



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

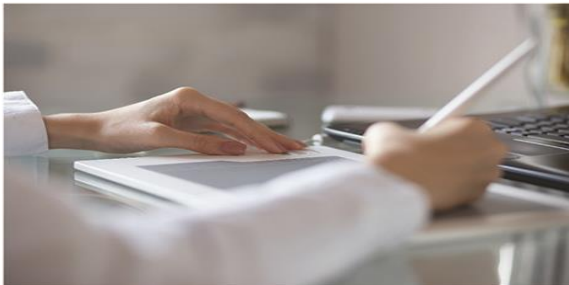
EDUCACIÓN
PROFESIONAL

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

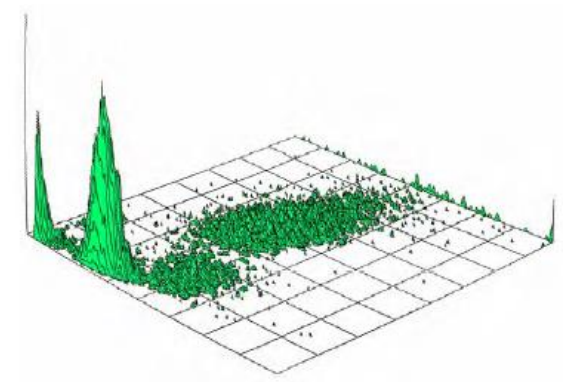
Minería para la extracción de información



¿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?

Jobs


Near Las Condes


Past 3 days Full-time Data scientist Academy Architect Data science Analyst Econo



Data Scientist
Especialistas de Reclutamiento
Santiago Metropolitan Region
via Computrabajo


🕒 Over 1 month ago 🧳 Full-time






Data Scientist in SimpliRoute
SimpliRoute
Providencia
via LinkedIn


🕒 14 hours ago 🚗 7 min 🧳 Full-time





Data Scientist
Tesseracto
Santiago
via Www.getonbrd.com

🕒 Over 1 month ago 🧳 Full-time



[➔ 100+ more jobs](#)

El *data scientist* moderno

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



PROGRAMMING & DATABASE

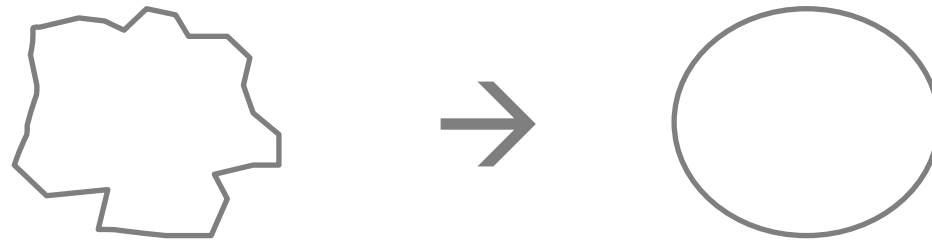
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

