Aprendizaje supervisado

Taller comercio y desarrollo regional

Módulo Aprendizaje de máquinas

Curso - Taller

Educación continua | Universidad de los Andes
octubre 26 – noviembre 5
2021

Photo by **ThisIsEngineering** from **Pexels**

Hoy

De qué vamos a hablar

1. **Nuevo plan:**
Algunos ajustes desde nuestra conversación pasada.

2. **Aprendizaje supervisado:**
¡Vamos a ver técnicas!

3. **Script de aprendizaje de máquinas:**
El proceso general del aprendizaje de máquinas.

4. **Taller diferenciado**

Nuevo plan

Con ajustes después de la conversación de la clase pasada...

Objetivos: al final del módulo

Objetivo de competencia general:

- Críticamente evaluar una situación y asociarla con algunas técnicas que pueden contribuir de forma concreta.

Objetivo de acompañamiento:

G1: En la elaboración del informe de auditoría.

G2: En la identificación de retos concretos contextualizados en la CGR en los que se puedan emplear estas herramientas.

Estructura del módulo 3

- Pequeños ajustes

Primera mitad	Segunda mitad
~~Clase 1 – Fundamentos del Aprendizaje de Máquinas~~	~~Exposición con partes de taller.~~
Modelos de aprendizaje supervisado – ¡Script…	… Script! – Taller diferenciado 1
Modelos de aprendizaje no supervisado - Ética de …	… la inteligencia artificial – Taller diferenciado 2
Datos no estructurados y evaluación - ¡Script…	... modelos! – Taller diferenciado 3

¿Cómo funciona el taller diferenciado?

- Vamos a dividir en grupos: auditoría de desempeño vs los demás
- Hay unas dinámicas prediseñadas para el aprendizaje autónomo
- El tutor va a rotar entre los dos grupos conversando, dando retroalimentación y facilitando

¿Cómo funciona el taller diferenciado?

- Vamos a dividir en grupos: auditoría de desempeño vs los demás
- Hay unas dinámicas prediseñadas para el aprendizaje autónomo
- El tutor va a rotar entre los dos grupos conversando, dando retroalimentación y facilitando

Grupo con auditoría de desempeño

1) En qué estamos, veamos los datos => orientación concreta para construir indicadores e identificar hallazgos sobre los objetivos

2) Tiempo para avanzar en el informe de auditoría

3) Tiempo para avanzar en el informe de auditoría

¿Cómo funciona el taller diferenciado?

- Vamos a dividir en grupos: auditoría de desempeño vs los demás
- Hay unas dinámicas prediseñadas para el aprendizaje autónomo
- El tutor va a rotar entre los dos grupos conversando, dando retroalimentación y facilitando

Grupo con interés en explorar aplicaciones del ML

1) Fuentes de datos posibles => Posibles retos de ML
2) Las propiedades de la inteligencia artificial => conectarla con procesos y procedimientos en la CGR
3) Un último script en R con técnicas de refinamiento de modelos

Aprendizaje supervisado

Un vistazo a las técnicas…

Lo que vamos a ver en un mapa

Situación

- ¿Qué datos tenemos?
  - ¿Estructurados?
  - ¿Etiquetados?
- Necesidad de análisis
  - Descripción
  - Predicción
  - Decisión [!]

Informes sobre el pasado

Herramientas para nuevos datos

* Ojo con la ética (otra clase)

Lo que vamos a ver en un mapa

Situación

- ¿Qué datos tenemos?
- Necesidad de análisis
  - Descripción
  - Predicción
  - Decisión [!]

Tipo de proyecto

- Análisis descriptivo
- Aprendizaje supervisado
- Aprendizaje no supervisado

Descriptivas univariadas: medias, frecuencias, por grupos, agregradas…
Descriptivas multivariadas: correlaciones, regresiones…

Técnicas de clasificación / regresión

Reducción de dimensionalidad, clustering…

EL PROCESO del ML…

Lo que vamos a ver en un mapa

Situación

- ¿Qué datos tenemos?
- Necesidad de análisis
  - Descripción
  - Predicción
  - Decisión [!]

→

Tipo de proyecto

- Análisis descriptivo
- Aprendizaje supervisado
- Aprendizaje no supervisado

→

Técnicas

- Requisitos, recomendaciones
- Alcances y tipo de conclusiones
- Limitaciones
- Detalles de código

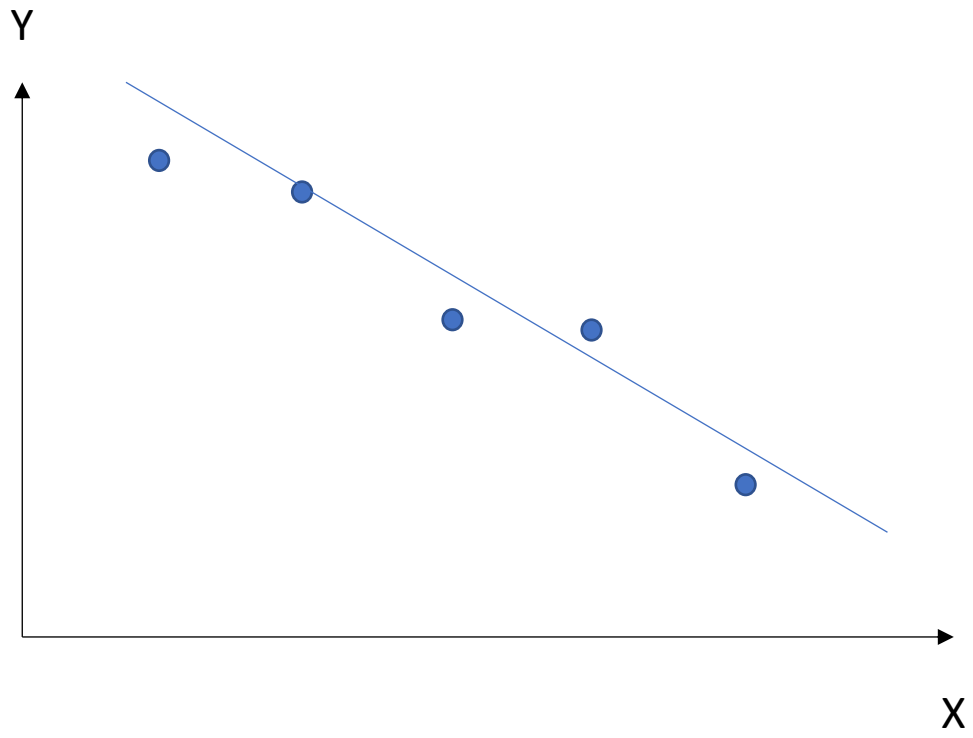
Esto lo podemos buscar en Google

Recordemos que contamos con:

