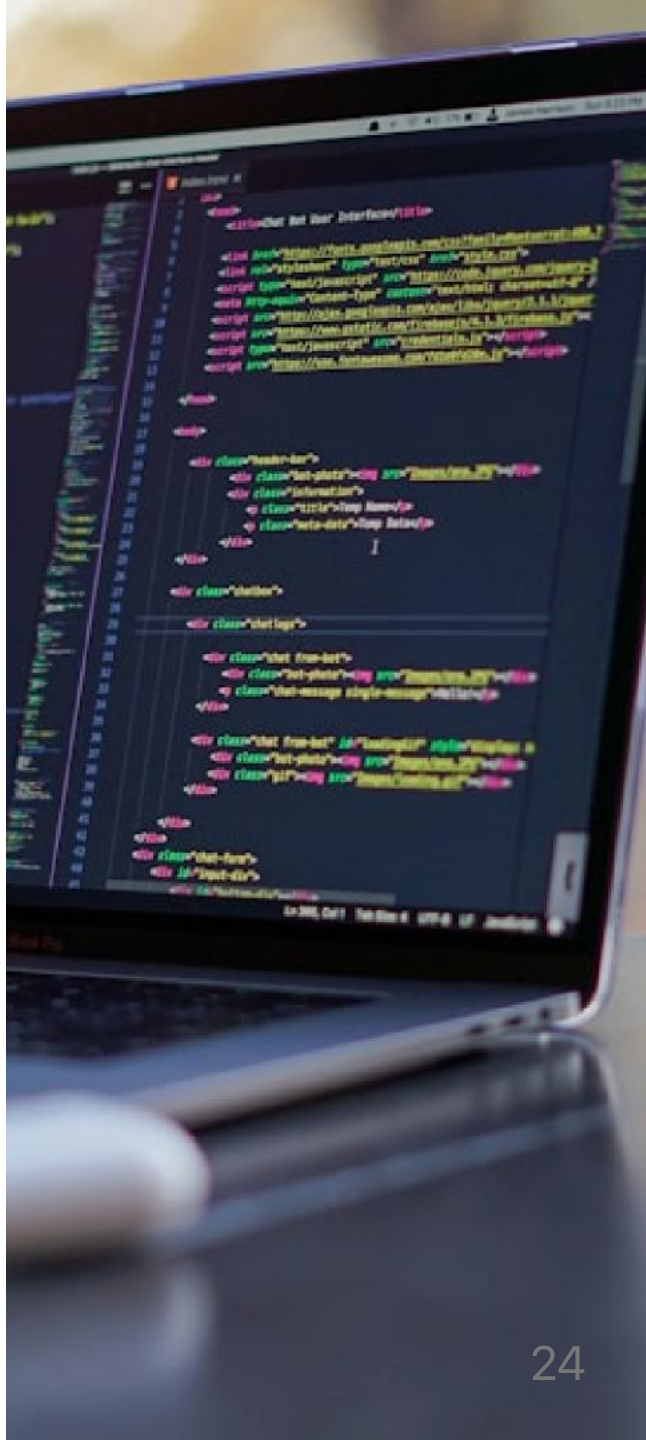
- **Fase Previa Indispensable: Formulación del Problema**
  - Definir: ¿Por qué quiero explorar datos? ¿Qué **pregunta** se intenta responder? | Identificar: ¿Cuál es el **uso potencial** del conocimiento extraído? | Determinar: ¿Qué **tipo de tarea analítica** está implicada?
- **Decisiones Clave al Inicio:**
  - Nivel de **automatización** vs. intervención humana. | Requisitos de **interpretabilidad**. | **Restricciones** temporales, computacionales, éticas. | **Criterios de éxito** (precisión, utilidad, novedad, robustez).
- **Producto Esperado de la Planificación:**
  - Un **Plan de Trabajo inicial** especificando: Objetivos, Fases, Datos y Entradas (volumen, contenido, calidad), Métodos candidatos, Criterios de validación.

Asegura que el esfuerzo técnico KDD esté alineado con un sentido, y por ende con objetivos relevantes.