¿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

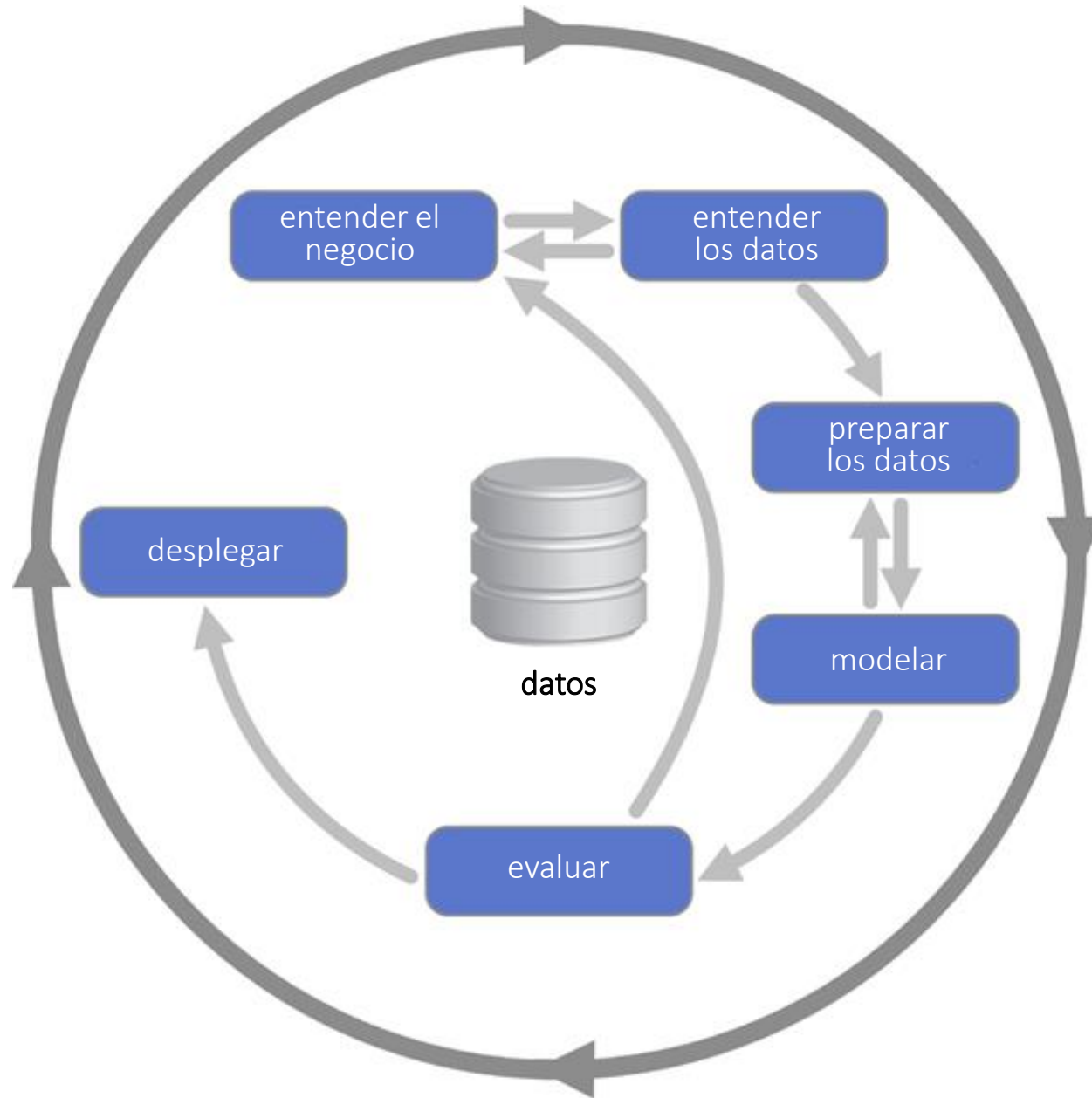
Determinar el plan de acción

Monitorear y mantener

Evaluar la implementación

Identificar aprendizajes

CRISP-DM



Ruta de aprendizaje

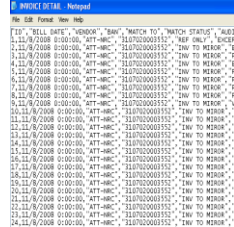
Bases de Datos



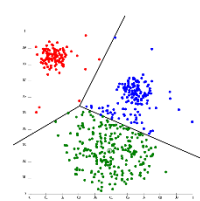
Integración de
Datos



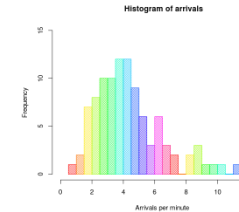
Transformación
y Selección



Minería de
Datos



Evaluación y
Visualización



Conocimiento



¿Qué aprenderán en este curso?