- Datos que tienen características de cada observación y una variable de “etiqueta” que queremos predecir en datos nuevos

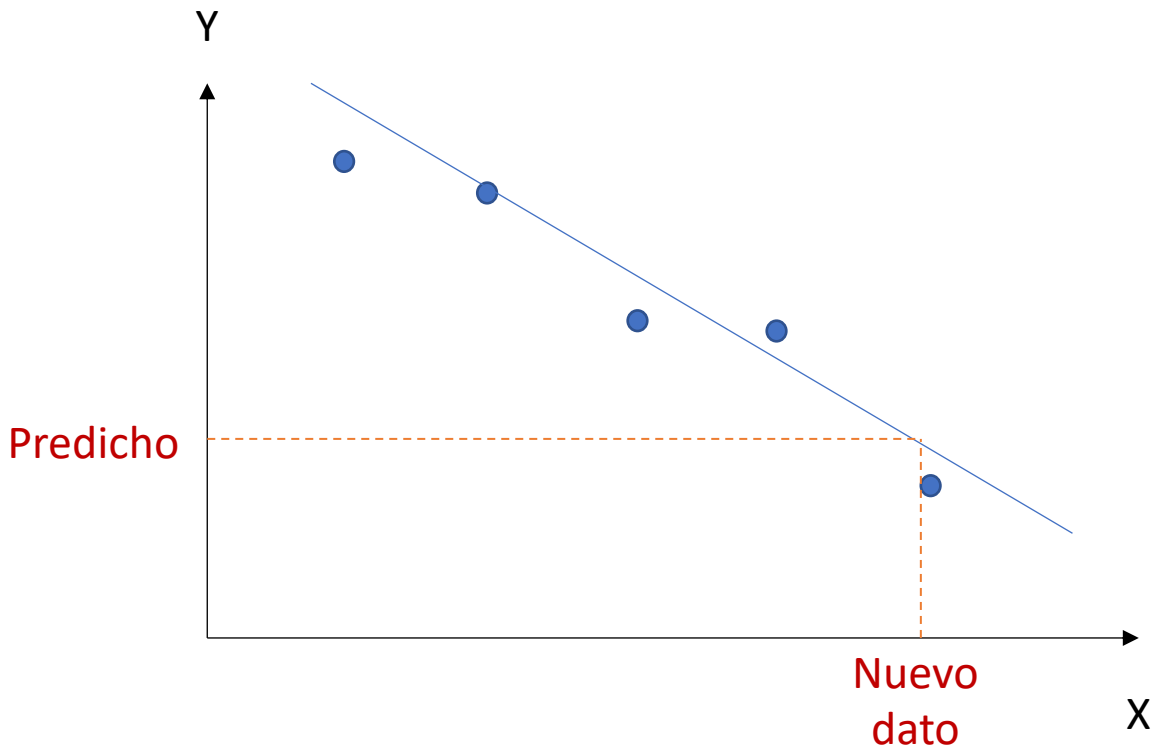
Identificador	Dato útil 1	Dato útil 2	ETIQUETA
Alfredo	12	7432	100.000
Juliana	16	9375	150.000

Regresión lineal



- El caso de clara y el más conocido por este grupo.
- Estudiamos la relación entre las cantidades características y la variable a predecir.
- Describimos esa relación con un modelo lineal.

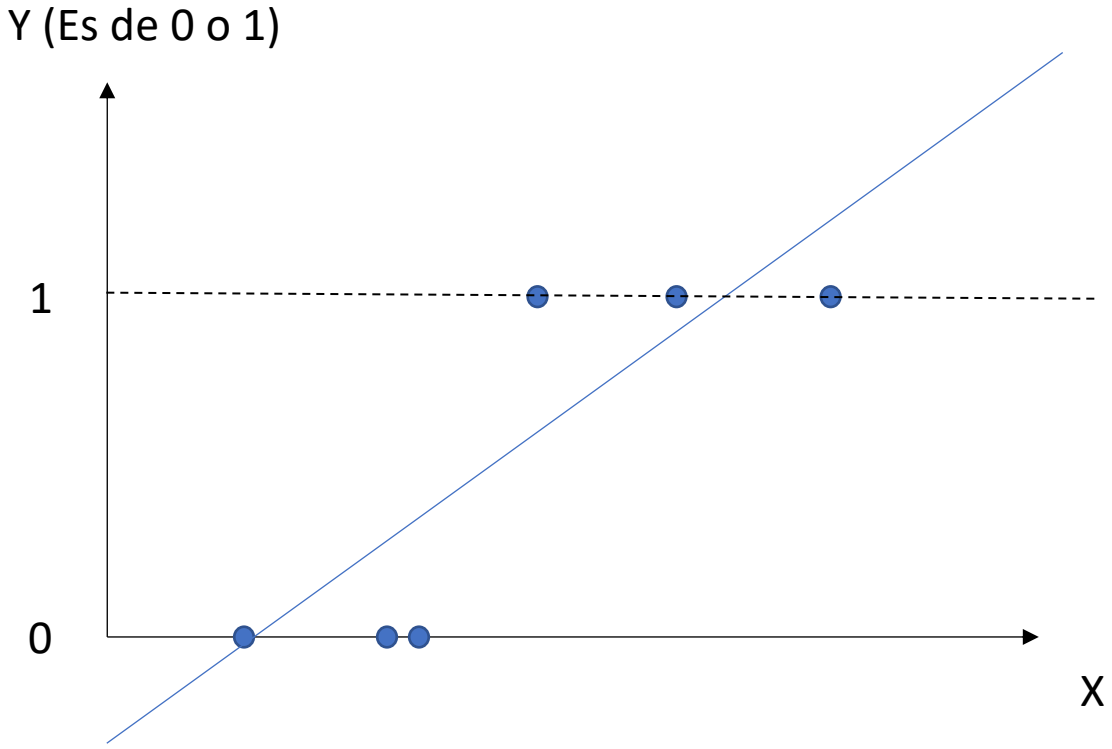
Regresión lineal



- El caso de clara y el más conocido por este grupo.
- Estudiamos la relación entre las cantidades características y la variable a predecir.
- Describimos esa relación con un modelo lineal.
- Cuando tenemos un nuevo conjunto de valores X, podemos consultar en el modelo y ver su correspondiente Y.
- Se ajustan los valores de la recta.