Fase de Selección y Preprocesamiento

1. Fase de Selección

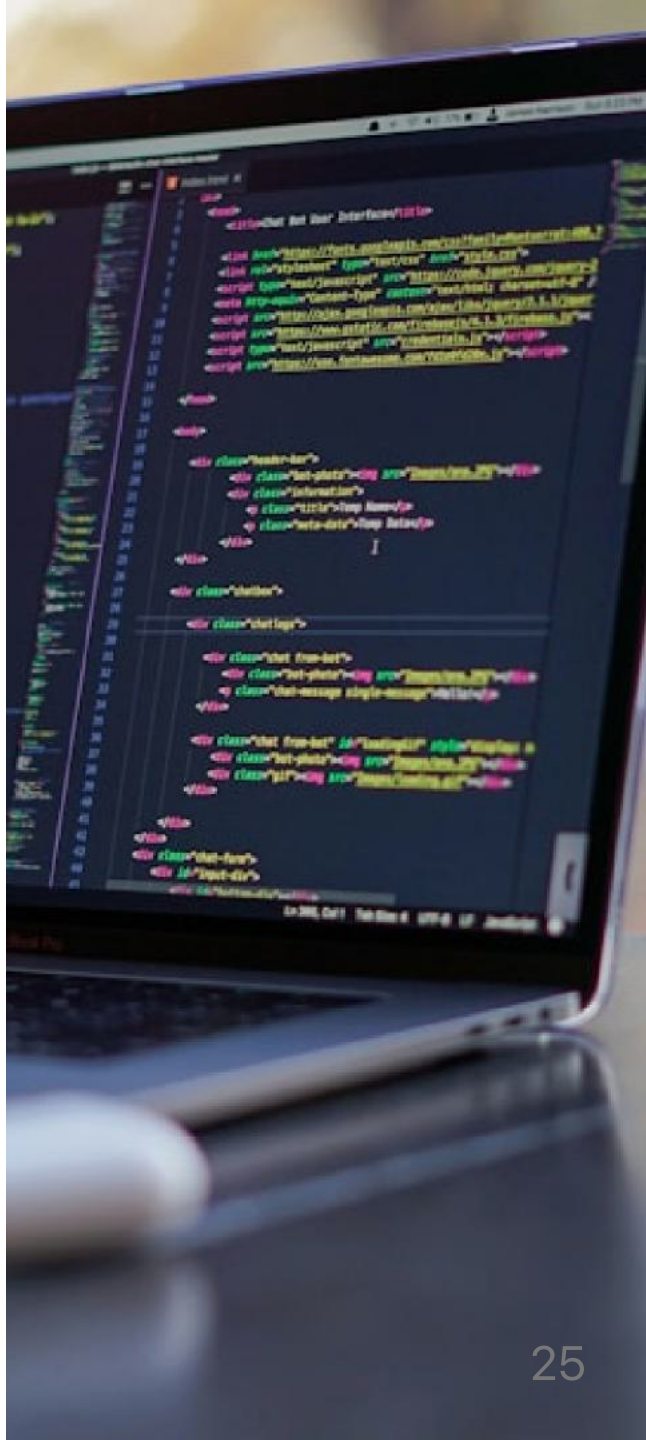
- **Objetivo:** Elegir los datos relevantes del universo disponible.
- **Tareas:**
  - Acceso a fuentes diversas (BBDD, APIs, archivos planos...).
  - Filtrado, submuestreo (simple, estratificado).
  - Extracción de variables (features) pertinentes al problema.
- **Técnicas:** SQL, Pandas (filtrado, `subsetting`), exploración de metadatos.
- **Resultado:** Subconjunto coherente, legible y relevante de datos para análisis.

