ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

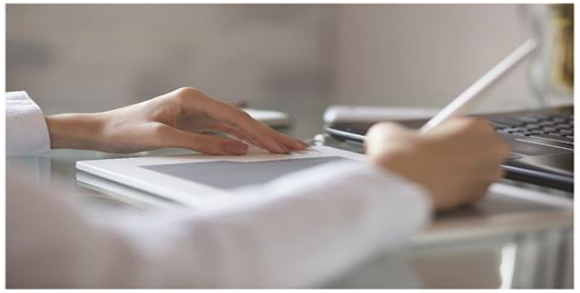
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Introducción

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




¿Qué es la Minería de Datos?

"El proceso que usa estadística, matemática, inteligencia artificial y técnicas de *machine learning* para extraer e identificar información útil y subsecuentemente adquirir conocimiento a partir de grandes bases de datos"

Han, J., Kamber, M., & Pei, J. (2012)

Data Mining Concepts and Techniques

Morgan Kaufmann

¿Qué es la Minería de Datos?

Área de investigación que pertenece a la Ciencia de la Computación y Estadística

… puede que a la Física y Matemática también

La Minería de Datos estudia cómo desarrollar herramientas automáticas para analizar datos

¿Qué es la Minería de Datos?

Las soluciones deben ser:

Lo suficientemente rápidas

(Big Data)

Lo suficientemente precisas

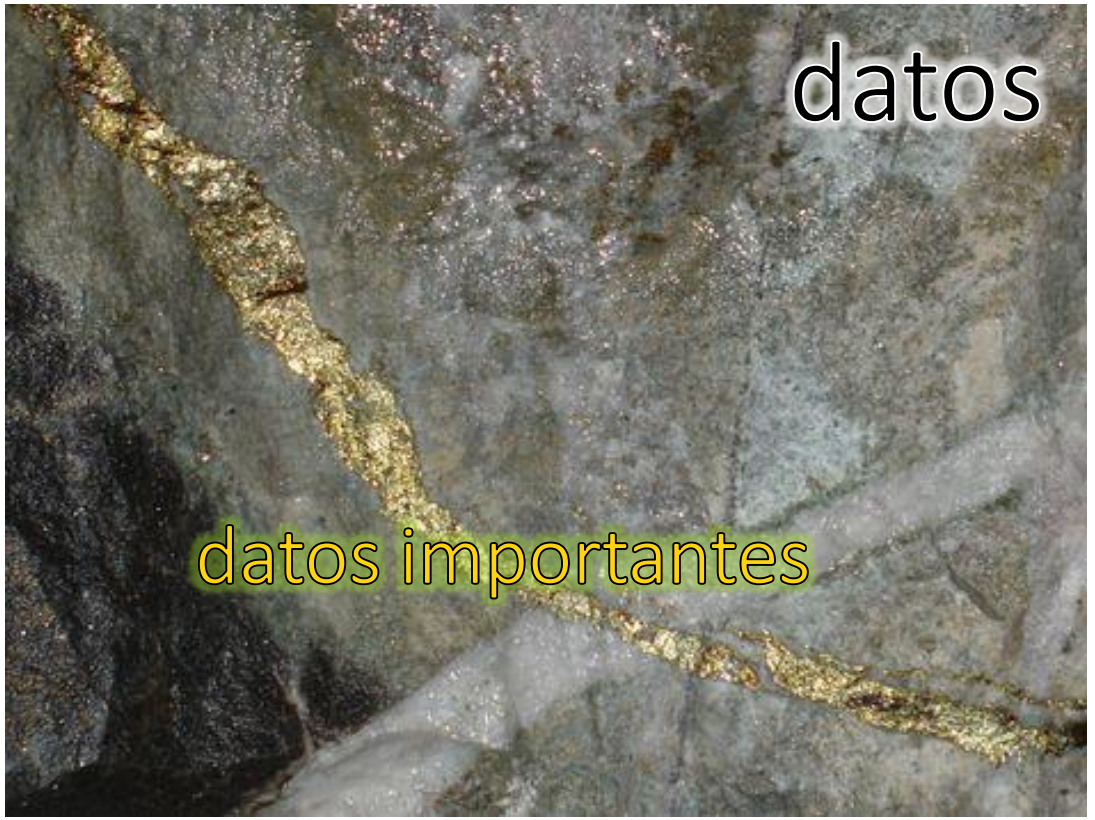
(modelos flexibles y adaptivos)

¿Qué es la Minería de Datos?

Minería para extracción de oro

versus

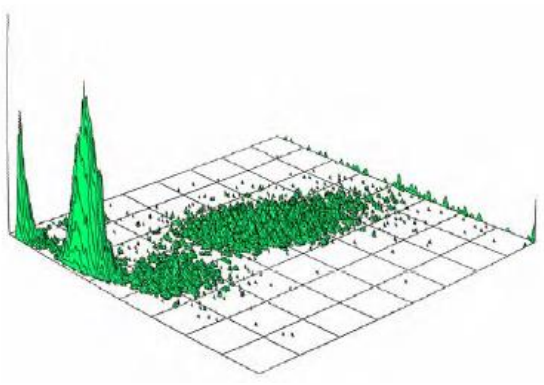
Minería para la extracción de información



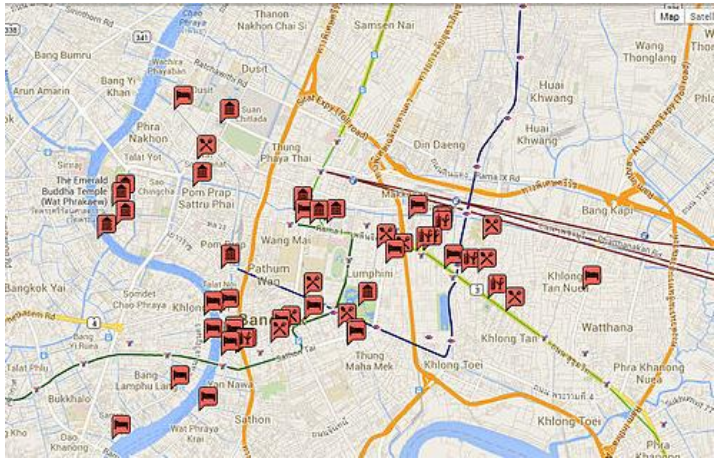
¿Por qué aplicar Minería de Datos?

