Big Data y Visualización
Aprendizaje de Máquina

Universidad Autónoma de Manizales

Docente: Alejandro Mora Rubio

Contenido

¿Qué es el aprendizaje de máquina?

Importancia de los datos

¿Por qué usarlo?

Tipos de sistemas

Tareas básicas

Etapas de desarrollo

Principales retos

Recursos

¿Qué es el aprendizaje de máquina?

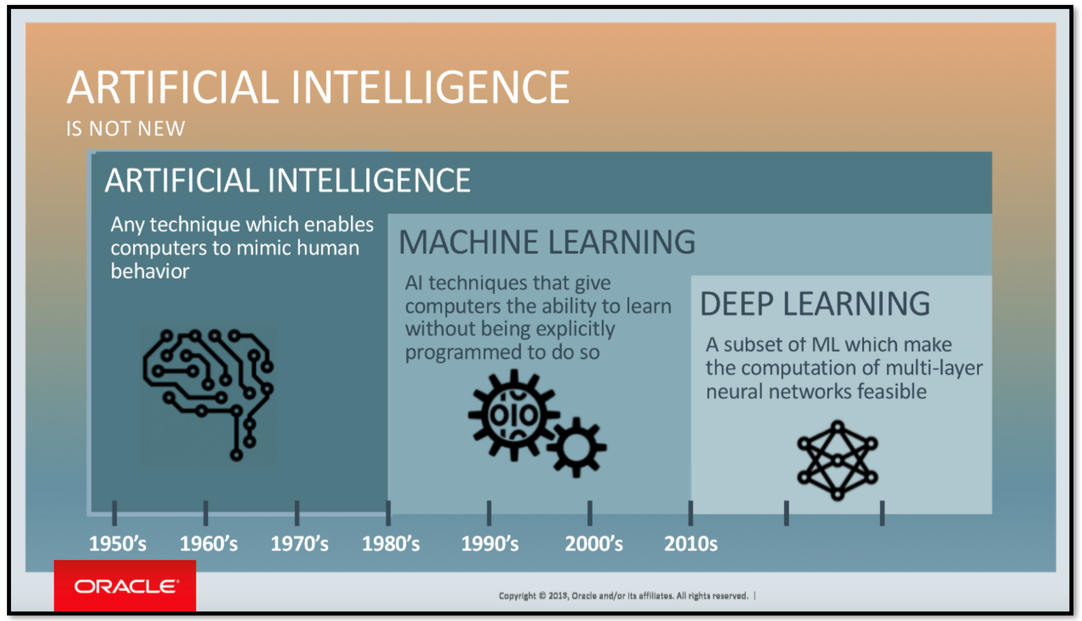


Imagen de Oracle

El aprendizaje de máquina o *Machine Learning*, es un **área de la inteligencia artificial que permite a los computadores aprender a partir de datos.**

Entendiendo inteligencia artificial como el conjunto de técnicas que permiten a los computadores imitar la inteligencia o el comportamiento humano.

Basadas principalmente en la matemática, estadística y programación.

Importancia de los datos

Imagen de Farolbi

Información sobre un hecho o suceso:

- Audio
- Texto
- Imágenes
- Medición de variables físicas

Los datos constituyen la base de los sistemas actuales de inteligencia artificial.

¿Por qué usarlo?

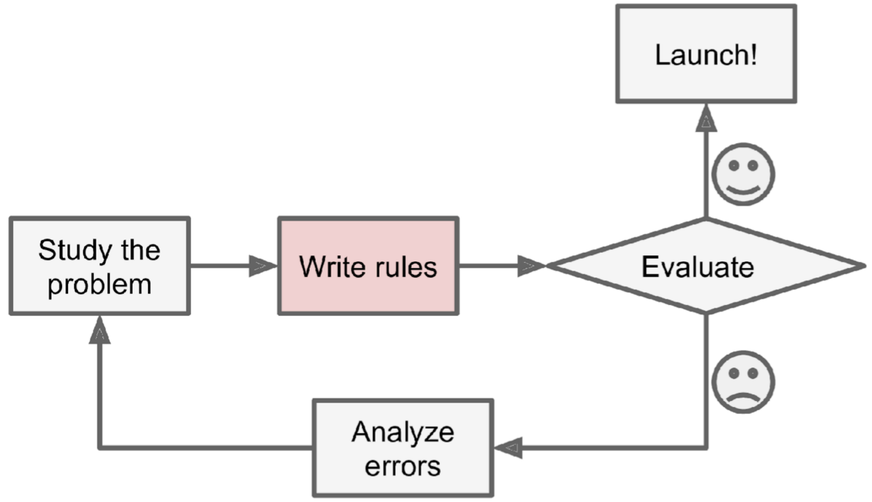
Consideremos la aplicación de un filtro de spam en el correo electrónico:

¿Cómo escribir un filtro de spam usando técnicas de programación tradicionales?