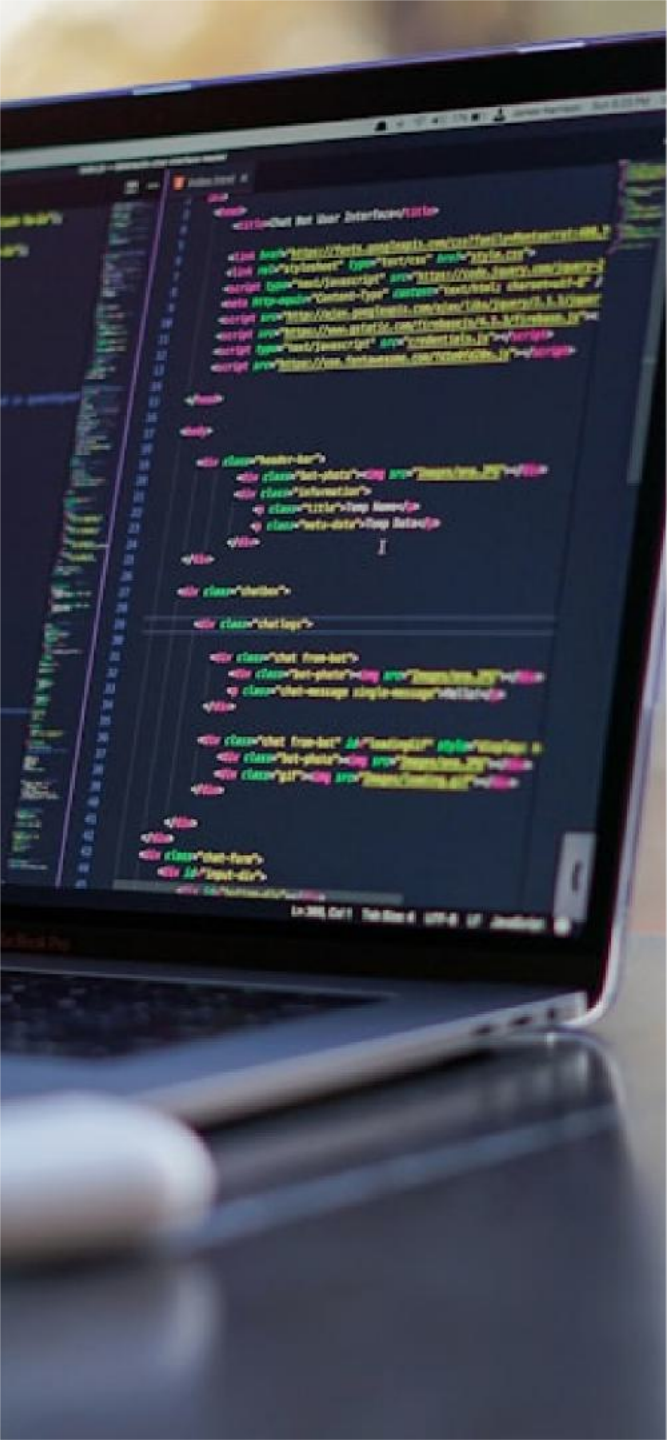


Fase de Selección y Preprocesamiento

2. Fase de Preprocesamiento

- **Objetivo:** Limpiar y preparar los datos seleccionados.
- **Tareas:**
  - Tratamiento de valores faltantes (NAs).
  - Corrección de errores (outliers, codificación incorrecta).
  - Homogeneización de tipos y formatos.
  - Detección/eliminación de duplicados.
  - Validación de consistencia.
- **Técnicas:** Imputación (media, KNN, etc.), conversión de tipos (`astype` , `to_datetime`), `drop_duplicates` , `fillna` , validaciones lógicas.
- **Resultado:** Dataset limpio, estructurado y sin errores sistemáticos graves.