Ciertas bases de datos no pueden ser analizadas manualmente

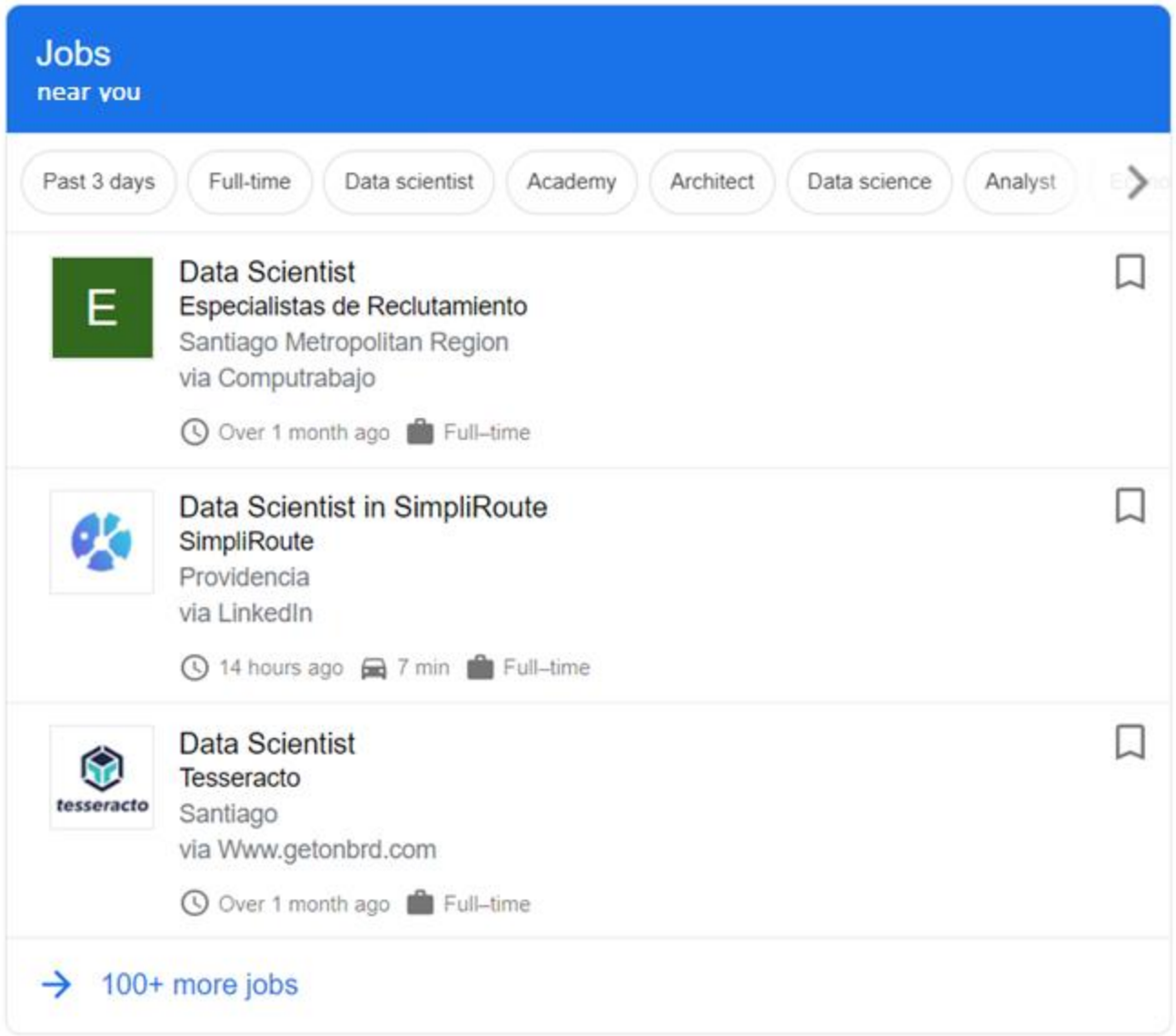
Se necesitan herramientas complementarias para revisar datos en diferentes partes del proceso de análisis



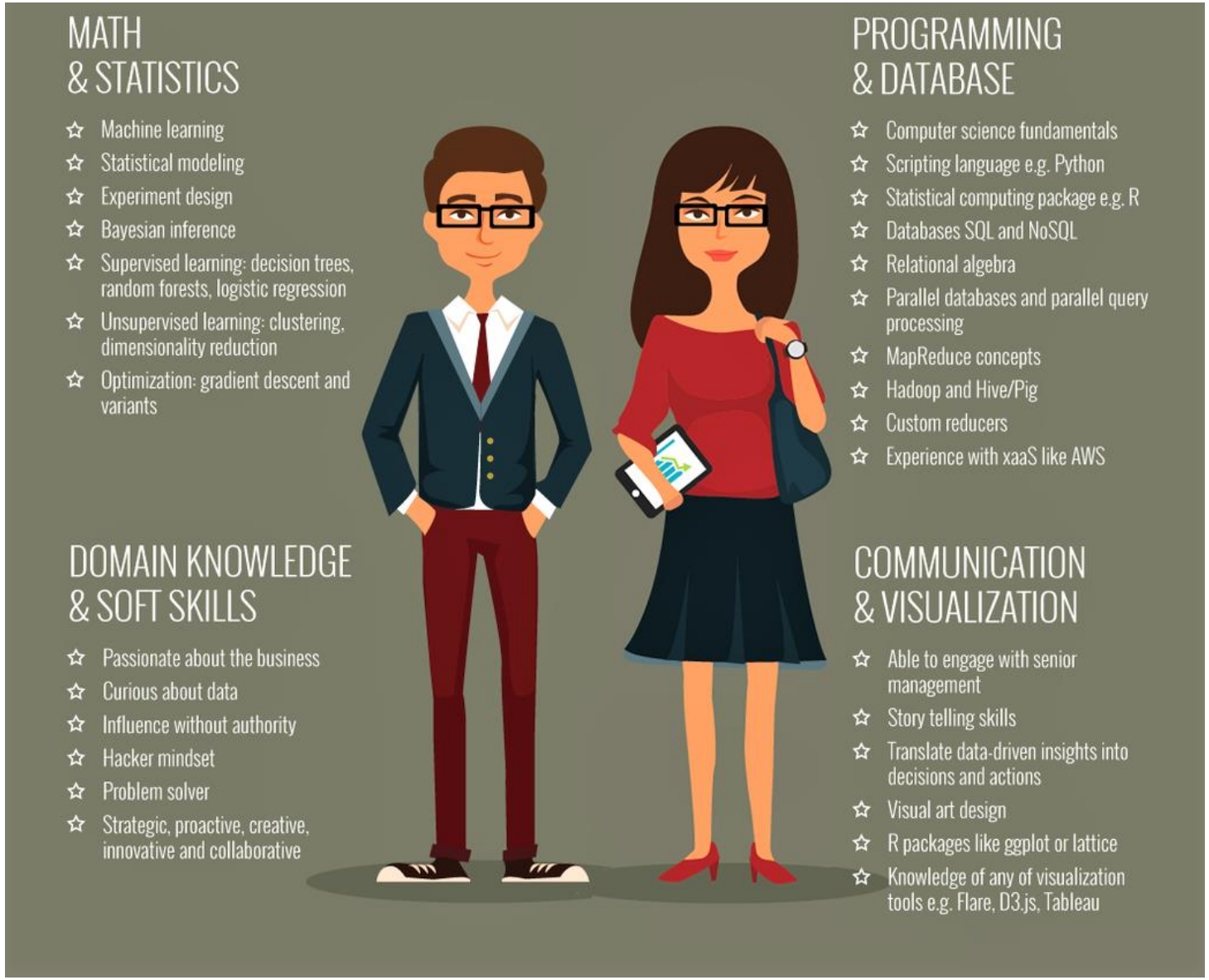
¿Por qué aplicar Minería de Datos?