1. Identificar las palabras o frases características de correos spam.

2. Identificar patrones en el remitente o cuerpo del correo.

3. Escribir un programa que detecte los rasgos identificados.

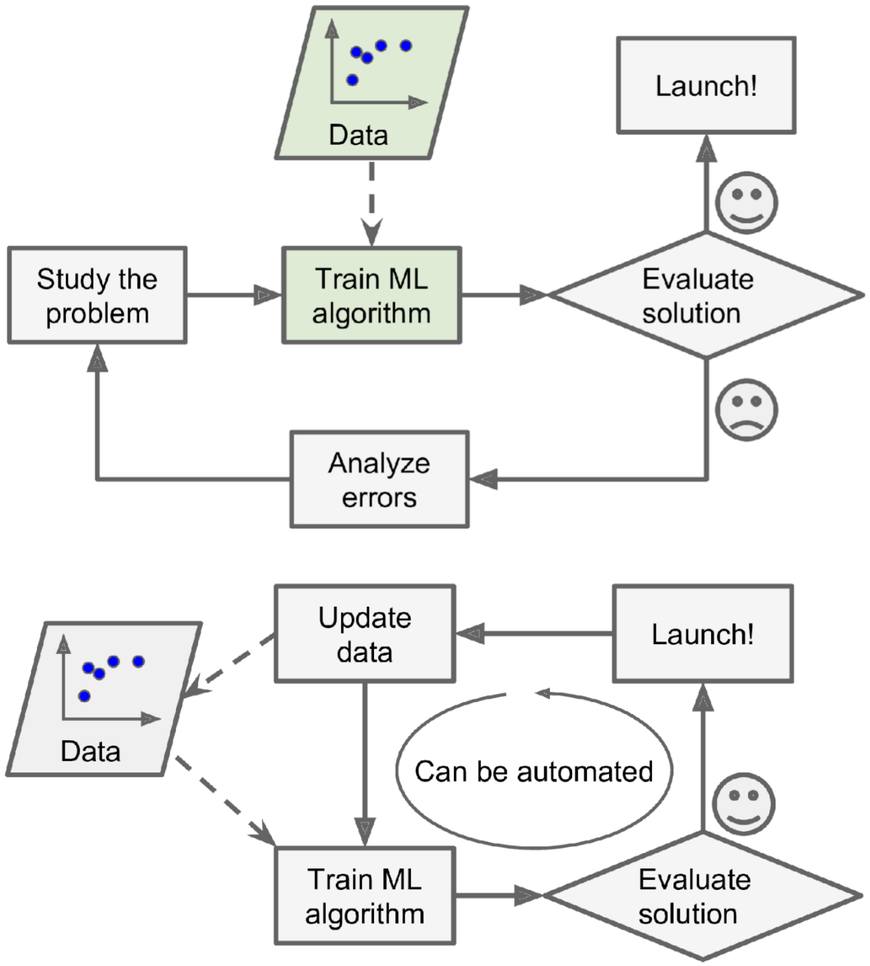


Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

¿Por qué usarlo?

En cambio, un filtro de spam basado en ML aprende automáticamente que palabras o frases son buenos predictores de un correo spam.

Puede ajustarse automáticamente a cambios en las palabras usadas por los spammers, sin necesidad de la intervenir o escribir nuevas reglas.



Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

¿Por qué usarlo?

También está el caso de tareas más complejas para las que definir un algoritmo o un conjunto de reglas se hace más complicado, por ejemplo:

- Reconocimiento del habla (Speech Recognition): se podría trabajar en el dominio de la frecuencia, pero considerando múltiples acentos e idiomas se vuelve una tarea muy compleja.
- Visión por computador (Computer Vision): los modelos de ML han sobrepasado el desempeño de algoritmos tradicionales basados en procesamiento digital de imágenes.