Regresión lineal

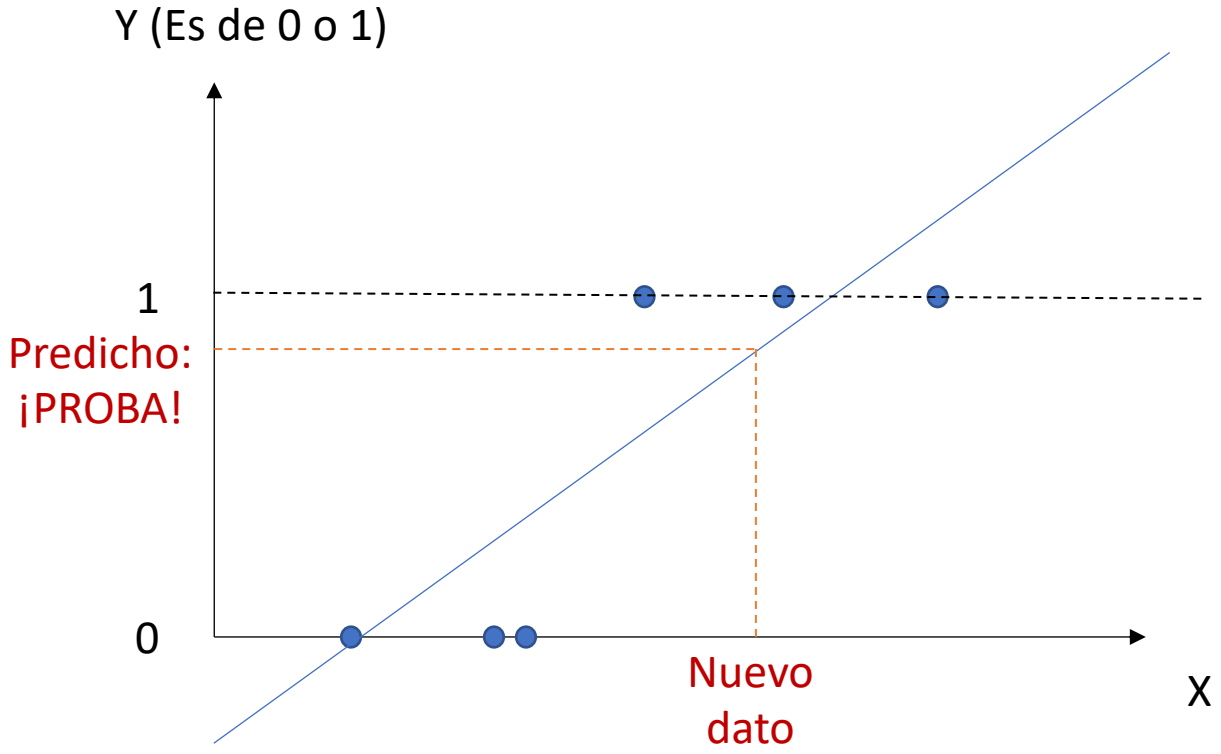
¡para clasificación binaria!



- Es igual que un modelo de regresión lineal, pero en lugar de tener una variable continua, nuestra etiqueta es una categoría.
- Describimos la relación entre las cantidades con un modelo lineal.

Regresión lineal

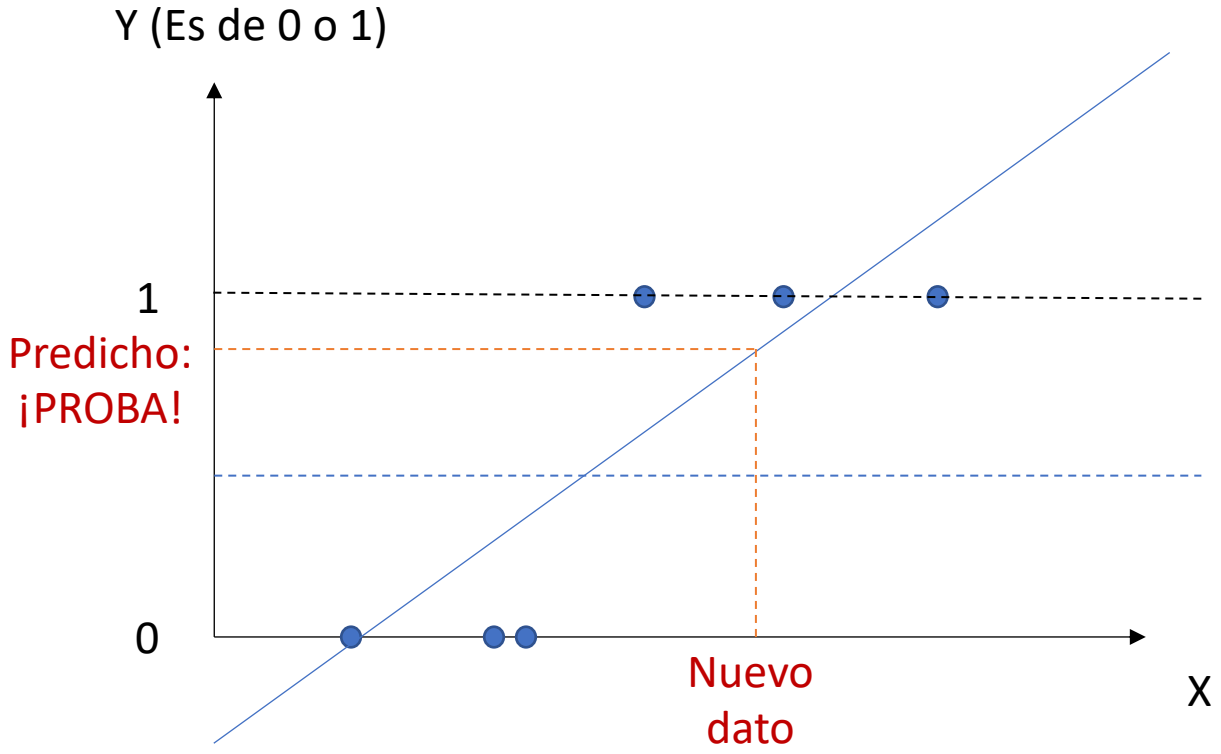
¡para clasificación binaria!



- Es igual que un modelo de regresión lineal, pero en lugar de tener una variable continua, nuestra etiqueta es una categoría.
- Describimos la relación entre las cantidades con un modelo lineal.
- Las predicciones de este modelo corresponden a la probabilidad de que la variable etiqueta tenga el valor de 1.