¿Por qué aplicar Minería de Datos?

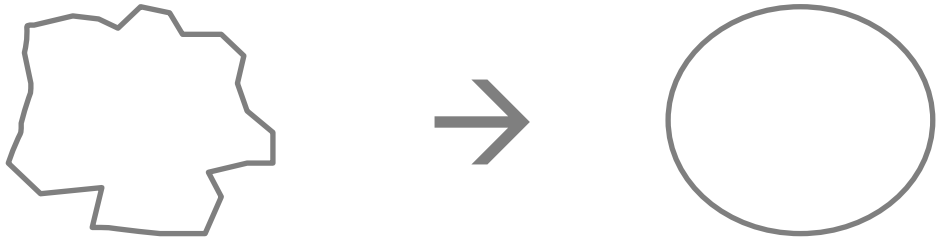


El *data scientist* moderno



¿Qué es un modelo?

Representación simplificada de parte del mundo real



Foco en los elementos más importantes

Relaciones causales entre variables

¿En qué consiste la Minería de Datos?

CRISP-DM

Cross-industry standard process for data mining

1. Entender el negocio
2. Entender los datos
3. Preparar los datos
4. Modelar
5. Evaluar
6. Desplegar

CRISP-DM

1. **Entender el negocio**

    Definir metas e indicadores de éxito

    Evaluar la situación actual

    Definir objetivos de minería de datos

    Definir el plan de acción

CRISP-DM

2. Entender los datos

Recolectar los datos

Describir los datos

Explorar los datos

Verificar la calidad y completitud de los datos

CRISP-DM

3. Preparar los datos

Seleccionar los datos a utilizar

Limpiar los datos

Procesar y construir nuevos datos

Integrar los datos

CRISP-DM

4. Modelar

Seleccionar la técnica de modelación

Realizar pruebas

Construir el modelo

Calibrar y entrenar el modelo

Validar el modelo

CRISP-DM

5. **Evaluar**

   Analizar los resultados

   Revisar el proceso

   Determinar los siguientes pasos