Tipos de sistemas

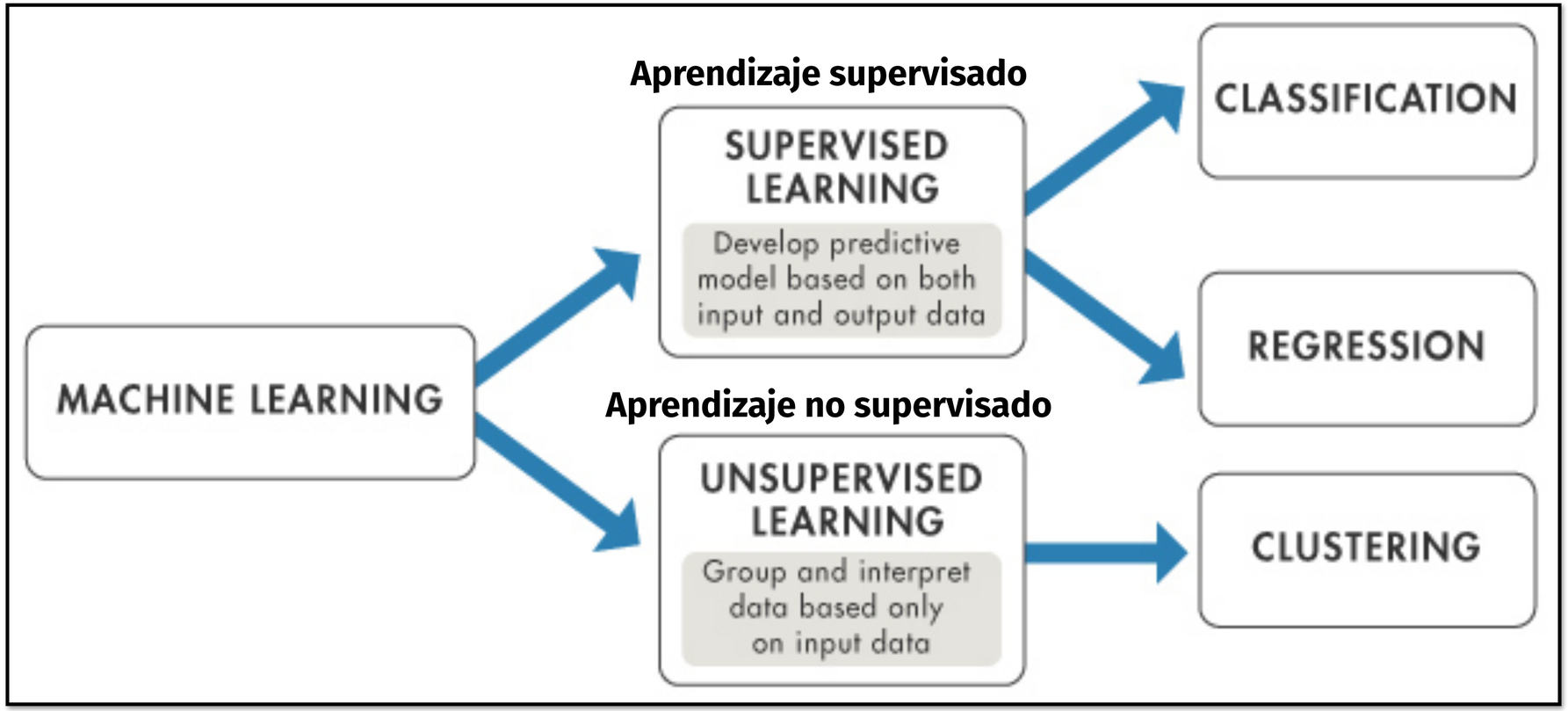
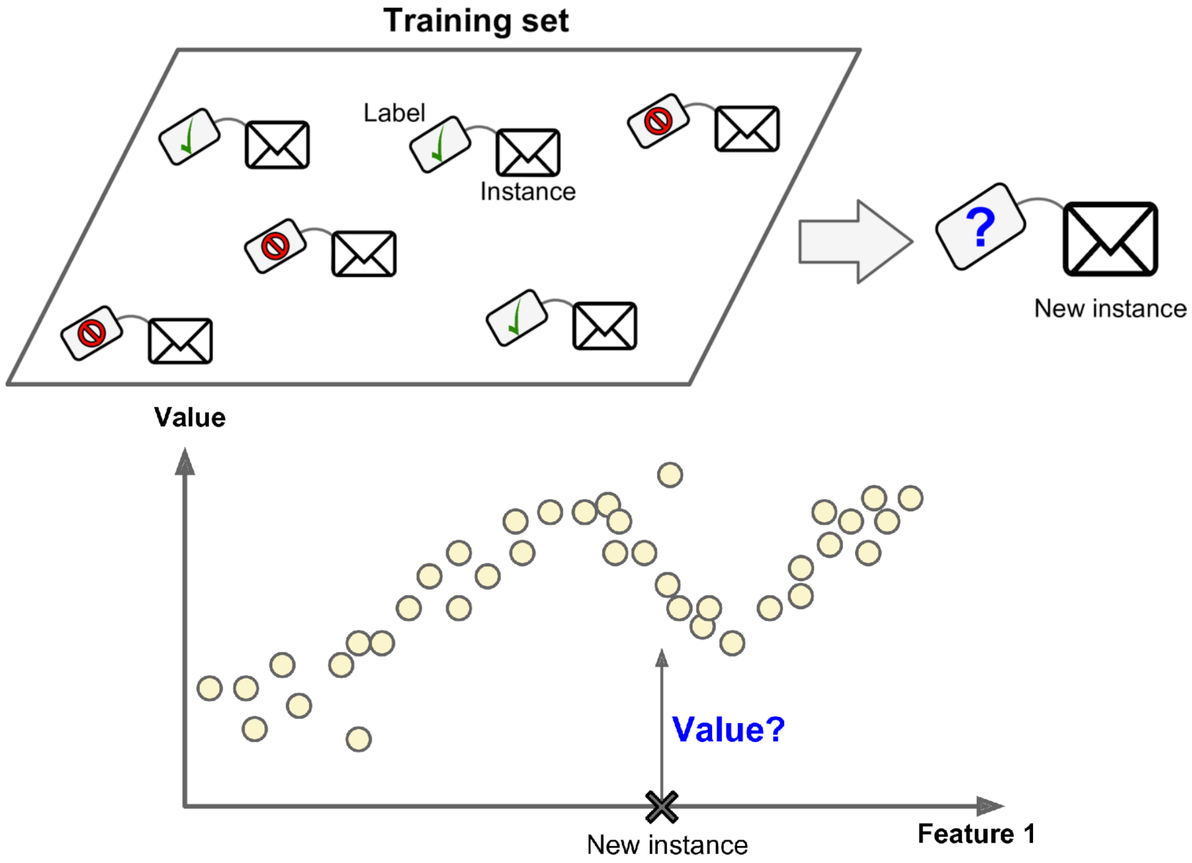