Contenidos

Clase 1 – Introducción a la Minería de Datos

Clase 2 – Preprocesamiento y análisis de datos

Clase 3 – Árboles de decisión

Clase 4 – Reglas de asociación

Contenidos

Clase 5 – Métodos de clasificación

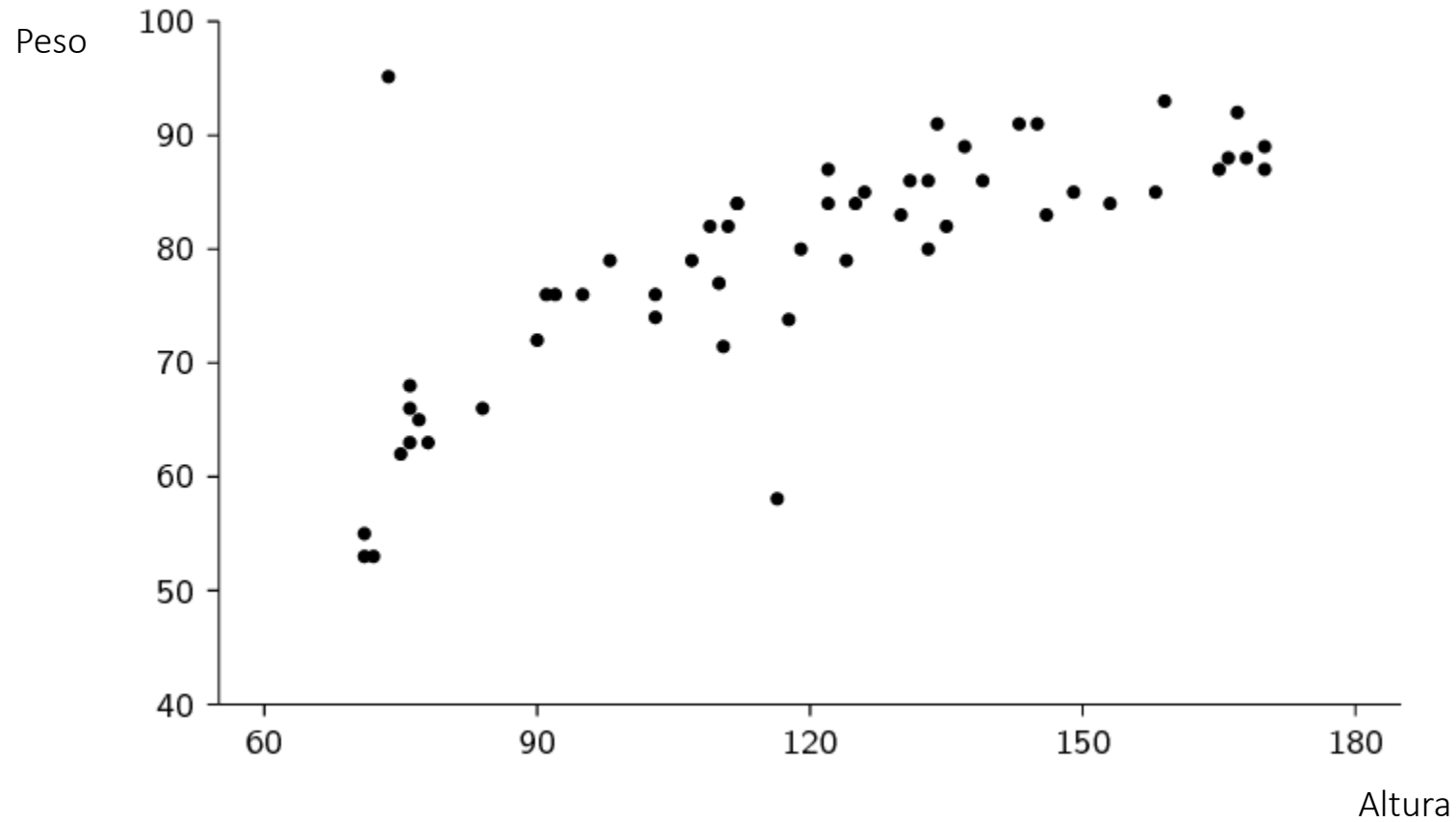
