PROYECTO APLICADO

"You can't manage what you don't measure",
Tom De Marco, 1982

"We´re drowning in data but starving for knowledge",
John Naisbitt, 1982

UNIVERSIDAD ICESI

DESCRIPCIÓN

En este curso corto (4 sesiones), utilizando un enfoque práctico, se repasarán los conceptos básicos de la analítica de datos, en especial los propios del aprendizaje supervisado (regresión y clasificación) y no supervisado (segmentación y reducción de dimensiones).

UNIVERSIDAD ICESI

DESCRIPCIÓN

Aprendizaje Supervisado

- Tareas de regresión
- Tareas de clasificación

Aprendizaje No supervisado

- Tareas de segmentación
- Tareas de reducción de dimensiones

UNIDADES

Unidad 1

Aprendizaje supervisado

Regresión

Unidad 2

Aprendizaje supervisado

Clasificación

Unidad 3

Aprendizaje no supervisado

Segmentación

Unidad 4

Aprendizaje no supervisado

Reducción de dimensiones

UNIVERSIDAD ICESI

METODOLOGÍA

Sesión de clase

- Repaso de los conceptos básicos a utilizar dentro del proyecto.
- Presentación y discusión del proyecto
- Revisión del conjunto de datos a utilizar en la tarea.
- Trabajo grupal

ANDRÉS A. ARISTIZÁBAL P.

Formación

- Ingeniero de Sistemas y Computación de la Universidad Javeriana Cali, 2006
- Doctorado en Informática, École Polytechnique de París, Francia, 2012

Experiencia académica

- Investigador en el grupo Avispa de la Universidad Javeriana Cali, 2006 - 2009
- Investigador Posdoctoral en el grupo de Lenguajes de Programación de la Universidad de Wrocław, Polonia 2014 - 2015
- Profesor hora cátedra de la Universidad Javeriana Cali, 2015
- Profesor hora cátedra de la Universidad Icesi Cali, 2016 – 2017
- Desde 2017, profesor tiempo completo Universidad Icesi, Facultad de Ingeniería

UNIVERSIDAD ICESI

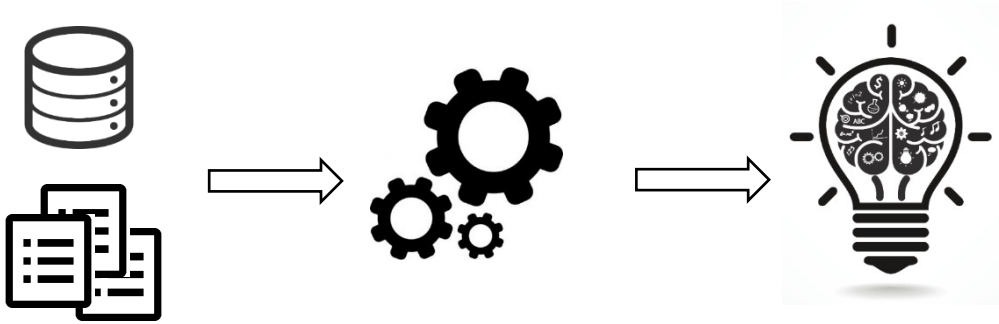
AGENDA DE HOY

- Aprendizaje automático
- Aprendizaje supervisado
- Tareas de regresión
- Modelos de regresión
- Proyecto de regresión

APRENDIZAJE AUTOMÁTICO

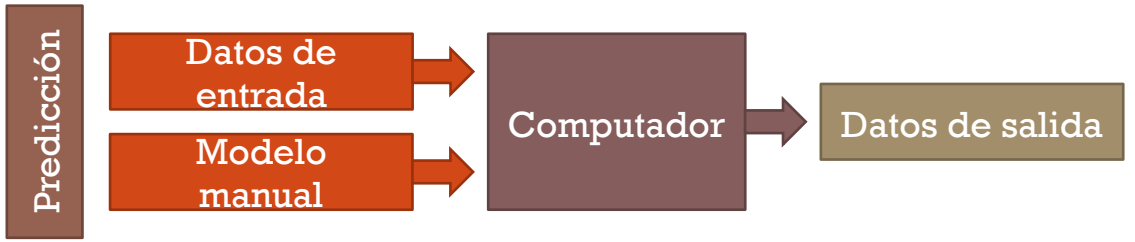
- Definición:

  El aprendizaje automático es la ciencia que permite a los computadores aprender, sin ser explícitamente programados[1]