Fase de Transformación y Reducción

1. Fase de Transformación

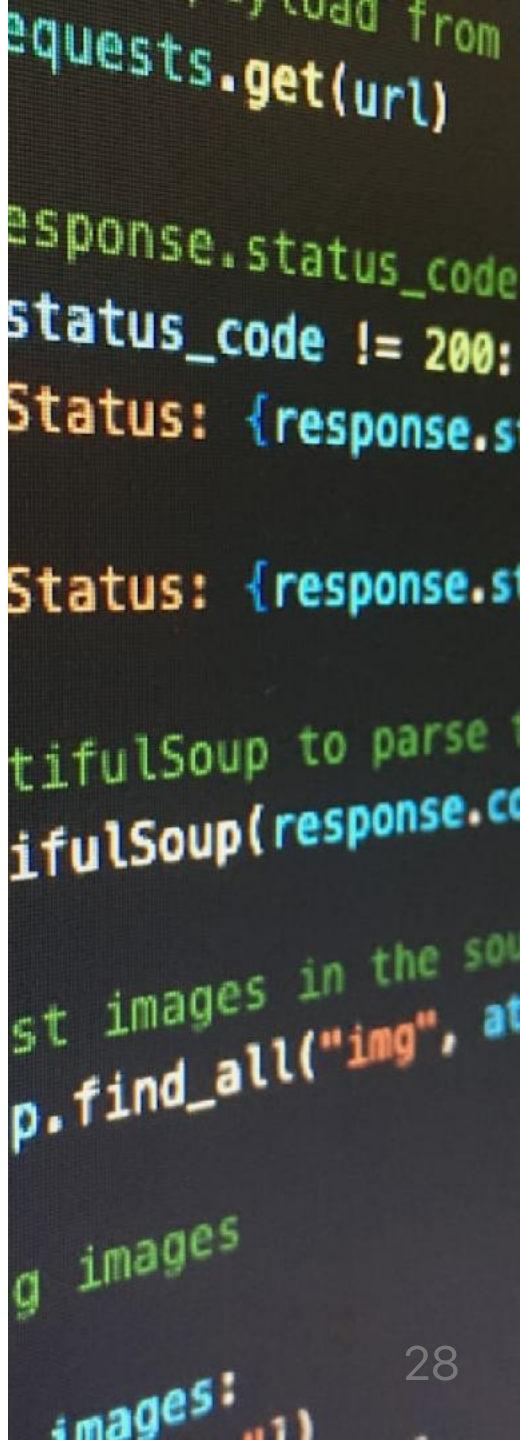
- **Objetivo:** Convertir datos preprocesados a formatos óptimos para algoritmos de minería.
- **Operaciones Típicas:**
  - trabajo con variables
  - Normalización/Estandarización (numéricas).
  - Codificación (categóricas: One-Hot, Label Encoding).
  - Ingeniería de Variables (creación de nuevas features: binning, interacciones, logs).
  - Manejo de formatos complejos (JSON anidado, XML).
- **Resultado:** Datos consistentes, variables adecuadas e informativas, variantes de variables informativas, categorización de datos, Matriz de datos (generalmente numérica) compatible con algoritmos, con variables informativas.

Fase de Exploración de Datos (EDA)

- **Objetivo:** Comprender la estructura, calidad y relaciones entre variables antes de transformar o modelar.
- **Tareas Principales:**
  - Estadísticas descriptivas.
  - Visualización univariada y bivariada.
  - Detección de outliers y valores faltantes.
  - Análisis de correlaciones.
- **Herramientas frecuentes:** Pandas, Seaborn, Matplotlib, Plotly, pandas-profiling.
- **Resultado:** Hipótesis preliminares, decisiones de limpieza, y guía para transformación posterior.

Fase de Transformación de Datos

- **Objetivo:** Adaptar los datos preprocesados a formatos compatibles con técnicas de minería.
- **Tareas Comunes:**
  - Normalización / estandarización de variables numéricas.
  - Codificación de variables categóricas.
  - Creación de nuevas variables (feature engineering).
  - Conversión de formatos (fechas, texto, listas, JSON).
- **Técnicas frecuentes:** MinMaxScaler, StandardScaler, One-Hot Encoding, Log-transform, Binning.
- **Resultado:** Matriz estructurada, numérica y lista para aplicar algoritmos.



Técnicas...