Regresión lineal

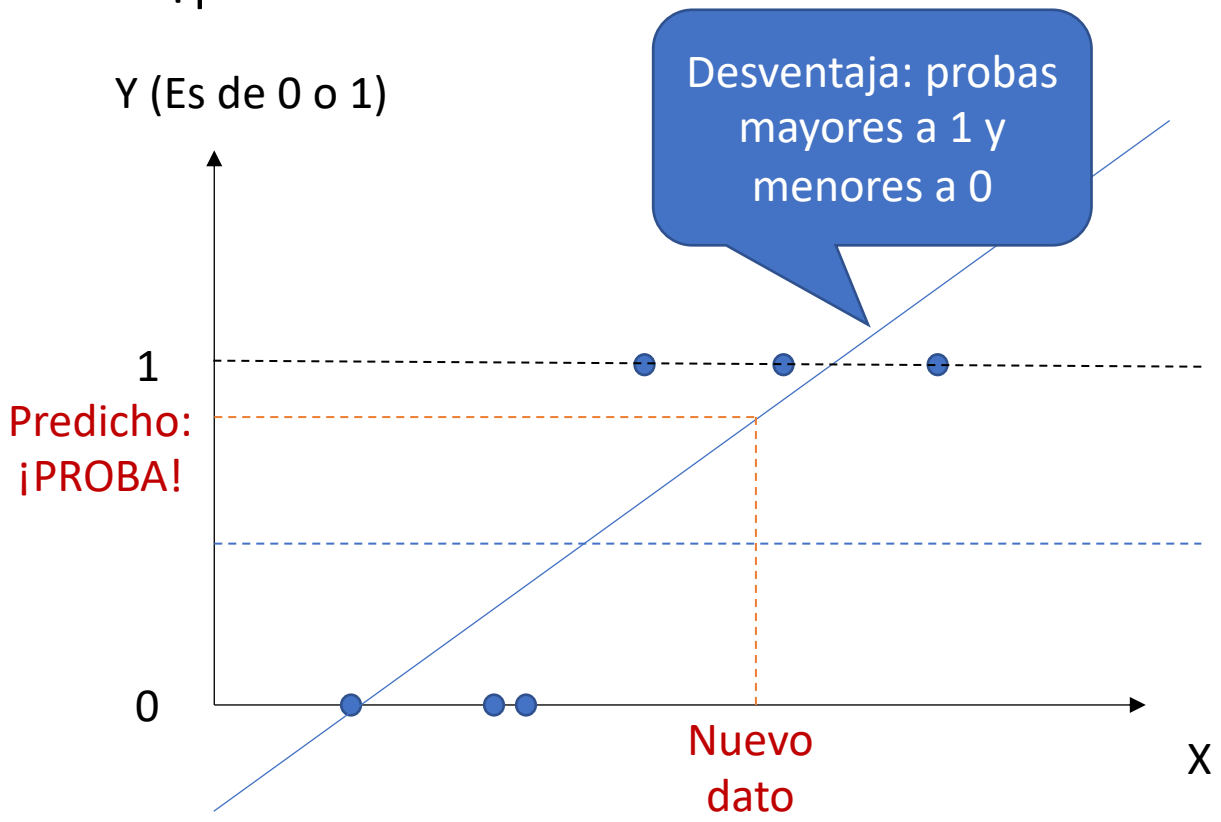
¡para clasificación binaria!



- Es igual que un modelo de regresión lineal, pero en lugar de tener una variable continua, nuestra etiqueta es una categoría.
- Describimos la relación entre las cantidades con un modelo lineal.
- Las predicciones de este modelo corresponden a la probabilidad de que la variable etiqueta tenga el valor de 1.
- Típicamente escogemos un umbral: si proba > 0.5 => y=1

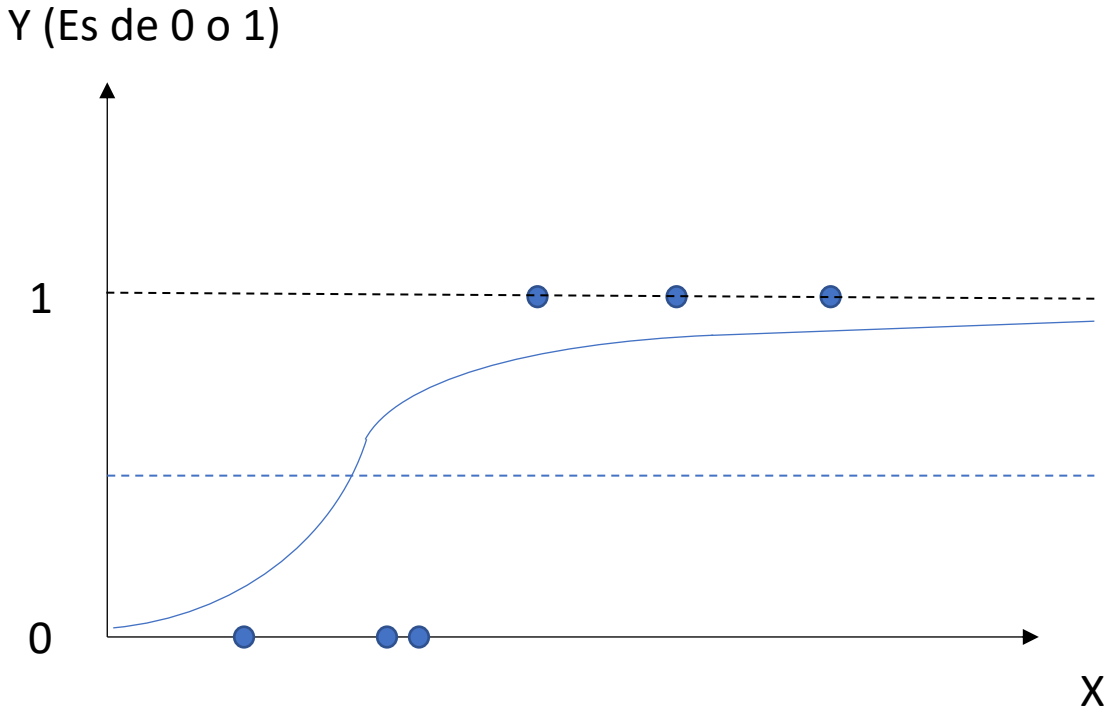
Regresión lineal

¡para clasificación binaria!