CRISP-DM

6. Desplegar

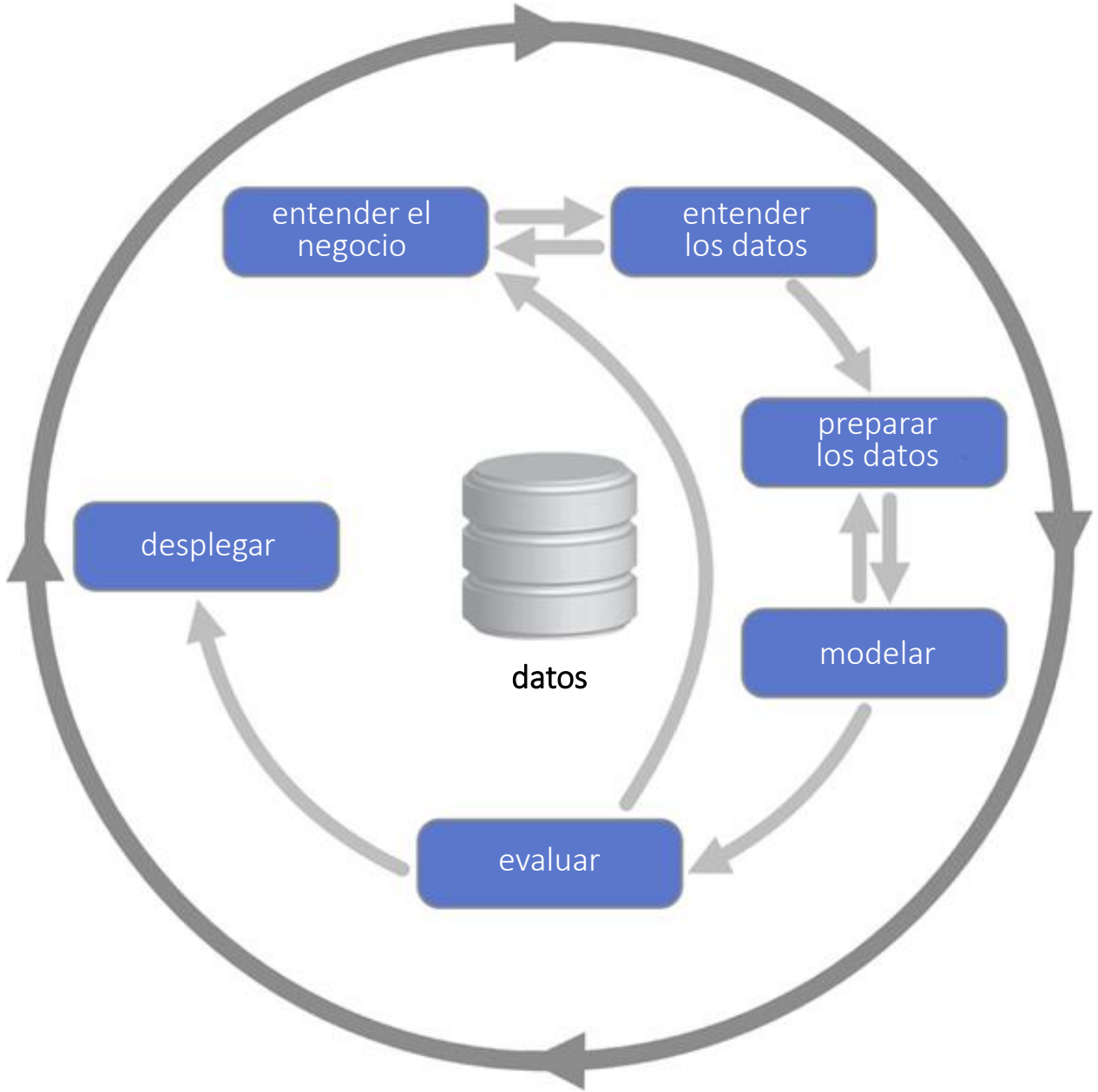
Determinar el plan de acción

Monitorear y mantener

Evaluar la implementación

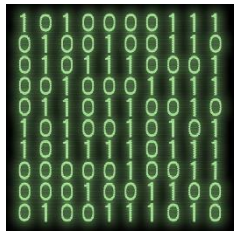
Identificar aprendizajes

CRISP-DM


entender el negocio
entender los datos
preparar los datos
modelar
evaluar
desplegar
datos

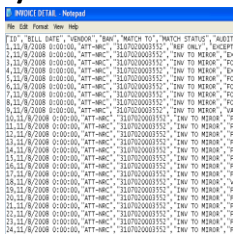
Ruta de aprendizaje

Bases de Datos

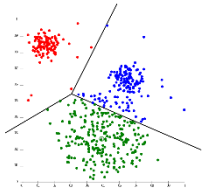


Integración de Datos

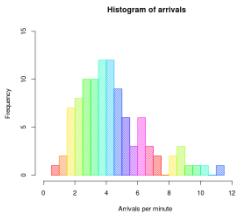
Transformación y Selección



Minería de Datos



Evaluación y Visualización



Conocimiento

¿Qué aprenderán en este curso?

Técnicas de Minería de Datos

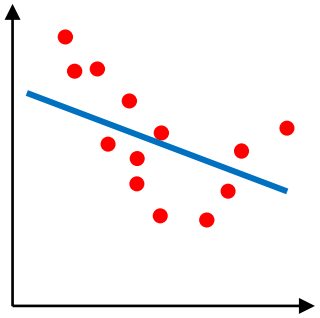
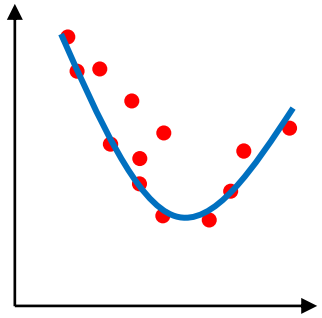
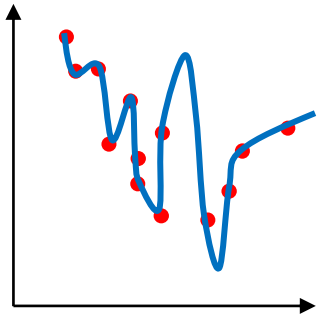
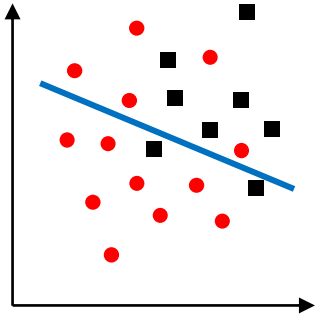
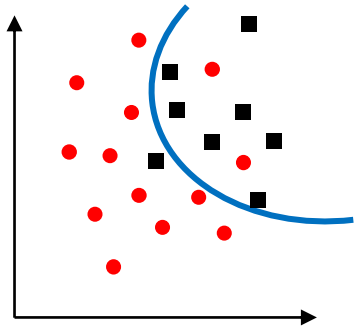
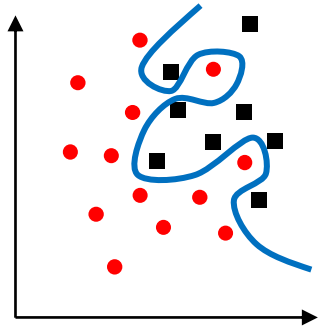
Técnicas Predictivas – Aprendizaje Supervisado

Regresión	ajustar variables/relaciones continuas
Clasificación	ajustar variables/relaciones discretas

Técnicas Descriptivas – Aprendizaje No Supervisado

Clustering	agrupar datos similares
Asociación	identificar patrones y coocurrencias