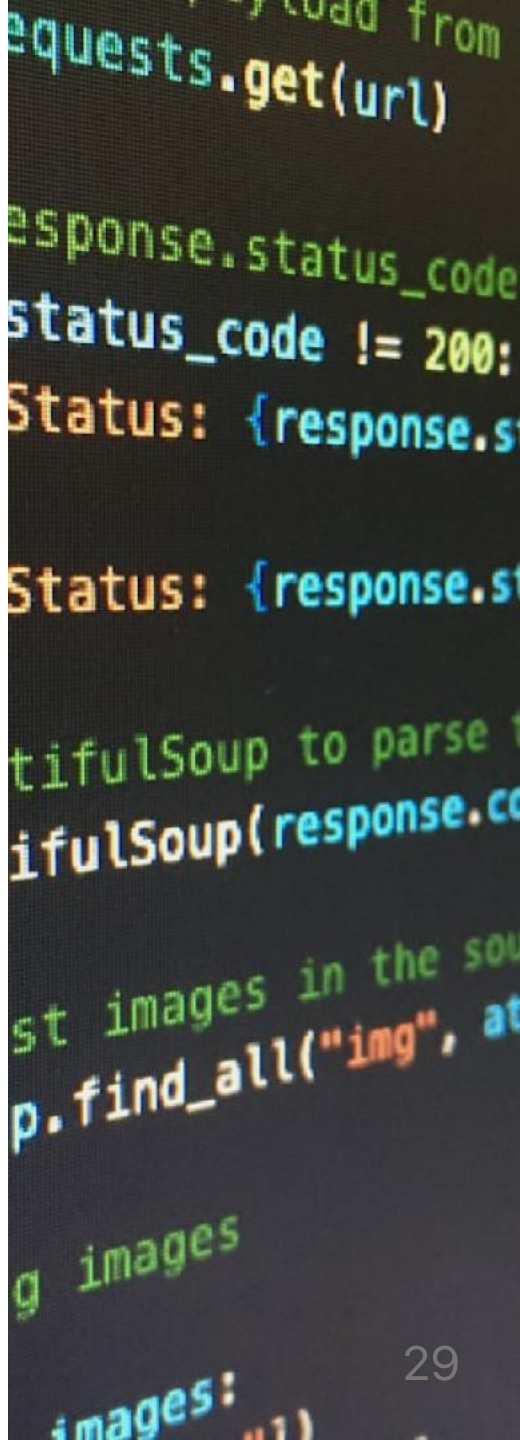
• MinMaxScaler
Escala los valores numéricos al rango 0, 1.
Útil cuando se requiere mantener la proporcionalidad y los algoritmos son sensibles a la escala (e.g., KNN, redes neuronales).

• StandardScaler
Estandariza los valores para que tengan media 0 y desviación estándar 1.
Recomendado para modelos que asumen datos centrados (e.g., regresión lineal, SVM, PCA).

• One-Hot Encoding
Convierte variables categóricas en variables binarias (dummies), una por cada categoría.
Evita que los algoritmos interpreten una relación ordinal inexistente.

• Log-transform
Aplica una transformación logarítmica a variables numéricas.
Útil para reducir asimetrías (skewness) y comprimir rangos muy amplios (e.g., ingresos, población).

• Binning
Agrupa valores numéricos continuos en intervalos discretos (bins).
Puede usarse para simplificar modelos o capturar no linealidades (e.g., edad en rangos).