Clase 6 – Análisis de clústeres

Clase 7 – Análisis de regresión

Clase 8 – Selección de modelos

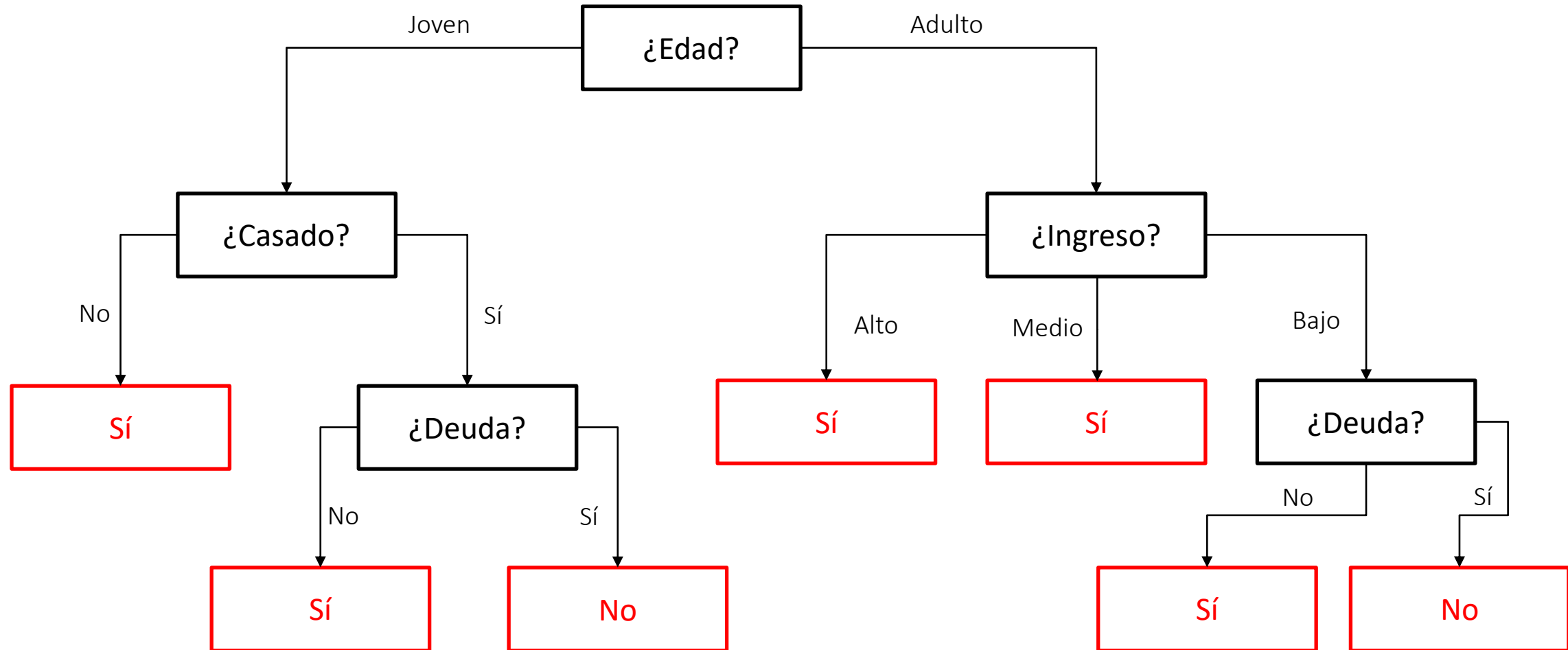
Preprocesamiento y análisis de datos

Todo modelo necesita buenos datos



Árboles de decisión