Técnicas de Minería de Datos

	bajo ajuste	ajuste óptimo	sobreajuste
Regresión			
Clasificación			

Contenidos

Clase 1 – Introducción a la Minería de Datos

Clase 2 – Preprocesamiento y análisis de datos

Clase 3 – Análisis de regresión

Clase 4 – Árboles de decisión

Contenidos

Clase 5 – Métodos de clasificación

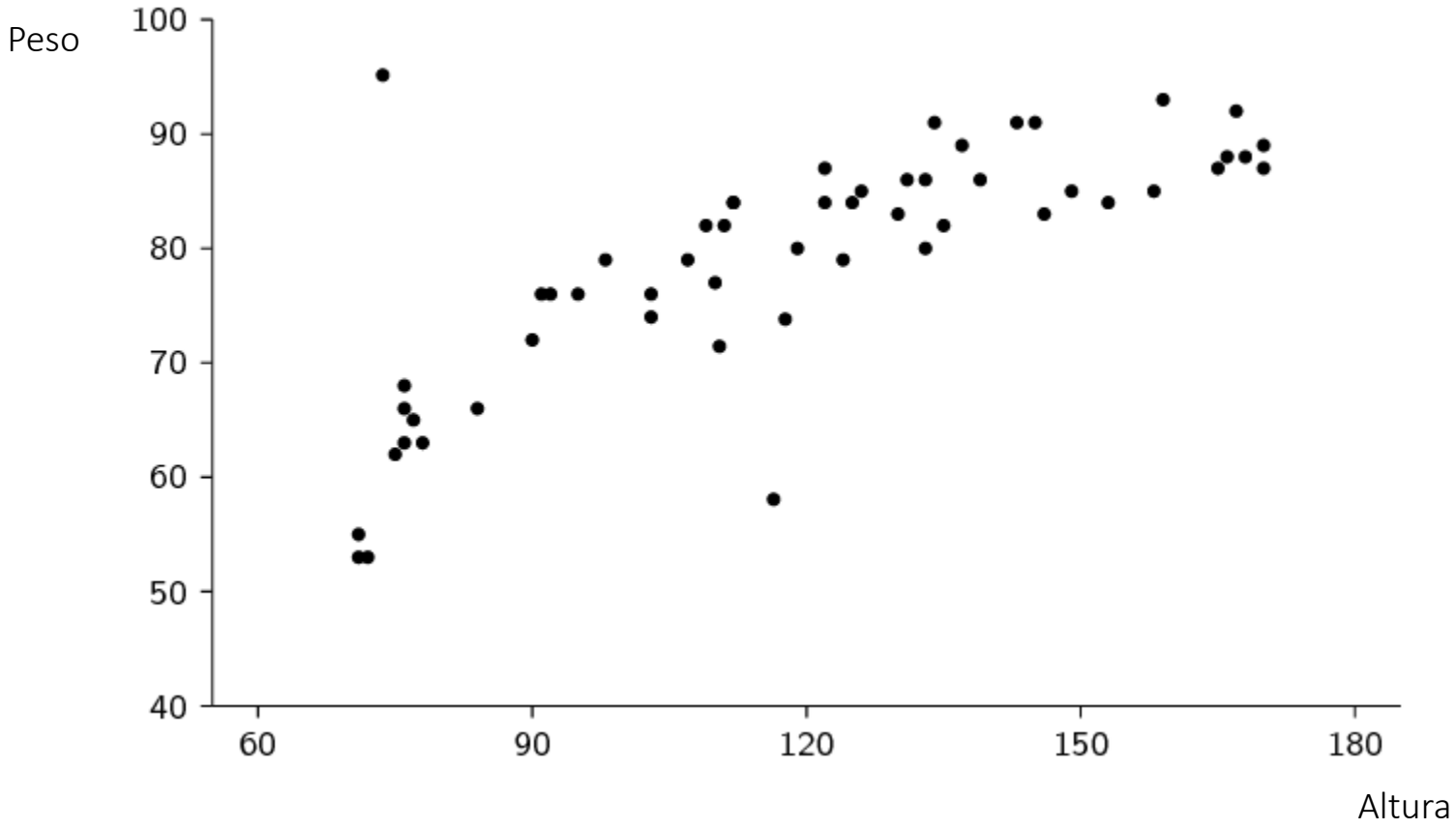