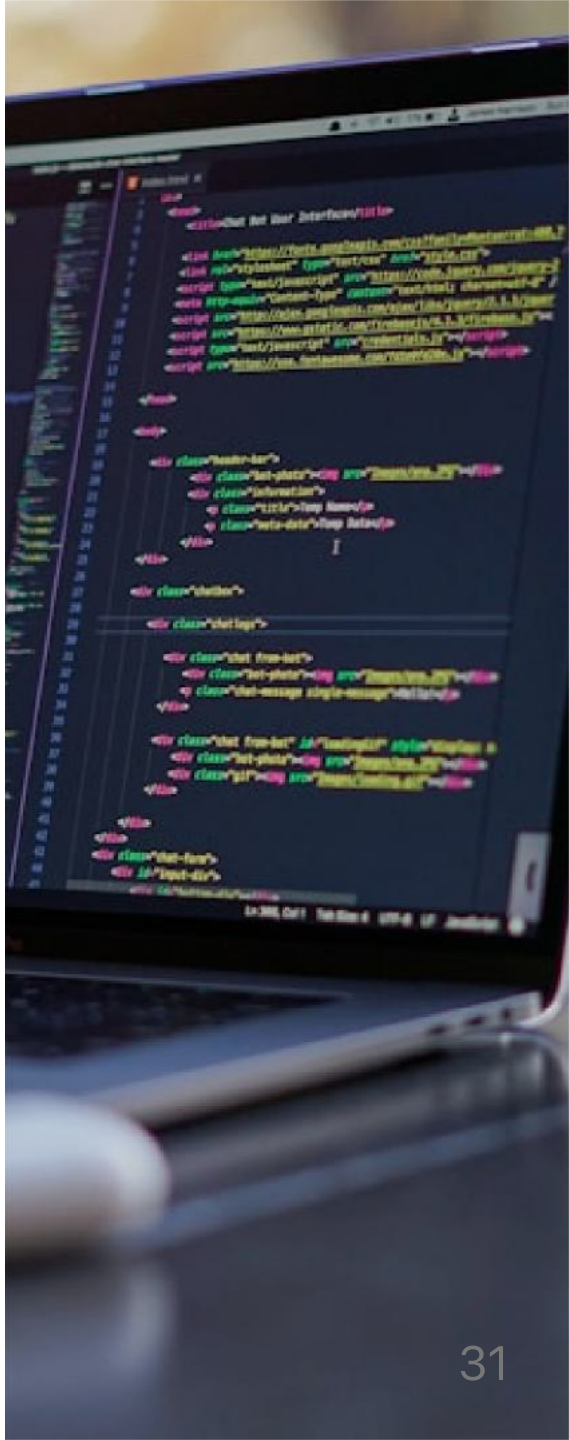


Técnica Clave: Feature Engineering

- **Objetivo:** Crear variables que capturen mejor la estructura del fenómeno analizado.
- **Técnicas comunes:**
  - Derivación: `ingreso_per_cápita = ingreso / miembros_hogar`
  - Interacciones: producto, razón, diferencia entre columnas.
  - Transformaciones: log(x+1), raíz, cuadrado.
  - Binning contextual: edad en rangos (18–29, 30–45, etc.).
  - Variables temporales: día, mes, año, diferencia temporal.
  - Indicadores booleanos: presencia/ausencia, condición cumplida.
  - Agrupamientos categóricos: etiquetas raras como "otros".
- **Importancia:**
  Una buena ingeniería de variables mejora más que cambiar de algoritmo.

Fase de Reducción de Dimensionalidad

- **Objetivo:** Simplificar la estructura del dataset preservando información clave.
- **Técnicas Comunes:**
  - PCA (Análisis de Componentes Principales).
  - Selección por varianza o correlación.
  - Selección por importancia de variables.
  - Autoencoders (avanzado).
- **Criterios:** Interpretabilidad, estabilidad, evitar colinealidad.
- **Resultado:** Dataset compacto, más eficiente para minería.



Fase de Minería de Datos (DM)

- **Objetivo:** Aplicar algoritmos para descubrir patrones o modelos útiles.
- **Tareas Analíticas posibles:**
  - **Clasificación:** Predecir clases (e.g., cliente leal o no).
  - **Regresión:** Estimar tendencias, relaciones o causalidad sobre valores (e.g., ingresos).
  - **Clustering:** Agrupar observaciones sin etiquetas.
  - **Asociación:** Encontrar reglas frecuentes (e.g., Si A, entonces B).
  - **Anomalías:** Detectar casos inusuales.
- **Tipo de enfoque:** Supervisado (clasificación, regresión) vs. No supervisado (clustering, asociación, anomalías).