Supongamos que queremos predecir si una persona comprará cierto producto



Reglas de asociación

Supongamos que queremos realizar promociones de productos

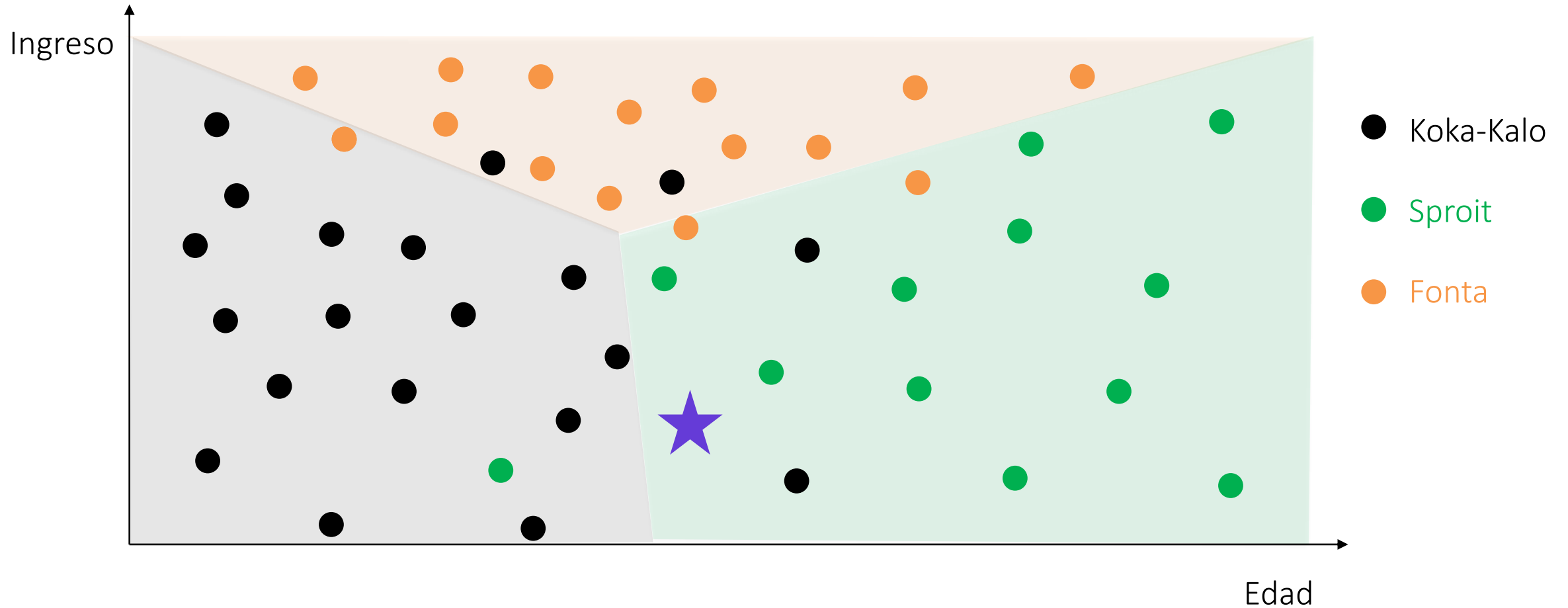


$$\boxed{A} + \boxed{B}$$

$$\boxed{F} \leftrightarrow \boxed{B}$$