Imagen de MathWorks

Tipos de sistemas

Aprendizaje supervisado

El conjunto de entrenamiento, además de las características (*features*) o atributos, contiene las soluciones u objetivos (*etiquetas/ labels)*. Las tareas típicas son la clasificación (las etiquetas son categorías) y la regresión (las etiquetas son valores numéricos).



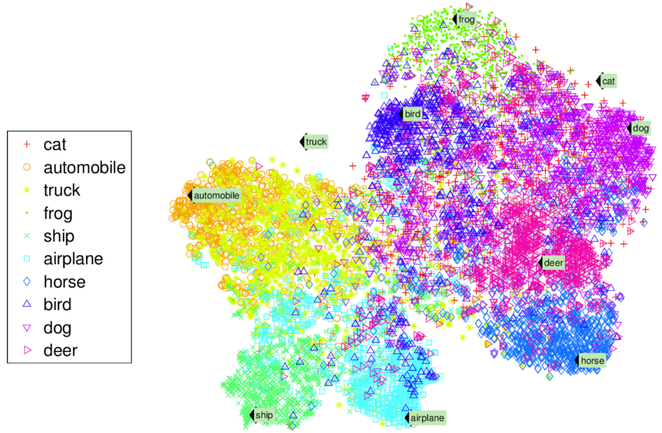
Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

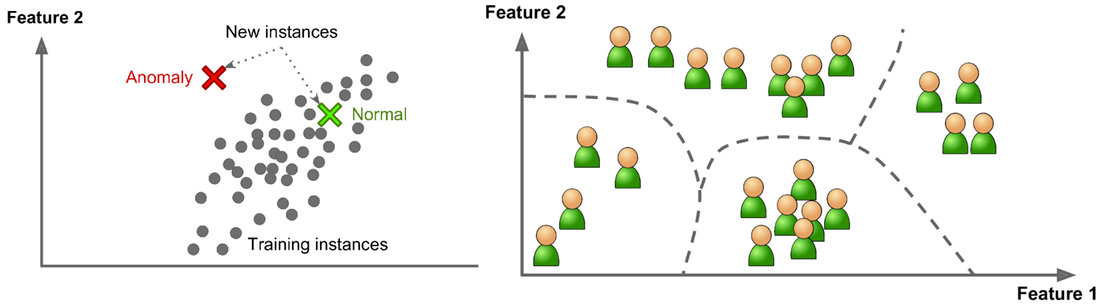
Tipos de sistemas

Aprendizaje no supervisado

Este tipo de aprendizaje trabaja con datasets no etiquetados. Las tareas más importantes son:

- Agrupamiento (*Clustering*)
- Detección de anomalías
- Visualización y reducción de dimensionalidad





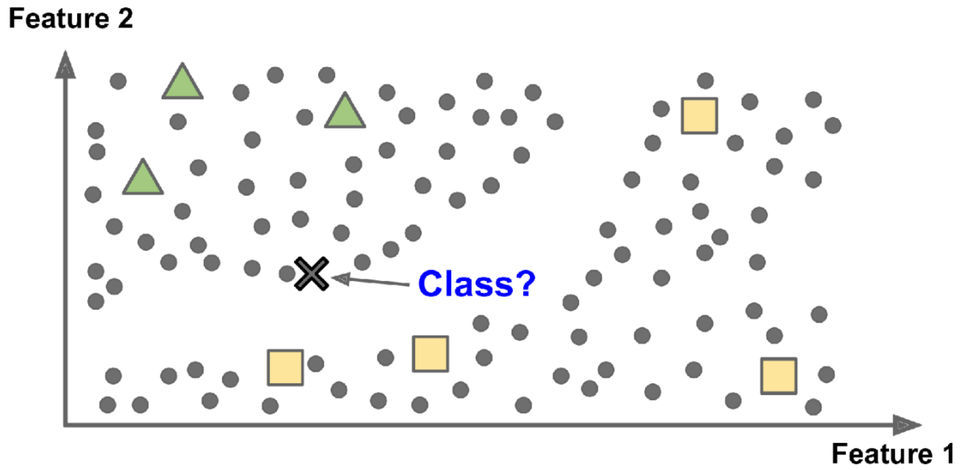
Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas

Aprendizaje semi-supervisado

En muchas ocasiones, la tarea de etiquetar los datos es costosa y requiere de mucho tiempo, una alternativa es etiquetar solo unas cuantas muestras y trabajar con datasets parcialmente etiquetados.

Por ejemplo, el reconocimiento de personas en fotos por Google Photos.



Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tareas básicas

Regresión

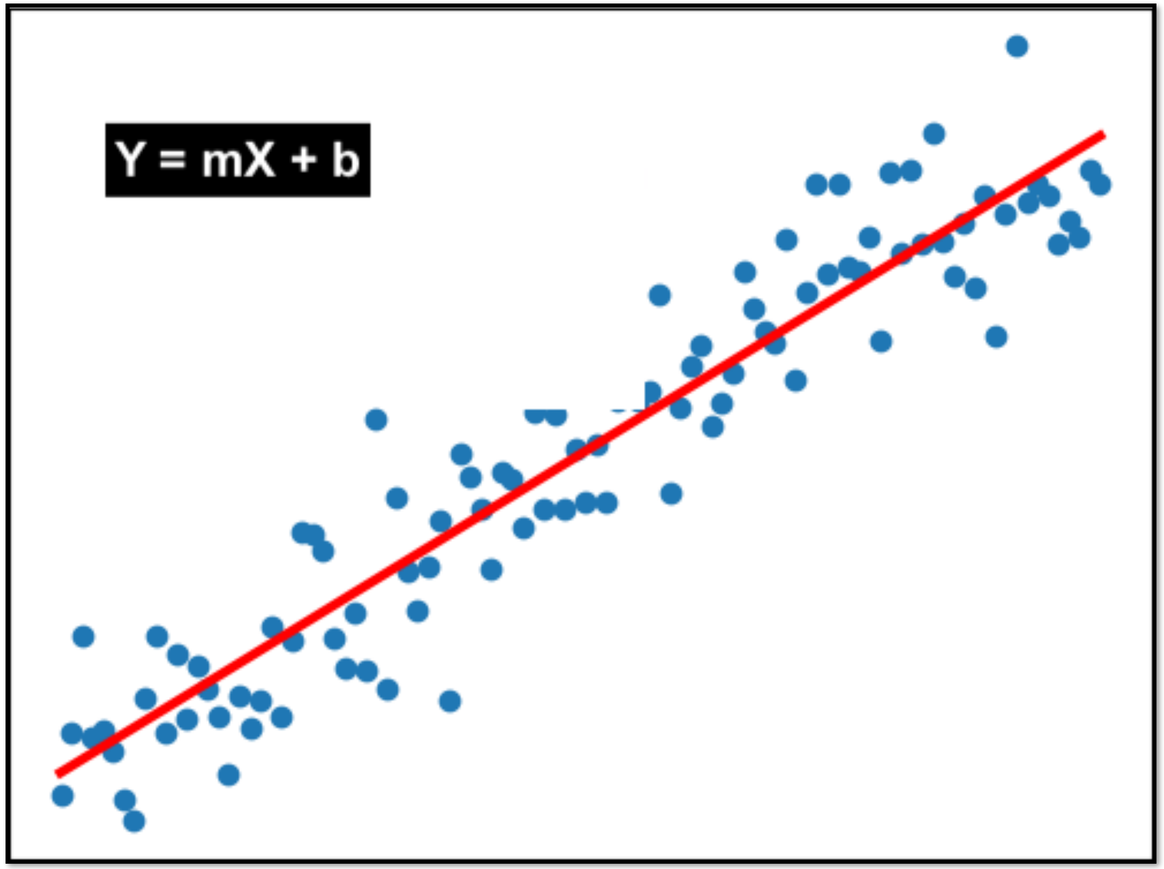


Imagen de Medium

Tareas básicas

Clasificación