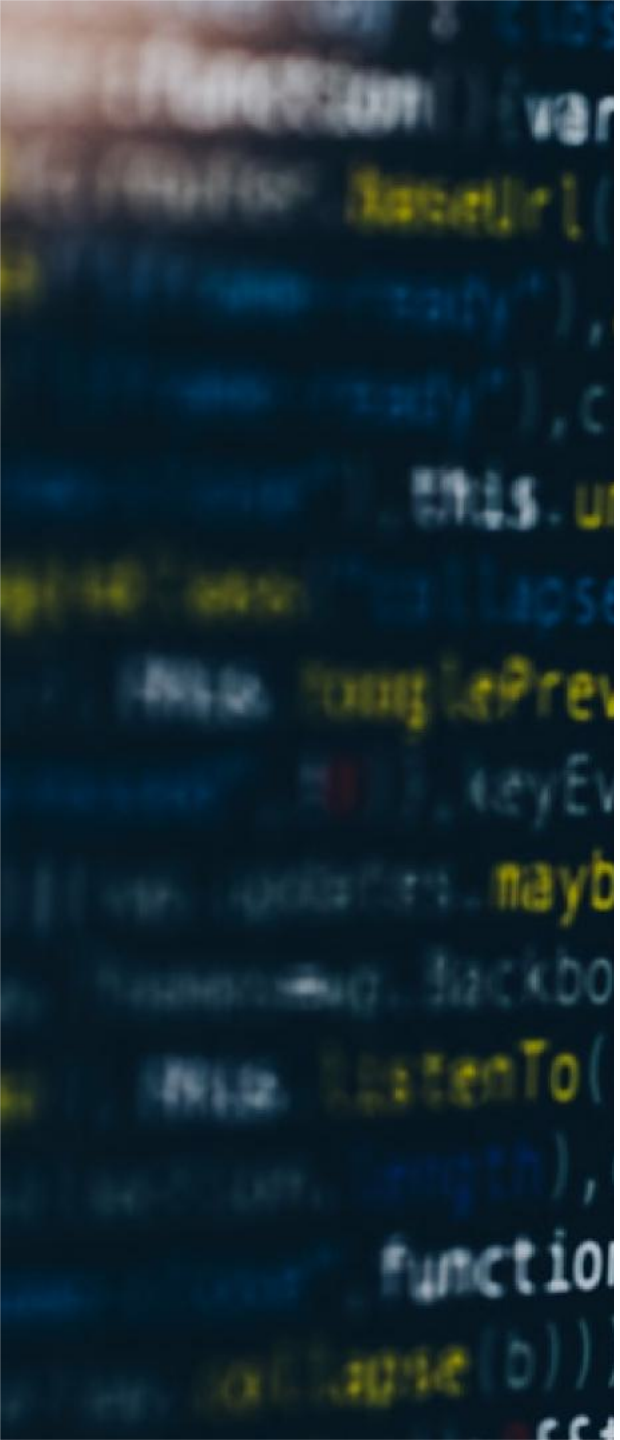
Fase de Minería de Datos (DM)

- **Técnicas Comunes por tipo de tarea:**
  - **Clasificación/Regresión:** Árboles de decisión, KNN, Regresión Lineal/Logística, SVM.
  - **Clustering:** K-Means, DBSCAN, Aglomerativo jerárquico.
  - **Asociación:** Apriori, FP-Growth.
  - **Anomalías:** Z-score, Isolation Forest, Local Outlier Factor.
- **Resultado esperado:** Modelos, patrones o segmentos candidatos a evaluación.