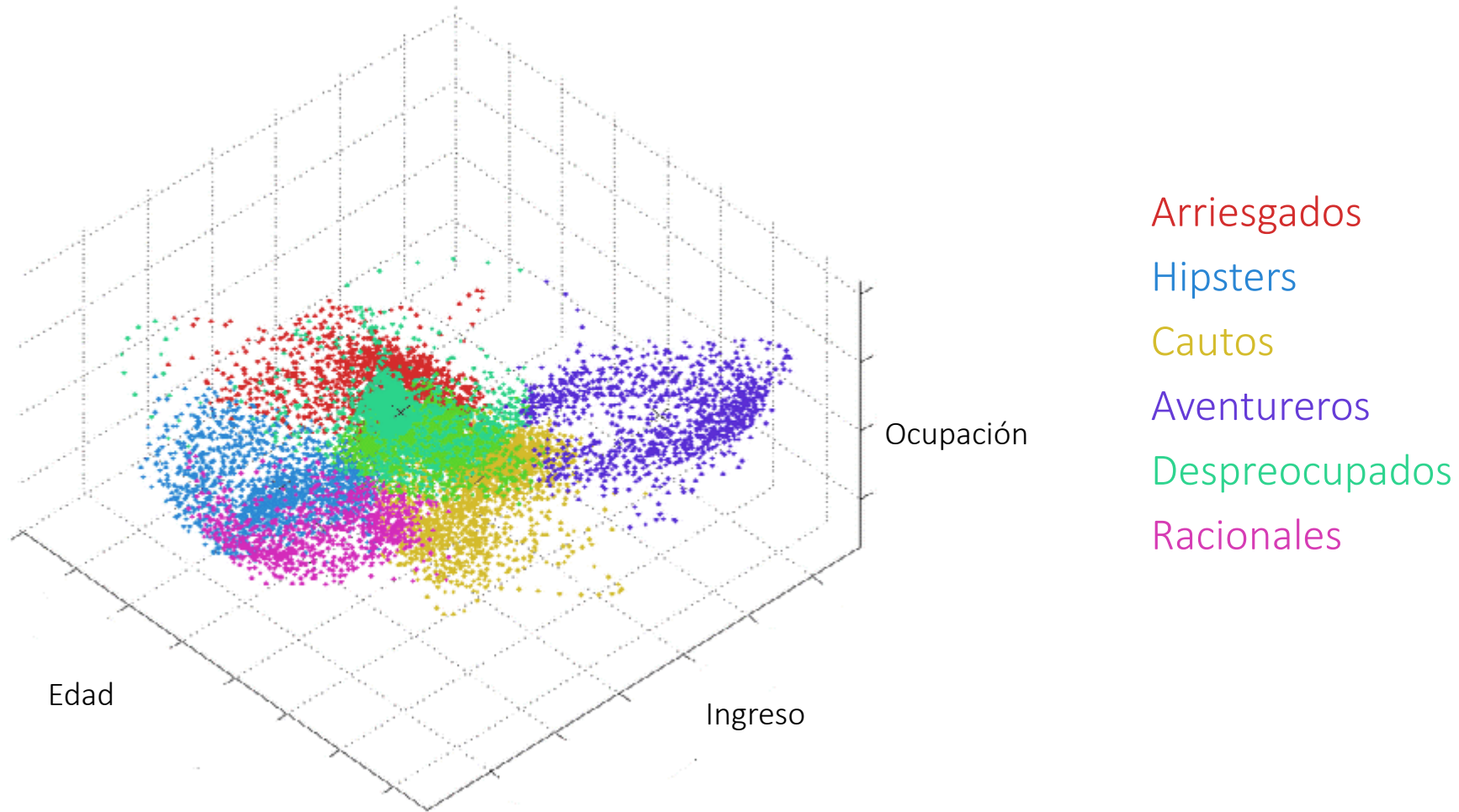
Métodos de clasificación

Supongamos que queremos predecir la elección entre ciertos productos



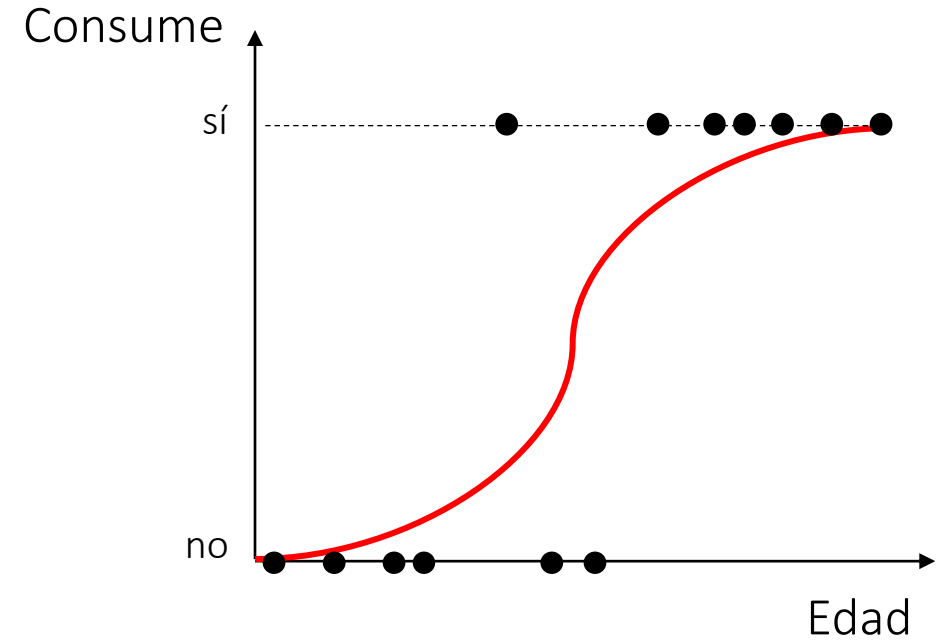
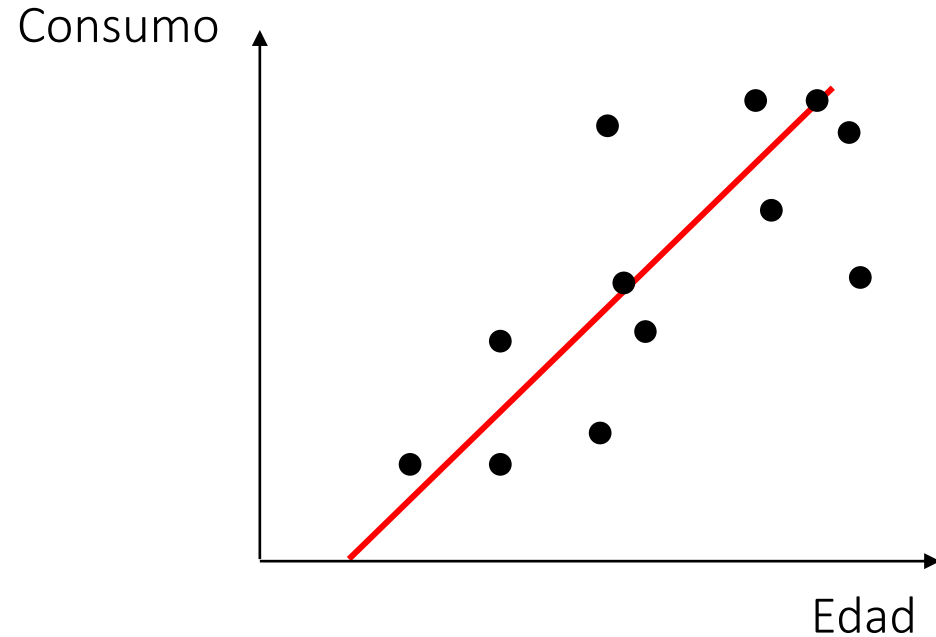
Análisis de clústeres