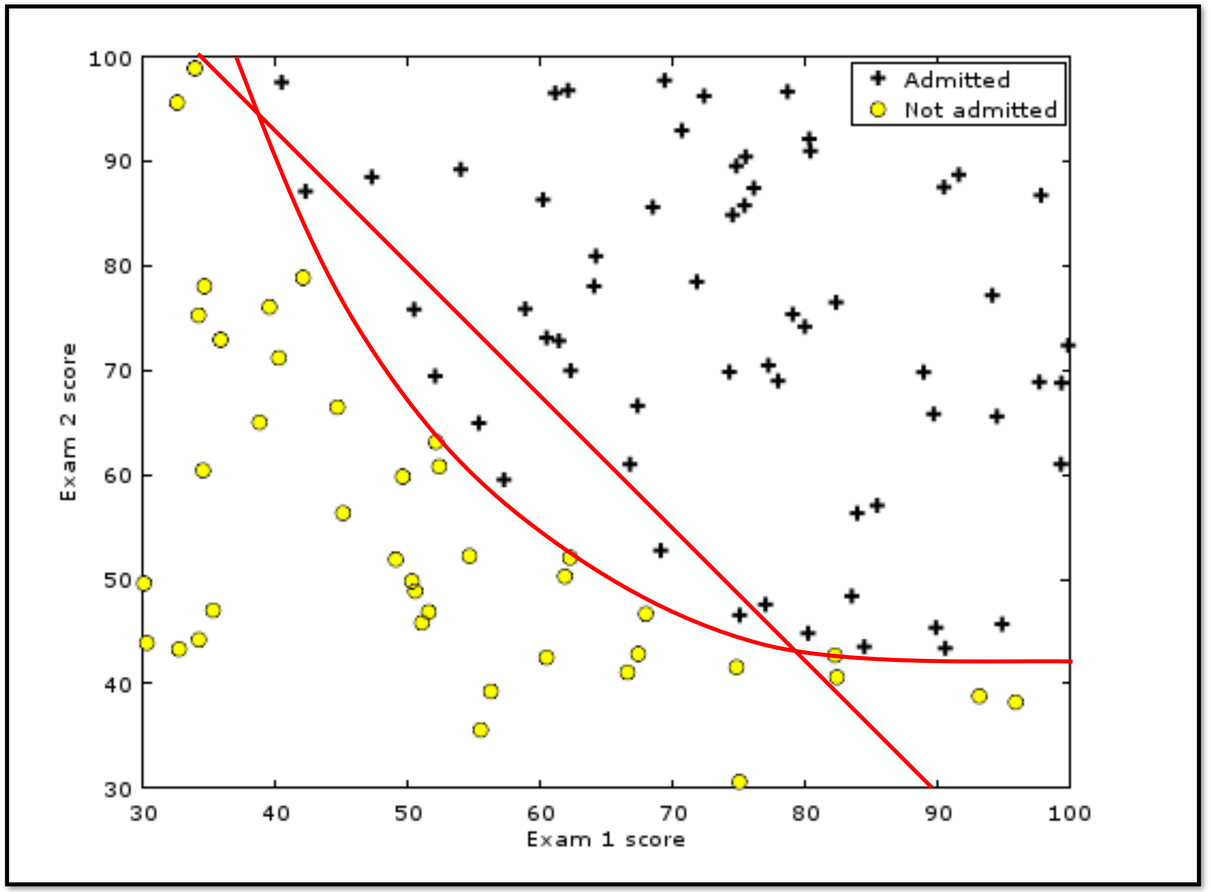


Imagen de Blog

Tareas básicas

Agrupación

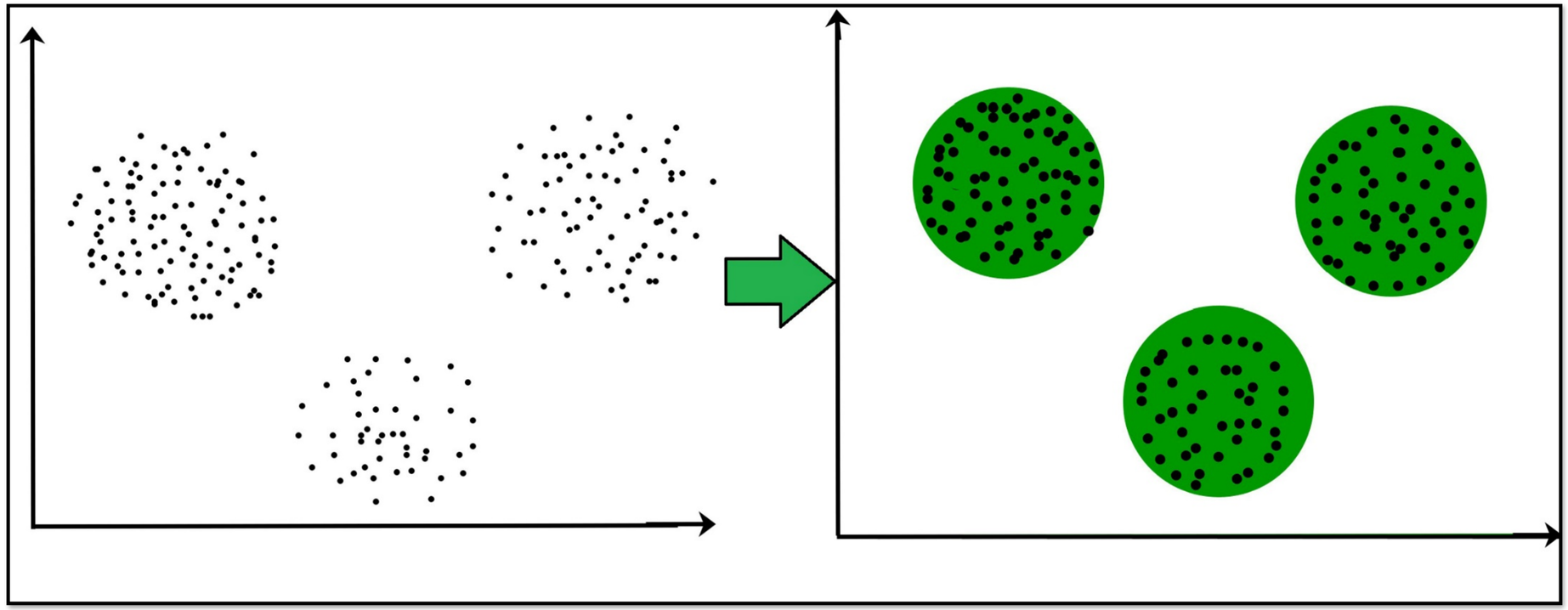
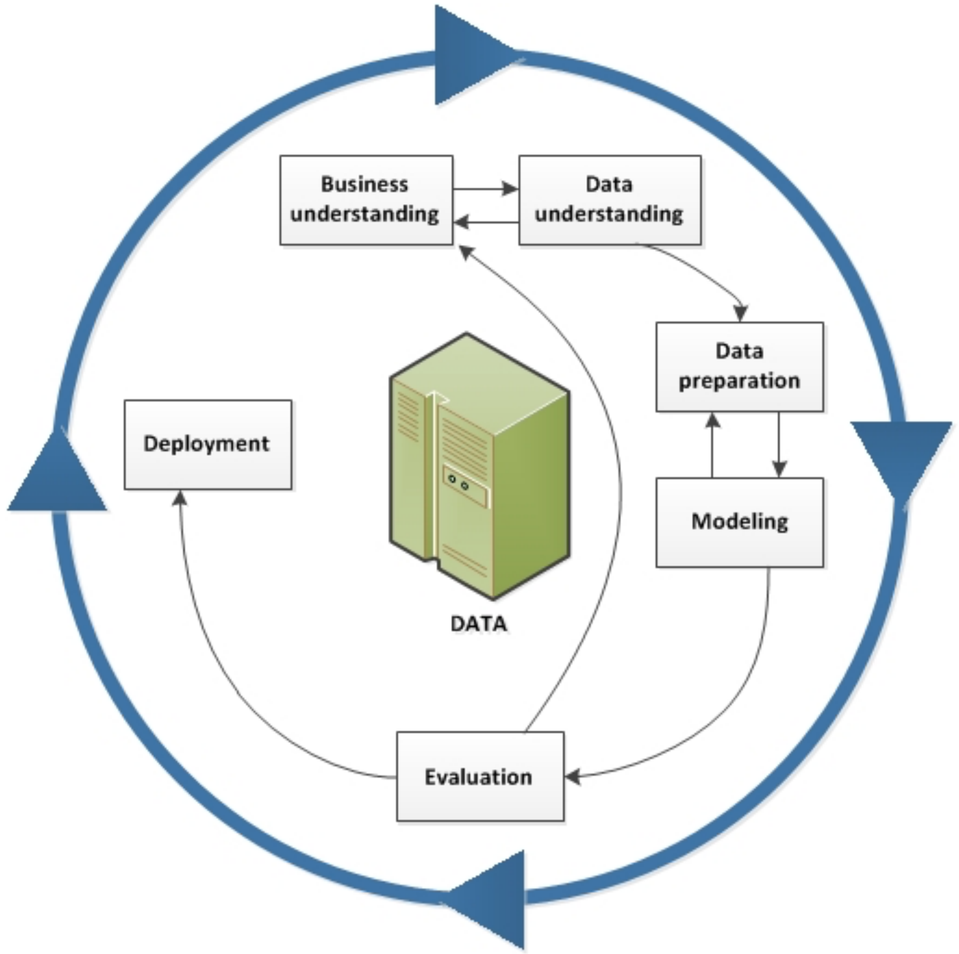