Importancia del Loop EDA y Preprocesamiento/Transformación

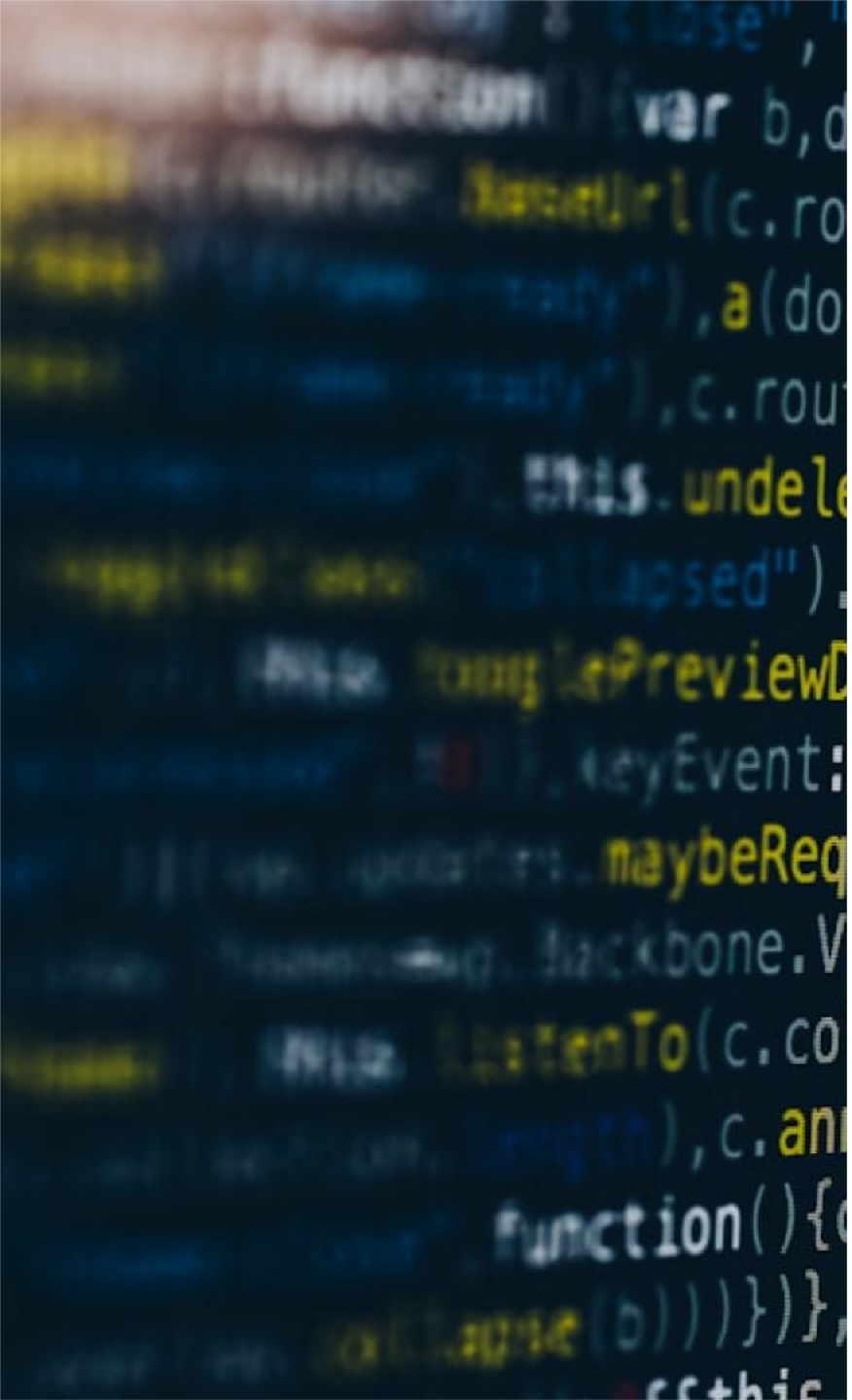
- La **exploración de datos (EDA)** no es una fase que finaliza luego de su implementación:
  actúa como una **retroalimentación continua** durante:
  - Limpieza y validación de datos.
  - Selección y creación de variables.
  - Detección de errores sistemáticos.
  - Evaluación del impacto de las transformaciones.
- Recomendación:
  Alternar entre **vista de datos** y **acciones técnicas** para evitar sobreajuste, redundancia o pérdida de información útil.

USAR FUNCIONES!



Fase de Evaluación e Interpretación

- **Objetivo:** Validar rigurosamente los resultados de la minería, interpretar su significado en el contexto del problema y decidir sobre su utilidad práctica. ¡Fase crítica!
- **Dimensiones de Evaluación:**
  - **Validez Estadística/Predictiva:** ¿Qué tan bien funciona el modelo en datos no vistos? (Precisión, Recall, F1, AUC, MSE, R^2, etc.).
  - **Estabilidad/Robustez:** ¿Es sensible a pequeños cambios en los datos?
  - **Interpretabilidad:** ¿Se entiende cómo/por qué el modelo toma decisiones? (Crucial para confianza y regulación).
  - **Novedad:** ¿El patrón descubierto es trivial o ya conocido?
  - **Utilidad/Valor Práctico:** ¿Es accionable? ¿Aporta valor real al negocio/objetivo?



Fase de Evaluación e Interpretación

- **Técnicas de Validación:**
  - Separación Train/Validation/Test (Hold-out).
  - Validación Cruzada (K-Fold, LOOCV).
  - Matrices de Confusión, Curvas ROC/PR.
  - Comparación con modelos Baseline o existentes.
  - Análisis de residuos (regresión).
  - Validación por expertos del dominio.
- **Resultado:** Juicio crítico sobre los modelos/patrones generados. Decisión informada sobre despliegue, refinamiento o descarte.

Seguimos en la próxima!

Bibliografía de Referencia

- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27–34.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
- Tan, P. N., Steinbach, M., Karpatne, A. & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.