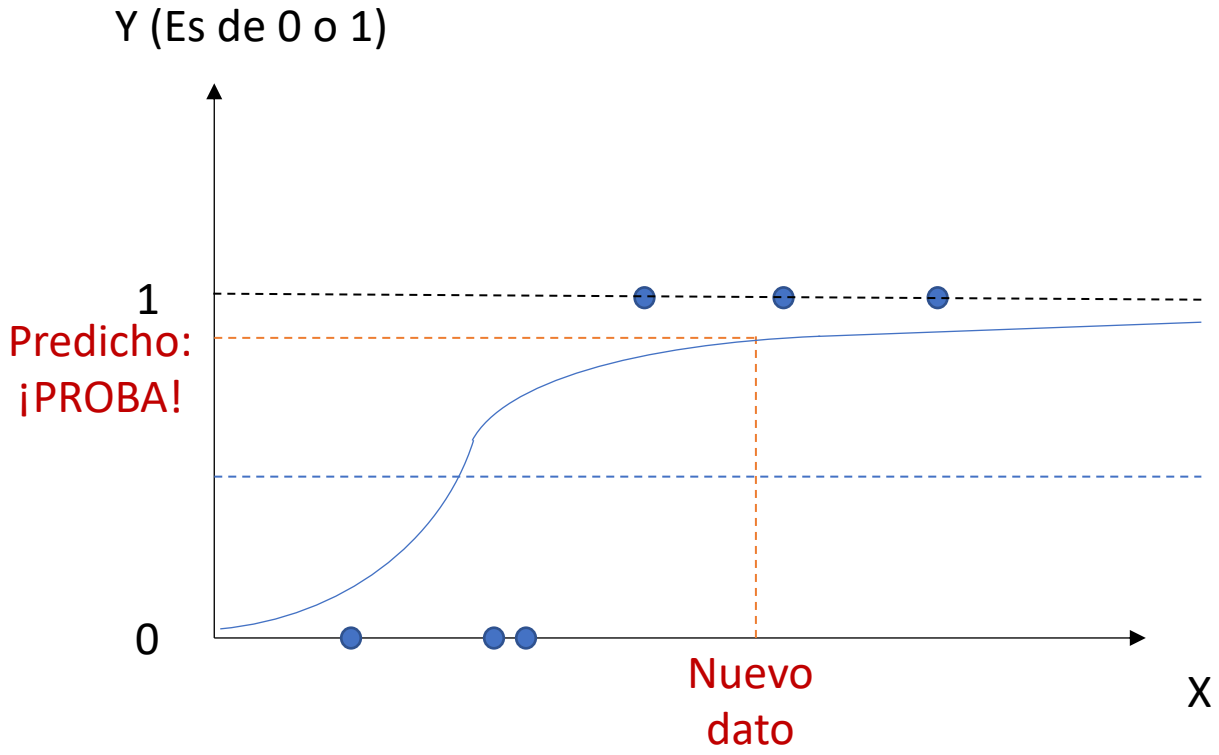
- Es igual que un modelo de regresión lineal, pero en lugar de tener una variable continua, nuestra etiqueta es una categoría.
- Describimos la relación entre las cantidades con un modelo lineal.
- Las predicciones de este modelo corresponden a la probabilidad de que la variable etiqueta tenga el valor de 1.
- Típicamente escogemos un umbral: si proba > 0.5 => y=1

Regresión logística



- Para solucionar el problema de proabilidades por encima y por debajo de los límites podemos usar una función logística en lugar de una línea.
- El procedimiento es similar pero ajustamos la forma de la curva en lugar de la forma de la línea.

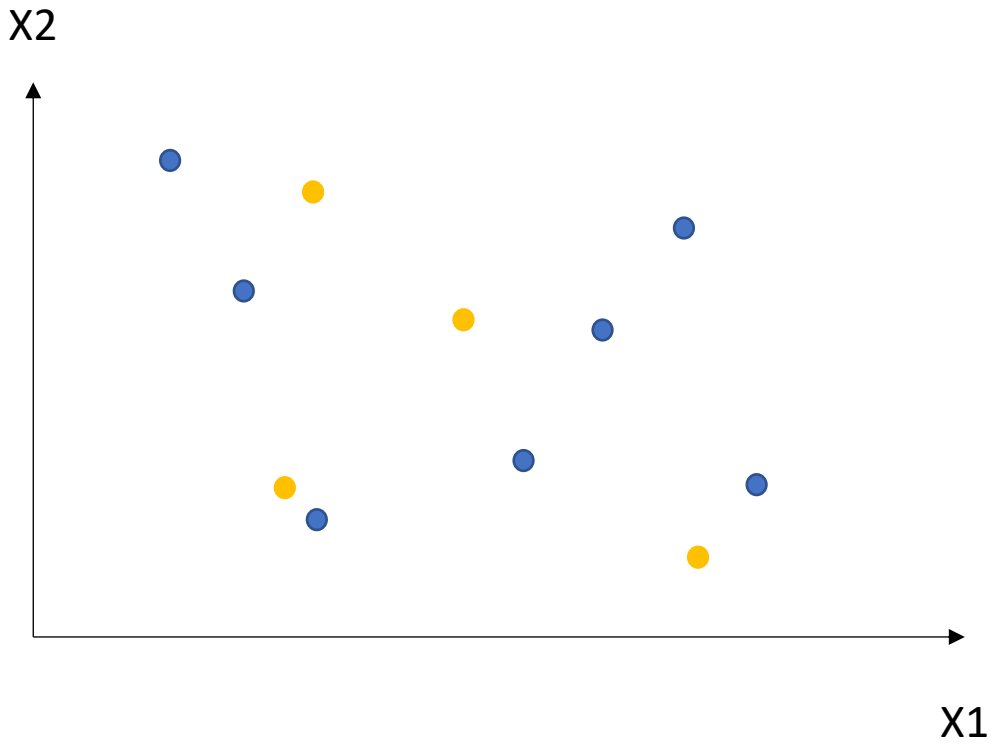
Regresión logística



- Para solucionar el problema de proabilidades por encima y por debajo de los límites podemos usar una función logística en lugar de una línea.
- El procedimiento es similar pero ajustamos la forma de la curva en lugar de la forma de la línea.
- Note que en ocasiones no aporta mucho cuando predecir es la prioridad.
- Igual: podemos usar más de una X.