Supongamos que queremos identificar perfiles de clientes



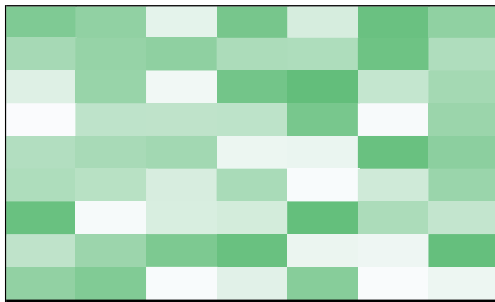
Análisis de regresión

Supongamos que queremos predecir el consumo de cierto producto

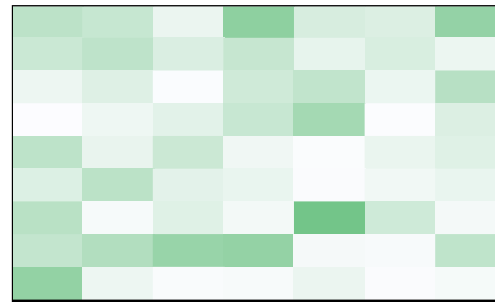


Selección de modelos

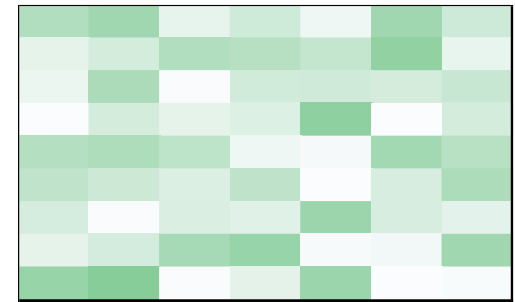
Todo modelo debe ser validado y evaluado