Clase 6 – Análisis de clústeres

Clase 7 – Reglas de asociación

Clase 8 – Selección de modelos

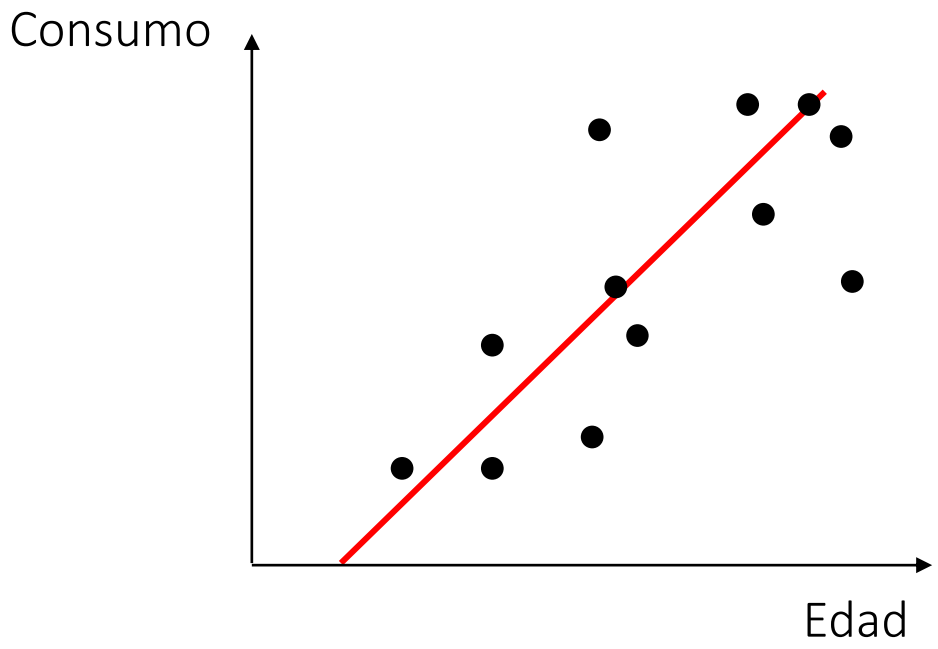
Preprocesamiento y análisis de datos

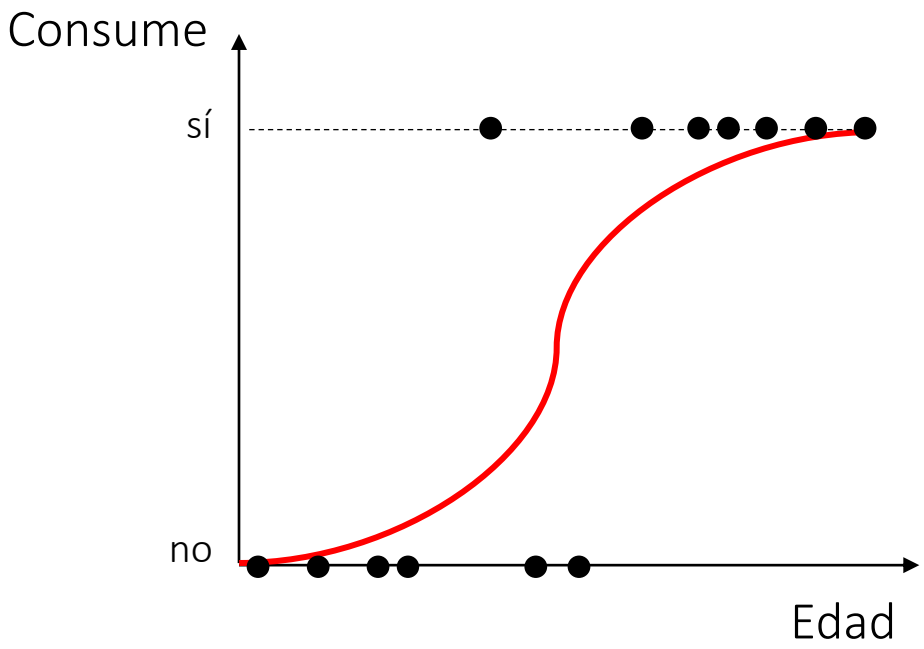
Todo modelo necesita buenos datos



Análisis de regresión

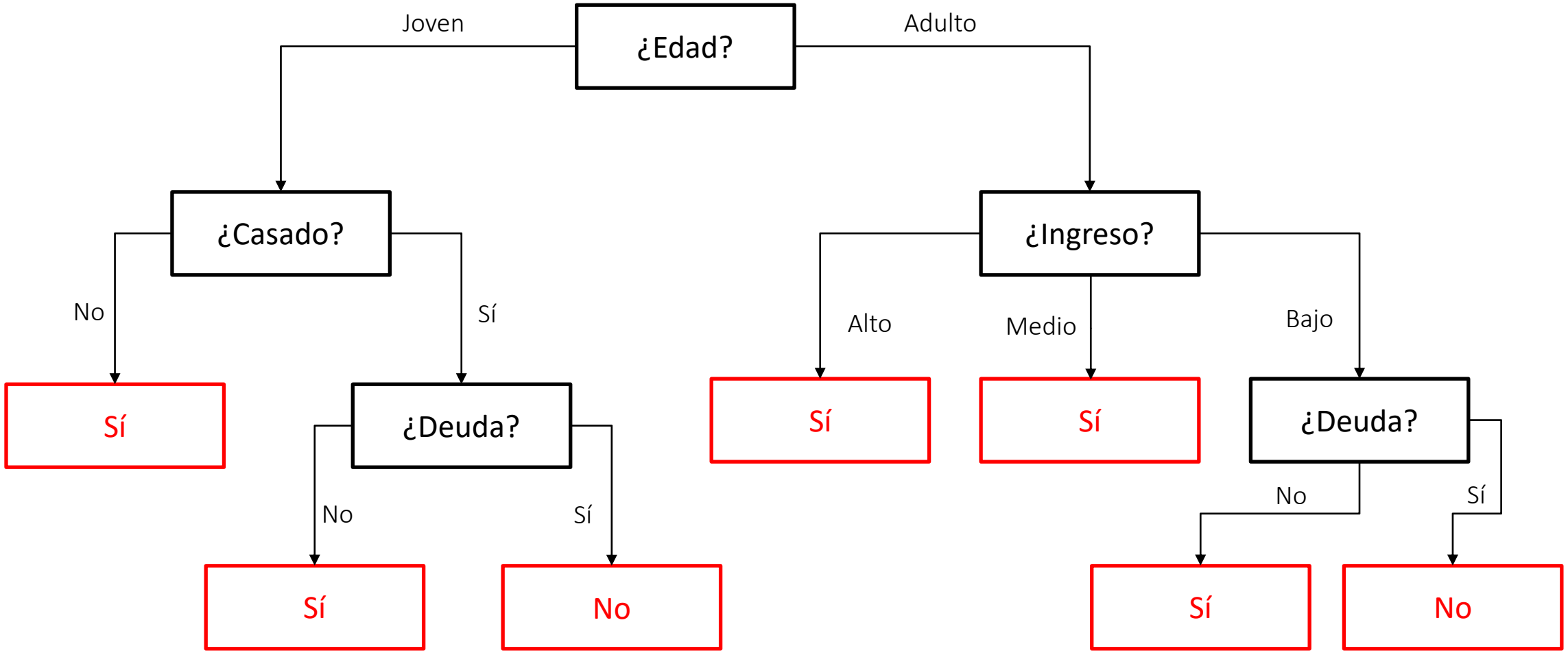
Supongamos que queremos predecir el consumo de cierto producto