K-vecinos más cercanos

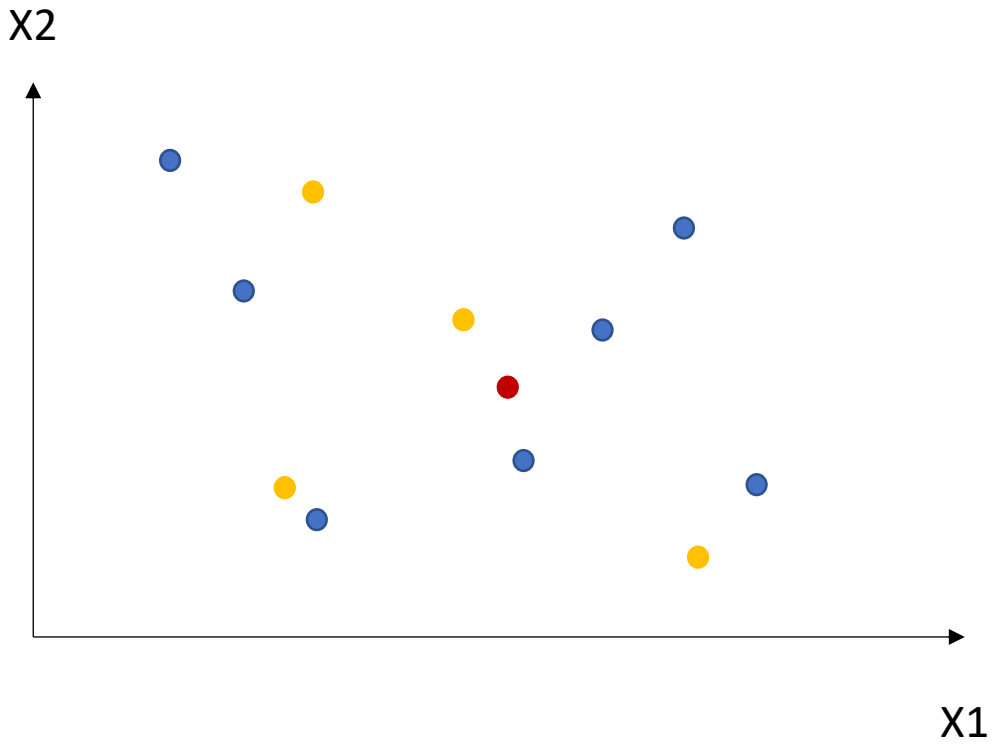
El primer algoritmo



- K vecinos más cercanos es un ejemplo de un algoritmo en lugar de un procedimiento mátemático.
- Graficamos las observaciones en un espacio de características.
- Es posible calcularlo para más de 3 características (aunque no dibujarlo).

K-vecinos más cercanos

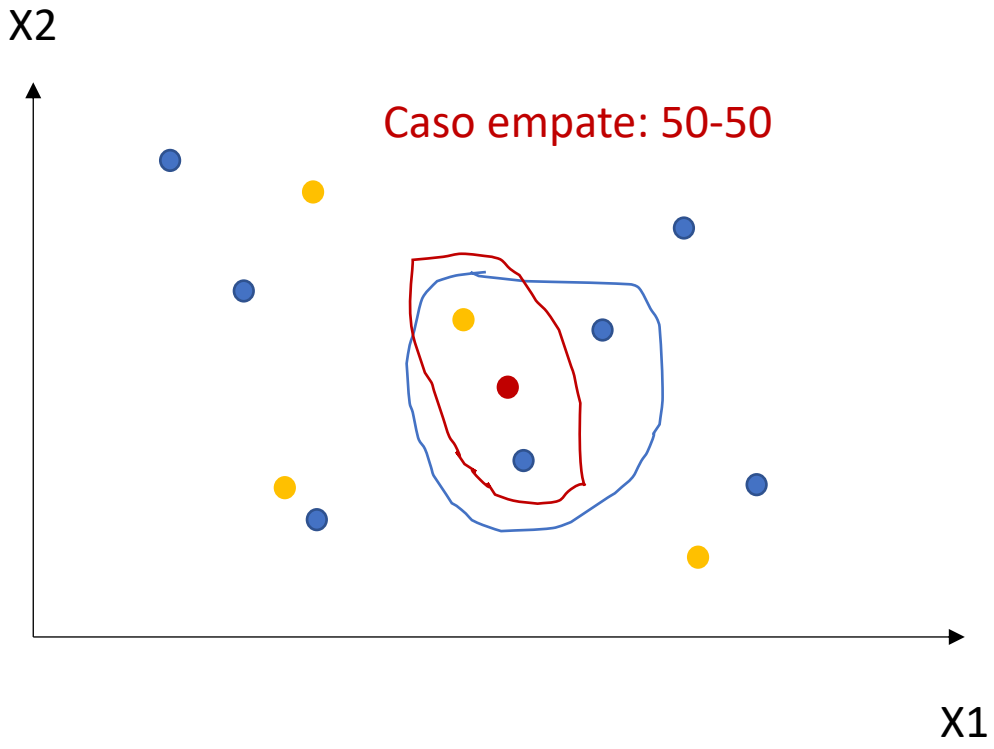
El primer algoritmo



- K vecinos más cercanos es un ejemplo de un algoritmo en lugar de un procedimiento mátemático.
- Graficamos las observaciones en un espacio de características.
- Es posible calcularlo para más de 3 características (aunque no dibujarlo).
- El nuevo dato se ubica en el espacio y se identifican los (1, 2, …) vecinos más cercanos.

K-vecinos más cercanos

El primer algoritmo

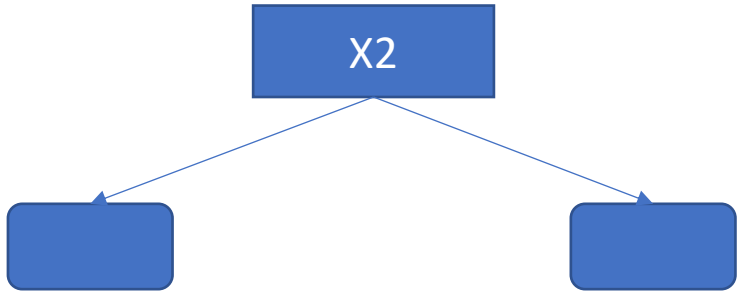


- K vecinos más cercanos es un ejemplo de un algoritmo en lugar de un procedimiento mátemático.
- Graficamos las observaciones en un espacio de características.
- Es posible calcularlo para más de 3 características (aunque no dibujarlo).
- El nuevo dato se ubica en el espacio y se identifican los (1, 2, …) vecinos más cercanos.
- Se ajusta el K en el entrenamiento.