Imagen de Geeks for Geeks

Etapas de desarrollo



Metodología CRISP-DM. Imagen de IBM

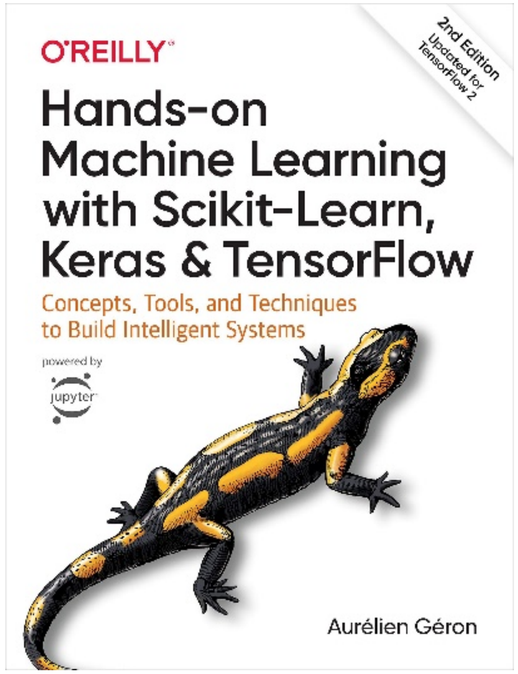
1. Entendimiento del negocio (objetivos)
2. Exploración de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

Este es un proceso iterativo.

Principales retos

- Cantidad insuficiente de datos
  - En la mayoría de los casos, tener más datos implica mejores resultados
- Datos de entrenamiento no representativos (*Data mismatch*)
  - El objetivo final es tener un buen desempeño en nuevas muestras o datos no conocidos, por lo que se requieren datos de entrenamiento con características similares
- Datos de baja calidad
- Características irrelevantes

Recursos



Hands-On Machine Learning with Scikit-Learn & TensorFlow

Machine Learning Library (MLlib) de Apache Spark - Documentación

Conceptos básicos de CRISP-DM por IBM - Guía