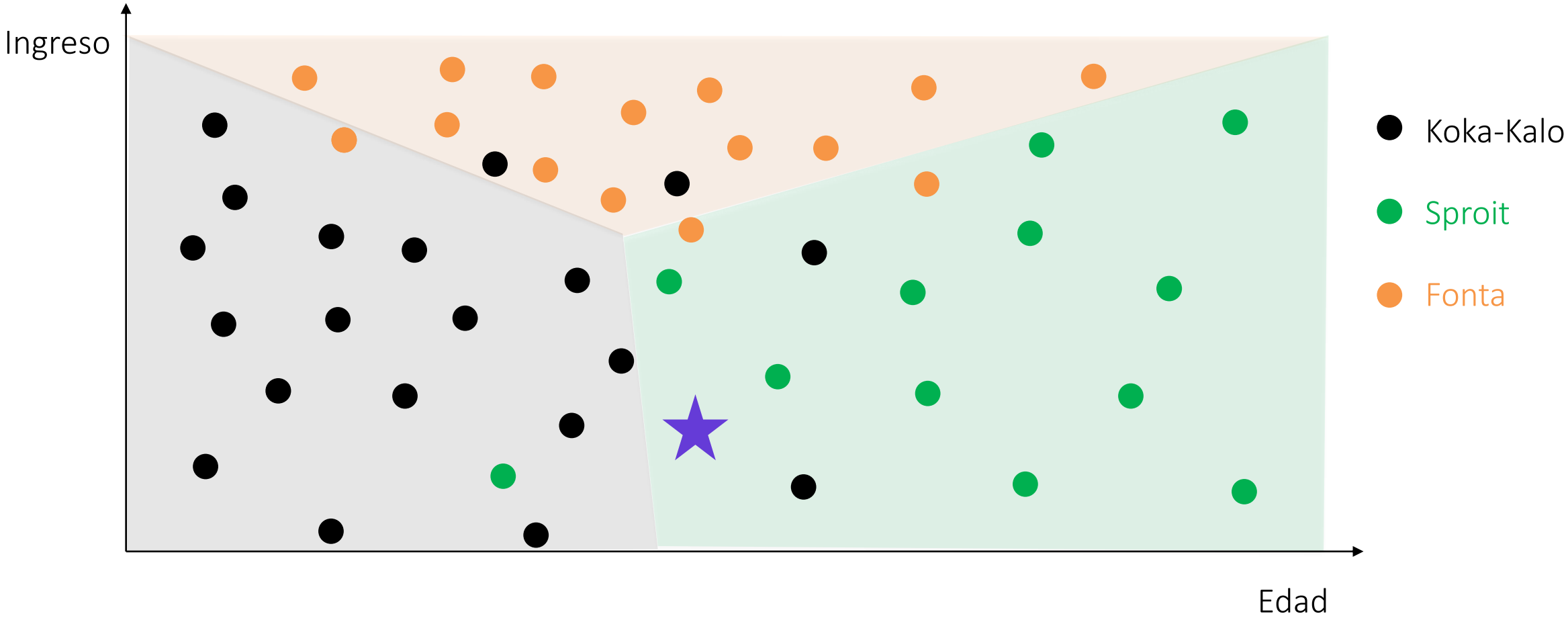
Árboles de decisión

Supongamos que queremos predecir si una persona comprará cierto producto



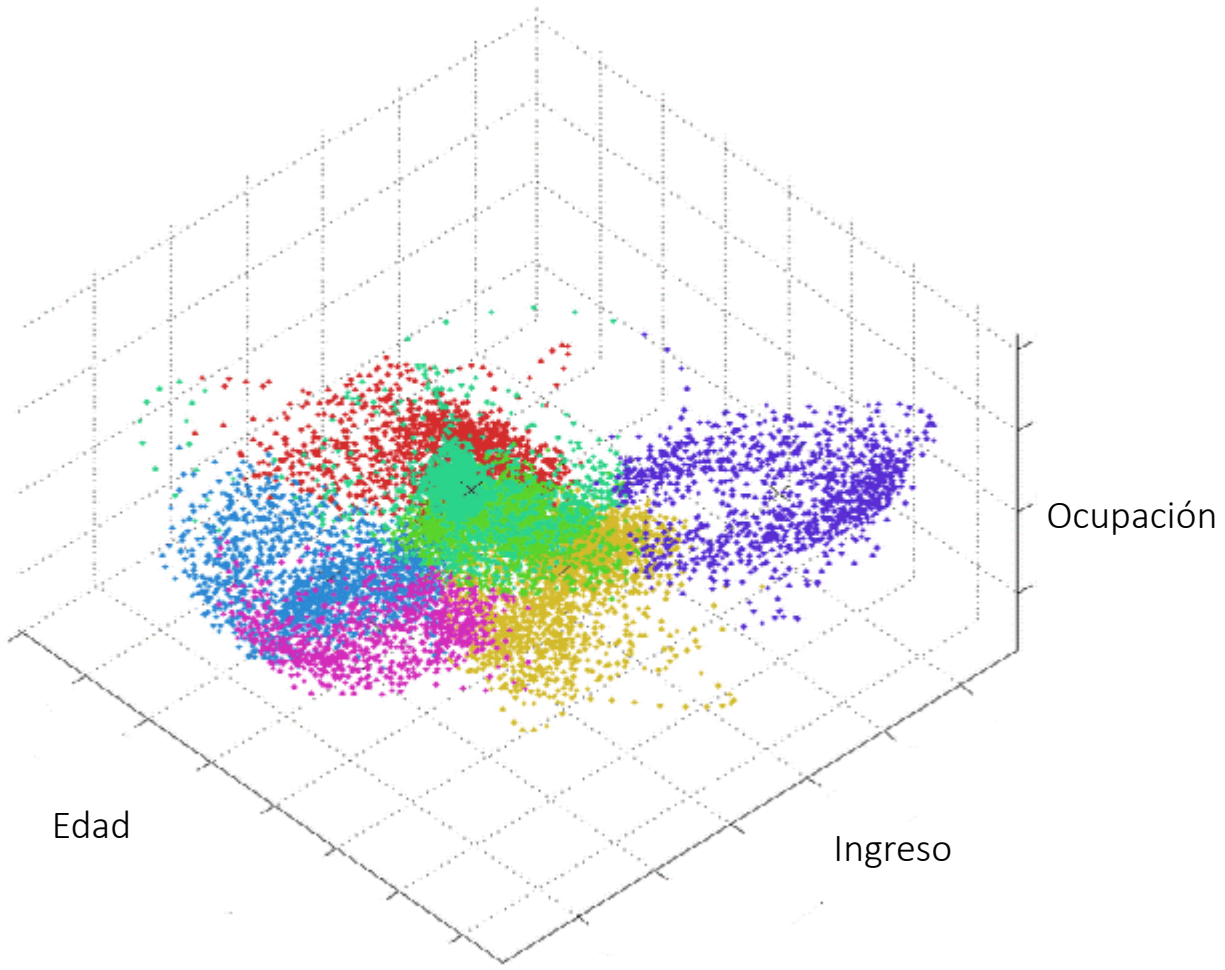
Métodos de clasificación

Supongamos que queremos predecir la elección entre ciertos productos



Análisis de clústeres

Supongamos que queremos identificar perfiles de clientes



Arriesgados

Hipsters

Cautos

Aventureros

Despreocupados

Racionales

Reglas de asociación

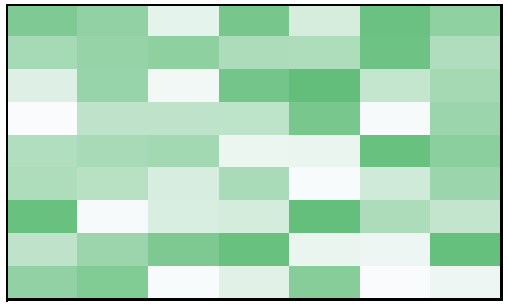
Supongamos que queremos realizar promociones de productos