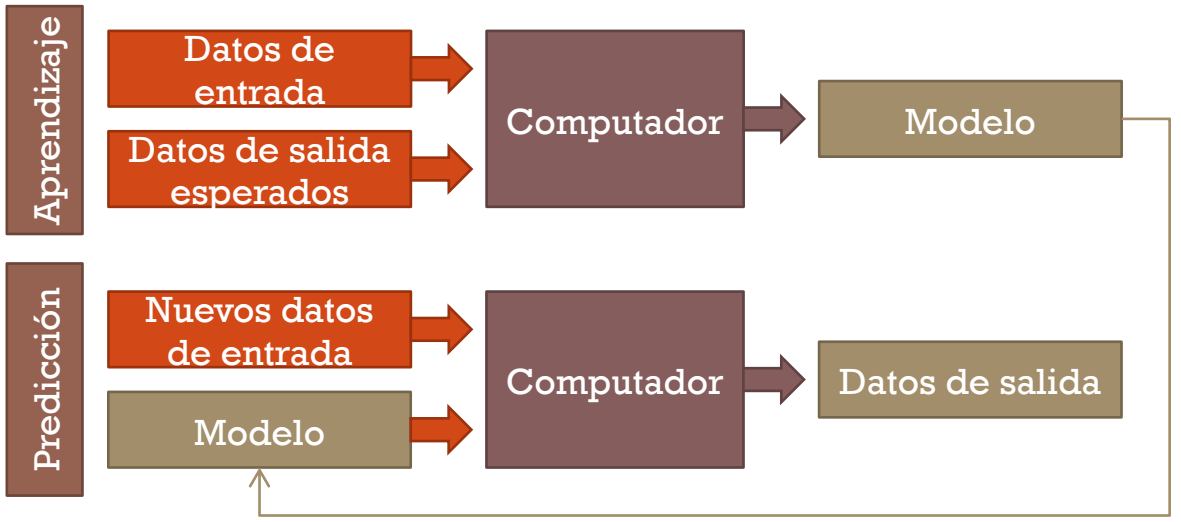


1. Andrew Ng, Stanford University, 2014

Modelo tradicional



Ciencia de datos



UNIVERSIDAD ICESI

APRENDIZAJE SUPERVISADO

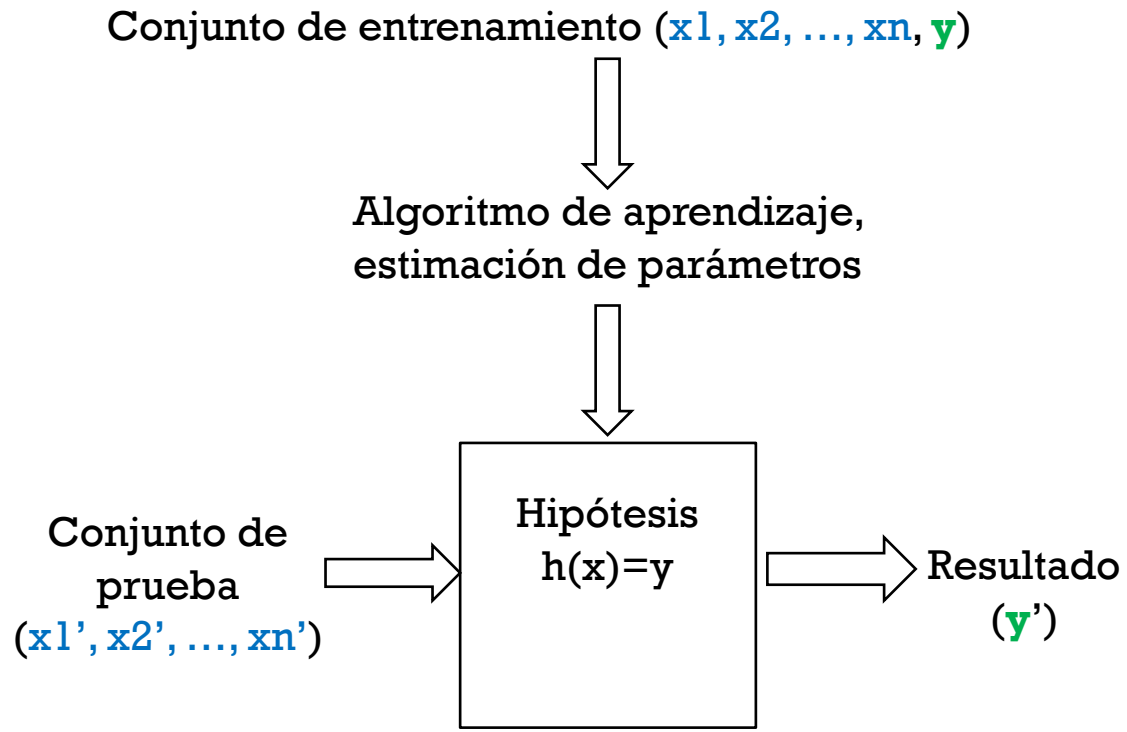
- Aprender a partir de un "experto"
- Datos de entrenamiento **etiquetados** con una clase o valor:

$(x1, x2, \ldots, xn, y)$

Predictores, variables de entrada (independientes)

Respuesta, variable de salida (dependiente)

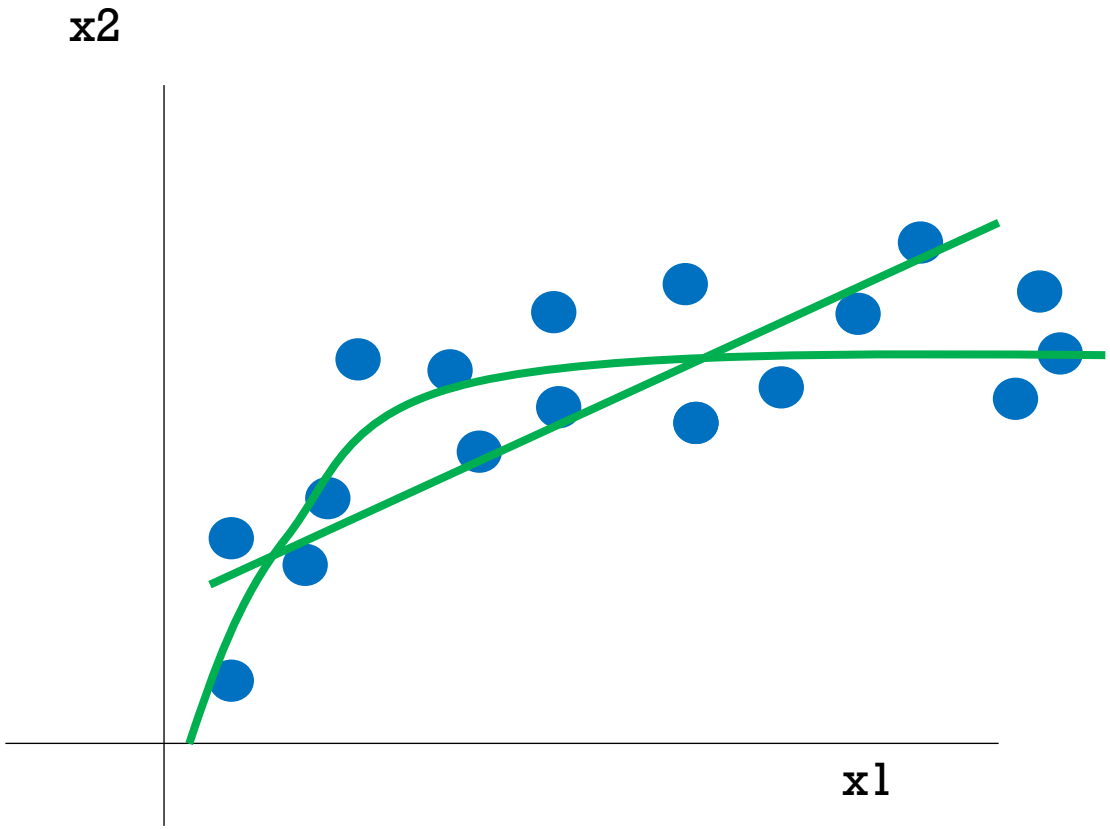
- **Meta**: predecir una clase o valor



UNIVERSIDAD ICESI

TAREAS DE REGRESIÓN

- Encontrar modelos, f, que permitan predecir valores continuos:
  - KNN
  - Regresión lineal
  - Regresión polinómica
  - Árboles de regresión
  - …
- Valores **numéricos** de la variable o función objetivo
- **Baseline**: medida de evaluación dada por un modelo que predice una medida de tendencia central (e.g. el promedio) (→"modelo **nulo**")