Árboles de decisión

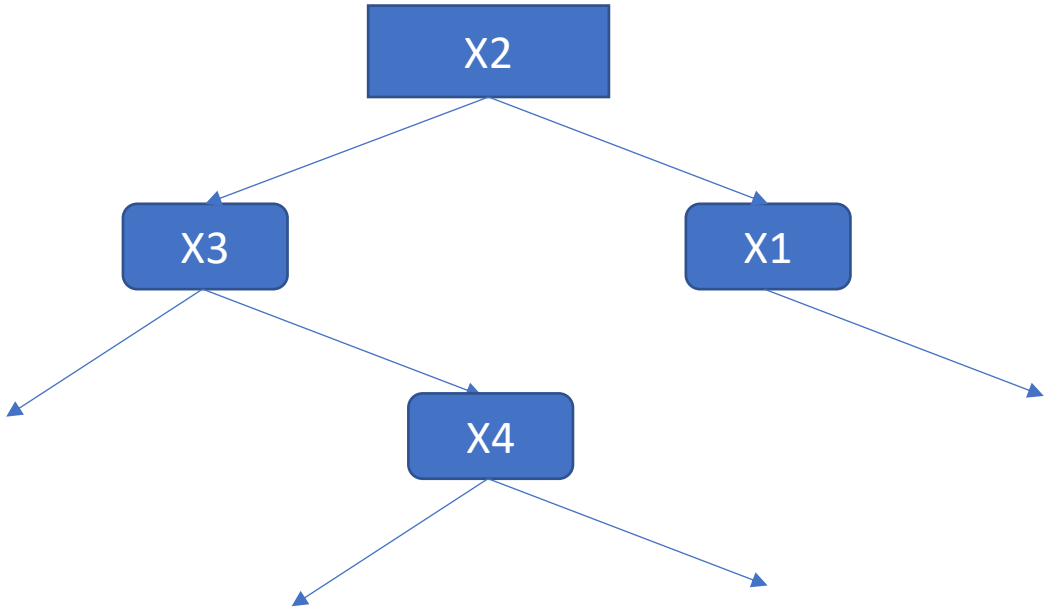
- Se toma la característica que más correlación tiene con la variable etiqueta.
- Esa es la raíz.

Árboles de decisión



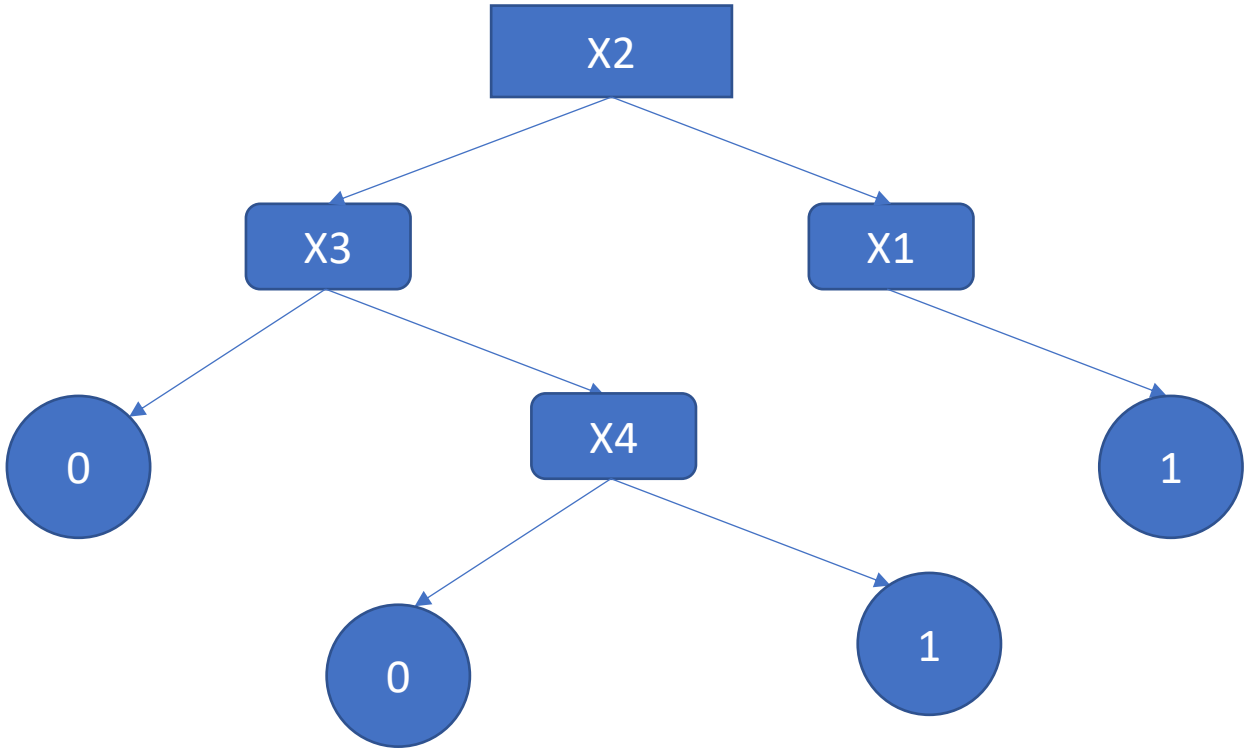
- Se toma la característica que más correlación tiene con la variable etiqueta.
- Esa es la raíz.
- De la raíz se hacen ramas para cada caso (intervalos, valores…)

Árboles de decisión



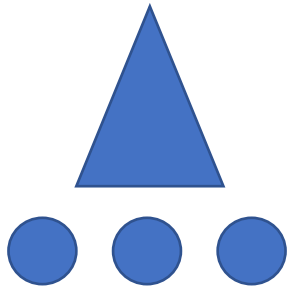
- Se toma la característica que más correlación tiene con la variable etiqueta.
- Esa es la raíz.
- De la raíz se hacen ramas para cada caso (intervalos, valores…)
- A partir de ahí se van haciendo consultas sobre otras variables características

Árboles de decisión



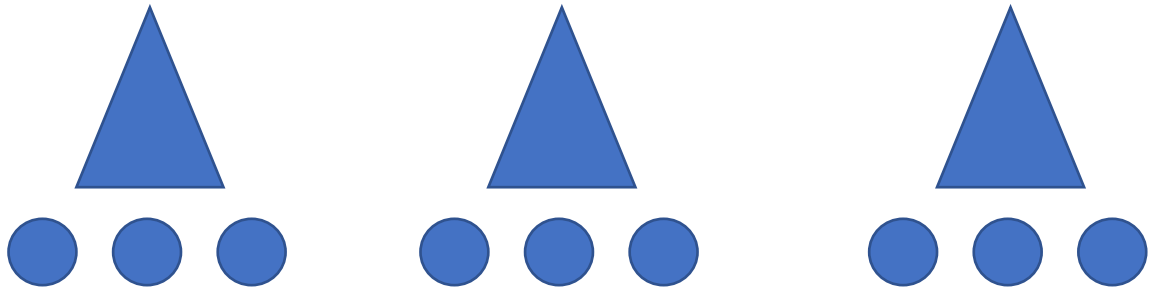
- Se toma la característica que más correlación tiene con la variable etiqueta.
- Esa es la raíz.
- De la raíz se hacen ramas para cada caso (intervalos, valores…)
- A partir de ahí se van haciendo consultas sobre otras variables características
- Los finales se llaman hojas.
- Se ajustan los parámetros de las preguntas.