A + B

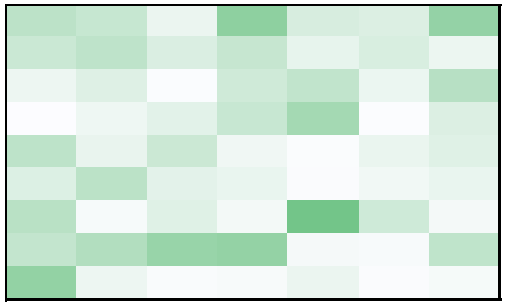
F ↔ B

Selección de modelos

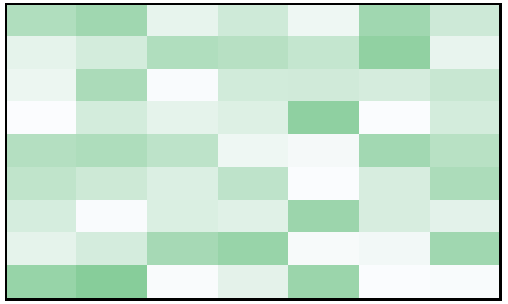
Todo modelo debe ser validado y evaluado



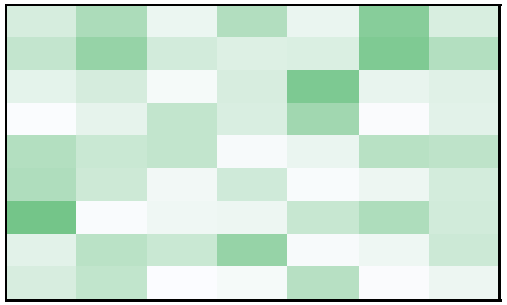
Realidad



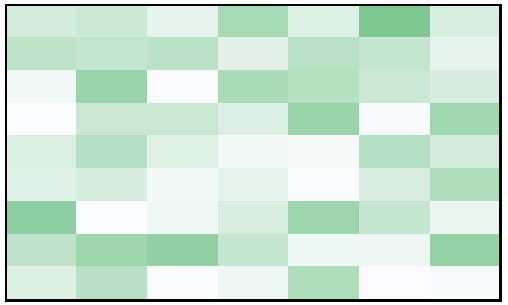
Modelo 1



Modelo 2



Modelo 3



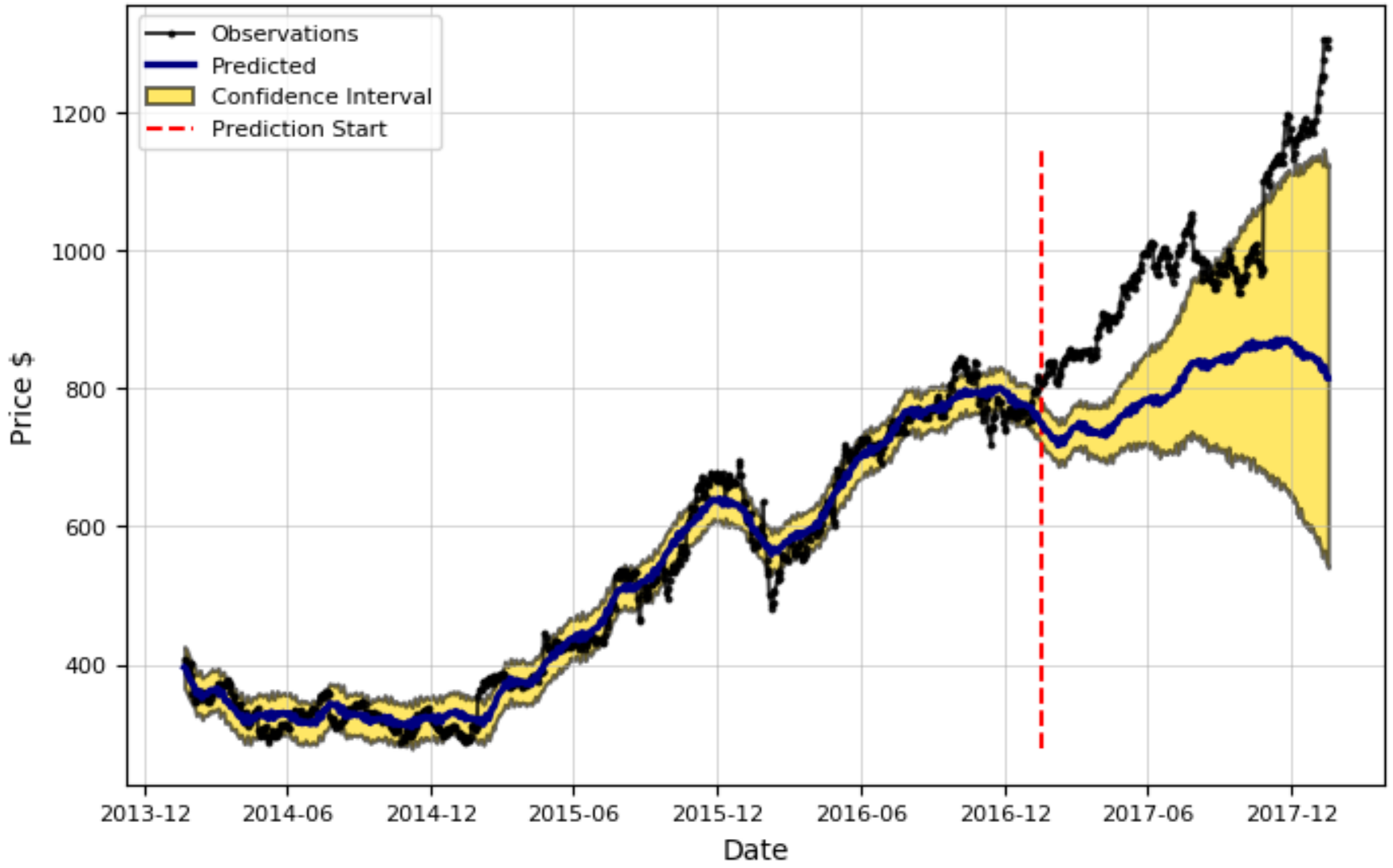
Modelo 4

Veamos algunas aplicaciones populares

Sistemas recomendadores

Descuentos personalizados

Pronóstico de acciones



Marketing

Perfiles de clientes

Fuga de clientes

Planificación de sistemas

Medicina

Clasificación de imágenes

gato

gato

gato

no gato

gato

no gato

Minería de texto