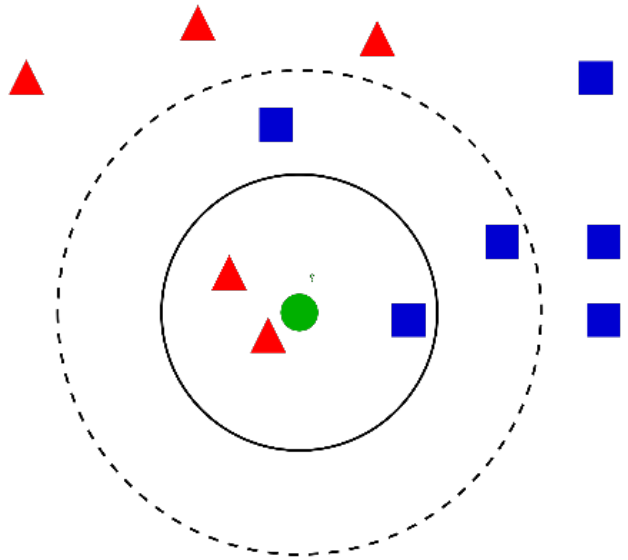
MODELOS DE REGRESIÓN

- **KNN**
- **Árboles de decisión**
- **Random Forest**
- **Boosting**

MODELOS DE REGRESIÓN

KNN:

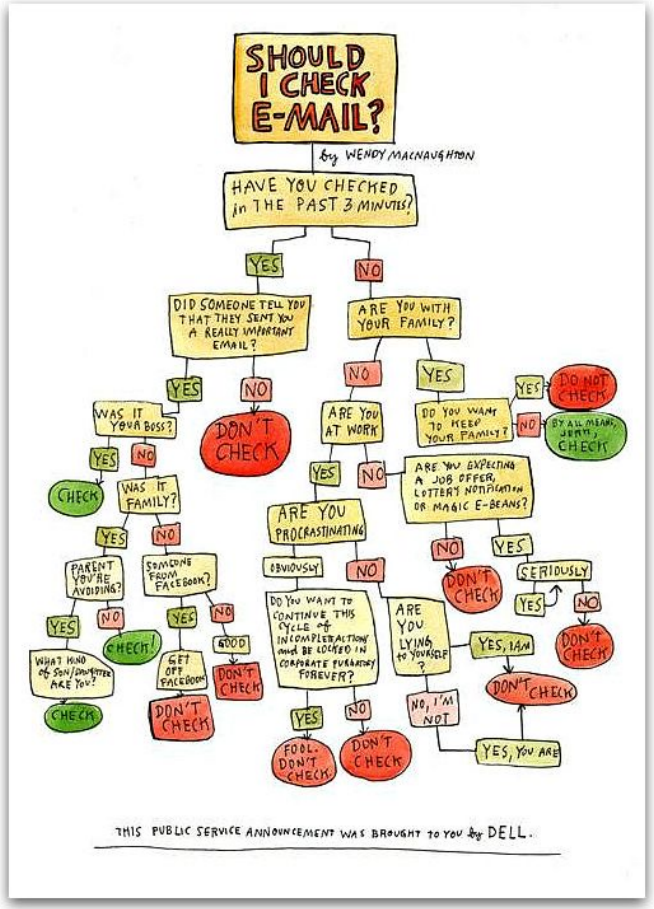
- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Asigna la clase o valor agregado de las instancias conocidas que se encuentran mas cerca de la instancia a predecir.



MODELOS DE REGRESIÓN

Árboles de decisión:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Posee una estructura de árbol, donde sus nodos internos representan las características, las ramas las decisiones y cada hoja el resultado.