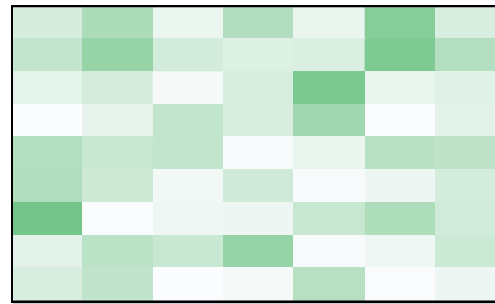
Realidad



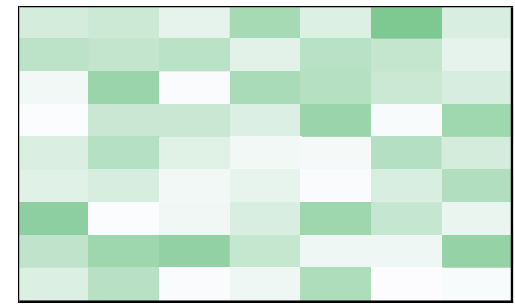
Modelo 1



Modelo 2



Modelo 3



Modelo 4

Veamos algunas aplicaciones populares

Sistemas recomendadores



Descuentos personalizados

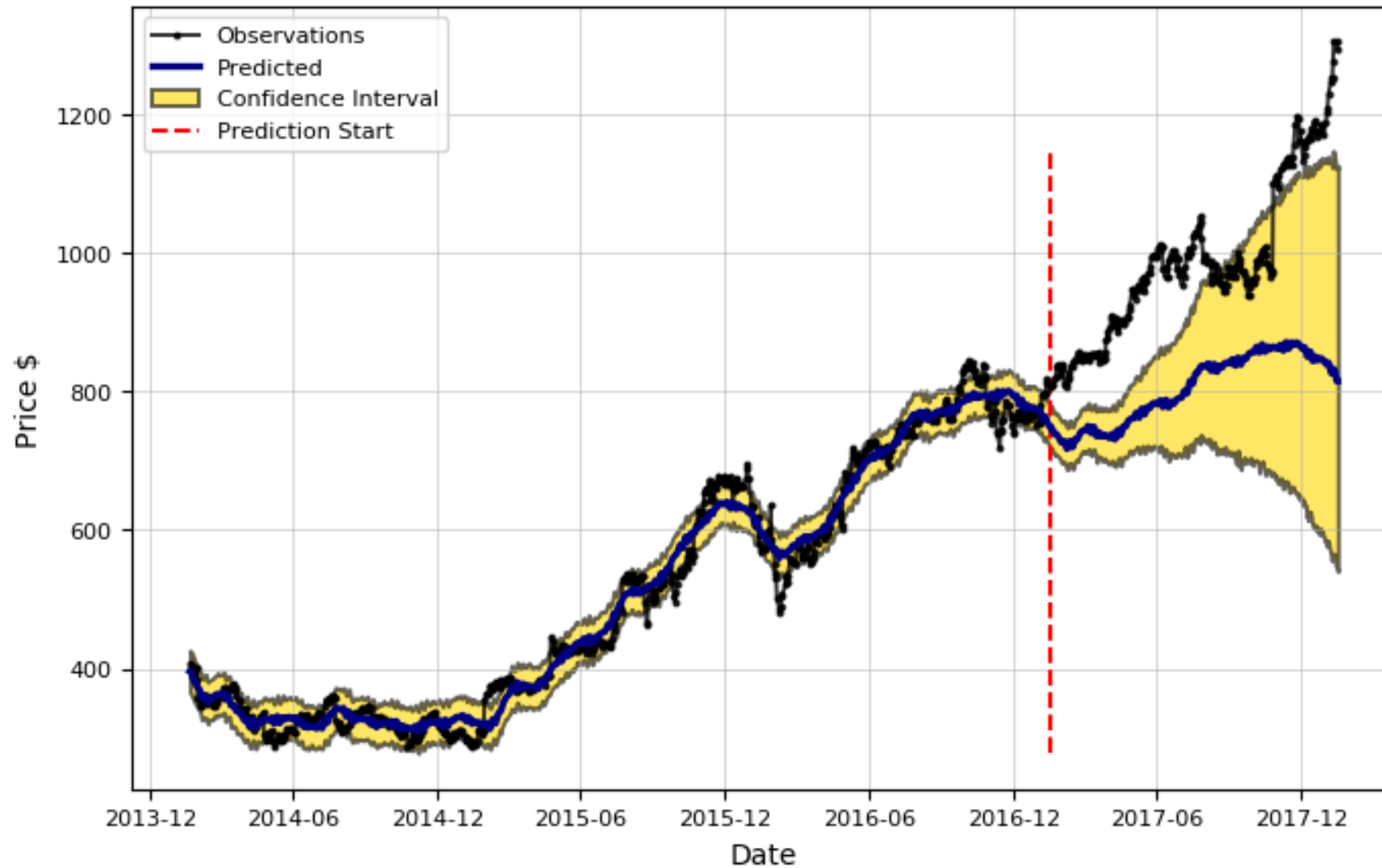
BUENAS NOTICIAS JUMBO

Productos Importados
te sorprende con

**30% DESCUENTO EN
BARRA DE CEREAL MARCA
CORN, CORNY FREE,
ESSENTIA
Y SACCIALIS.
CON TODO MEDIO DE PAGO**

Dicta tu RUT en caja
Oferta exclusiva para ti
Válido hasta el 09/09/19

Pronóstico de acciones



Marketing



Perfiles de clientes



Fuga de clientes



Planificación de sistemas



Medicina



Clasificación de imágenes



gato



gato



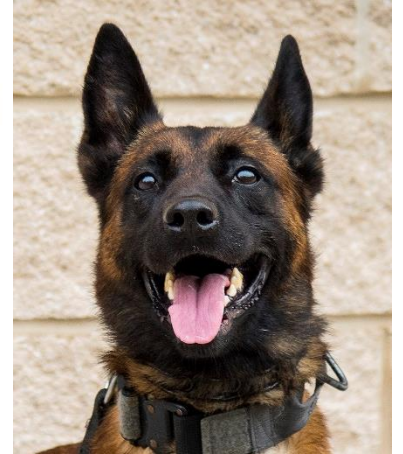
gato



no gato



gato



no gato

Minería de texto