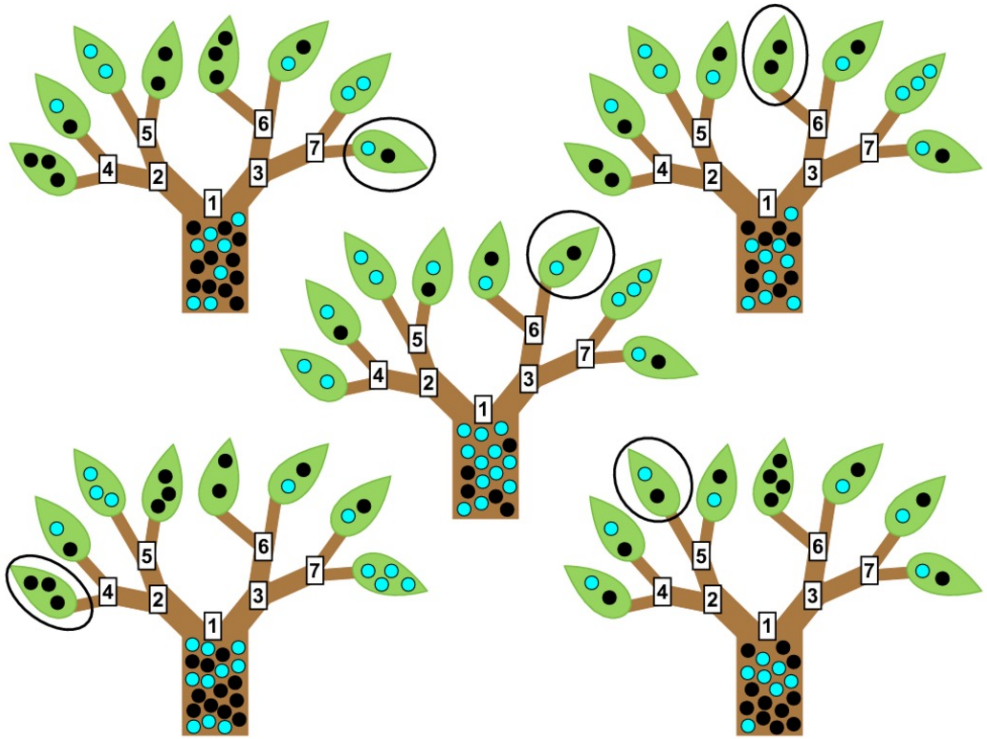


UNIVERSIDAD ICESI

MODELOS DE REGRESIÓN

Random Forest:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Basado en el concepto de ensambles, el cual es un proceso de combinación de múltiples regresores o clasificadores para resolver un problema complejo y mejorar el rendimiento del modelo.

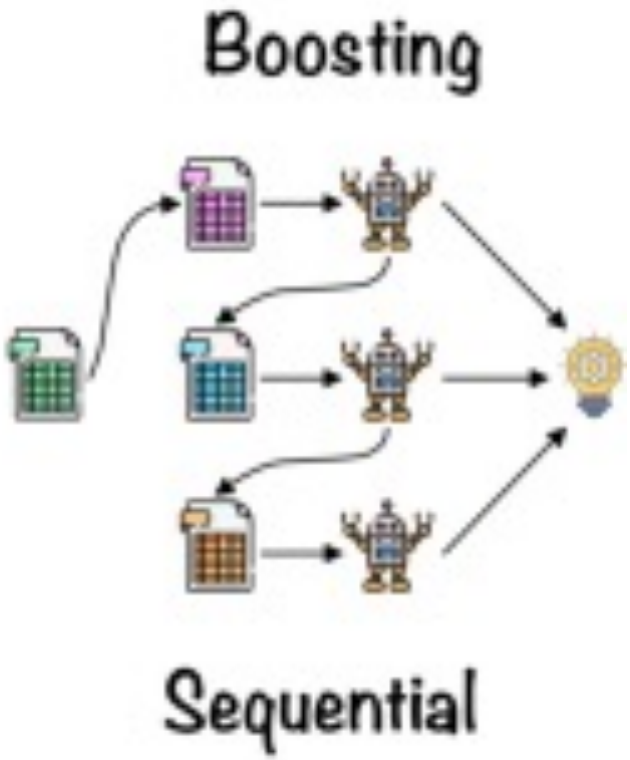


UNIVERSIDAD ICESI

MODELOS DE REGRESIÓN

Boosting:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Mejora la exactitud y rendimiento de los modelos de aprendizaje automático convirtiendo múltiples modelos débiles en uno mucho más potente.



Proyecto de regresión

- Análisis Exploratorio de Datos
  - Limpieza de datos
    - Formatos
    - Valores faltantes
      - Identificación (explícitos e implícitos)
      - Imputación (media y KNN)
      - Eliminación (columnas y/o registros)
    - Valores atípicos
      - Visualización (boxplots e histogramas)
      - Limitación inferior y superior (winsorizing)
  - Exploración de datos
    - Análisis univariado
      - Variables continuas
      - Variables categóricas
    - Análisis bivariado
      - Independientes continuas vs variable objetivo
      - Variables independientes continuas entre sí
      - Variables categóricas con respecto a la variable objetivo
  - Ingeniería de características
    - Creación de nuevas características
      - Dummificación
    - Eliminación de características
      - Variables independientes continuas altamente correlacionadas entre sí
      - Variables independientes categóricas con baja correlación con la objetivo

Proyecto de regresión

- Protocolos de evaluación
- Métricas de evaluación
  - R^2
  - RMSE
- Implementación de modelos
  - Baseline
  - Persistencia
    - Uso de pickle para guardar modelos y métricas
  - KNN
  - Árbol de regresión
  - Random forest
  - Gradient Boosting
  - XGBoosting
- Comparación de modelos

Ejercicio

Replicar los mismos pasos realizados con el conjunto de datos de la esperanza de vida, pero en este caso con respecto al conjunto de datos de admisiones universitarias.

Con este dataset se busca predecir la probabilidad de admisión universitaria de un estudiante, en particular, a partir de varios parámetros.