Bosques aleatorios



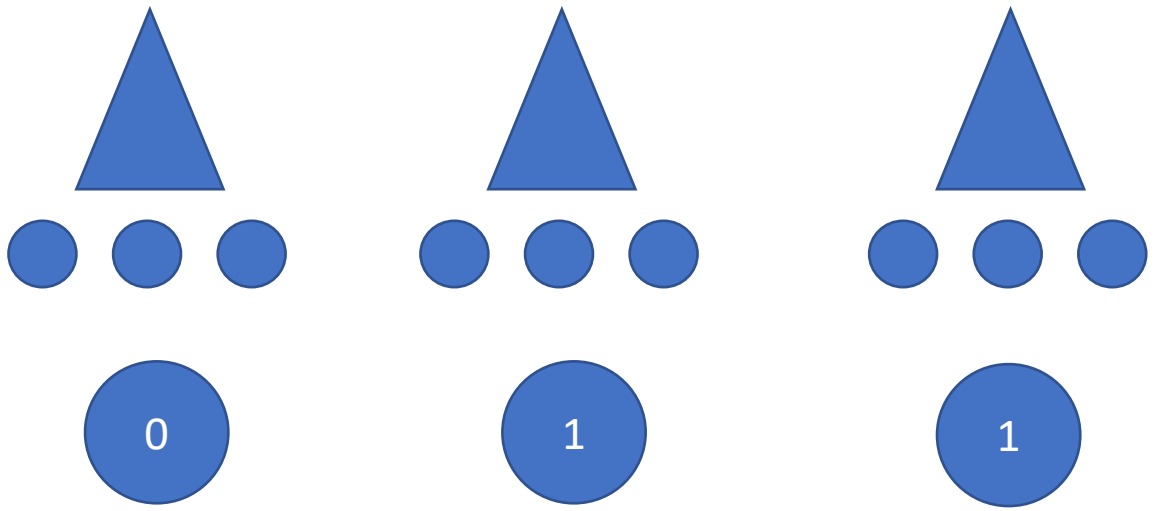
- Es la misma noción de árboles de clasificación.

Bosques aleatorios



- Es la misma noción de árboles de clasificación.
- Pero se seleccionan observaciones aleatorias.
- Para cada una se arma un arbol…

Bosques aleatorios



- Es la misma noción de árboles de clasificación.
- Pero se seleccionan observaciones aleatorias.
- Para cada una se arma un arbol…
- Y se clasifica por votación.
- Usualmente iniciamos por árboles de decisión individuales y exploramos con bosques aleatorios para ver si mejora el desempeño.

La nota es:

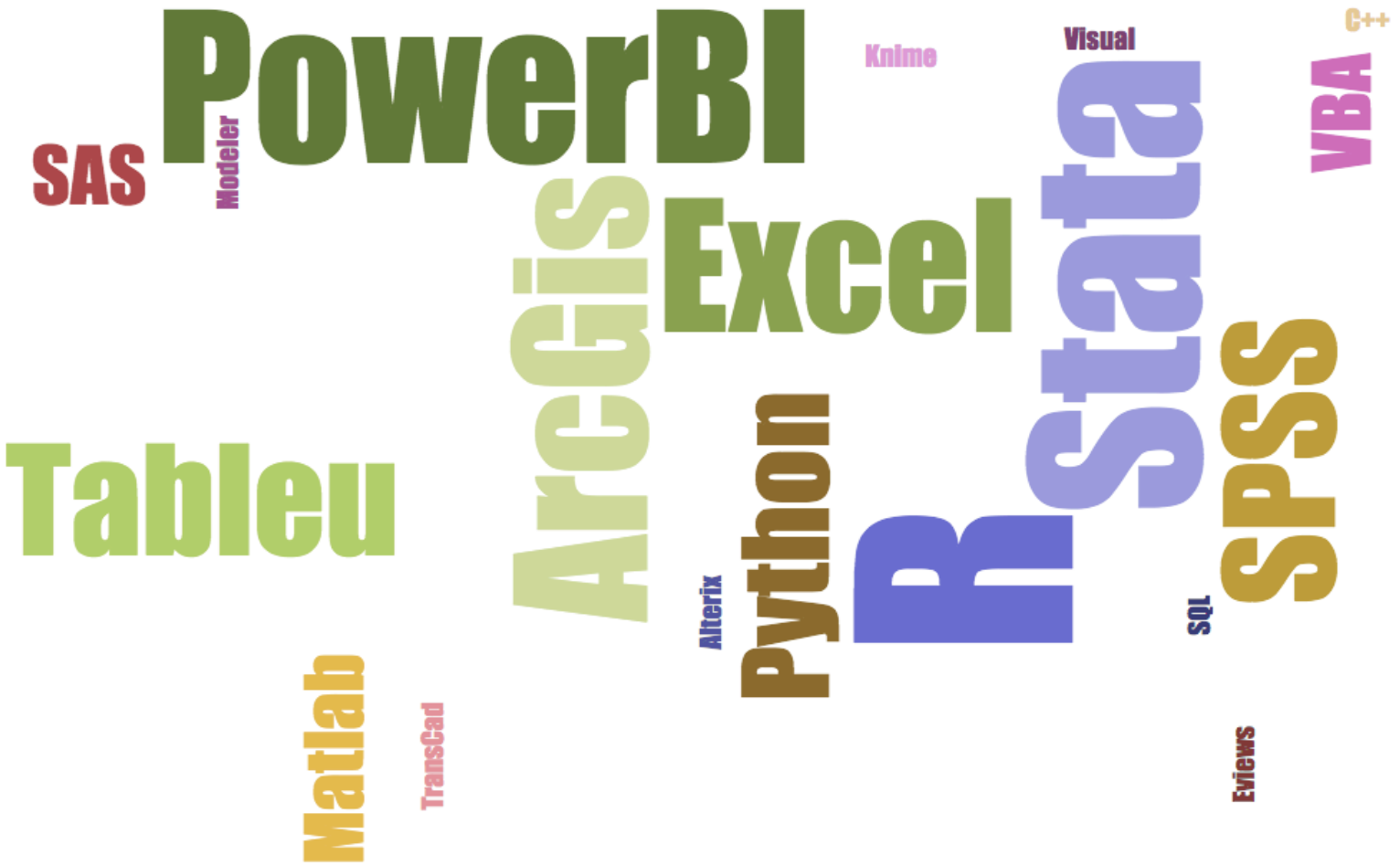
- Existen muchas otras técnicas pero estas capturan algo de la esencia
- También existen formas para mejorar estas técnicas
  - Regularización
  - Boosting
  - Validación cruzada

  (Vamos a ver algo de esto en la última clase)

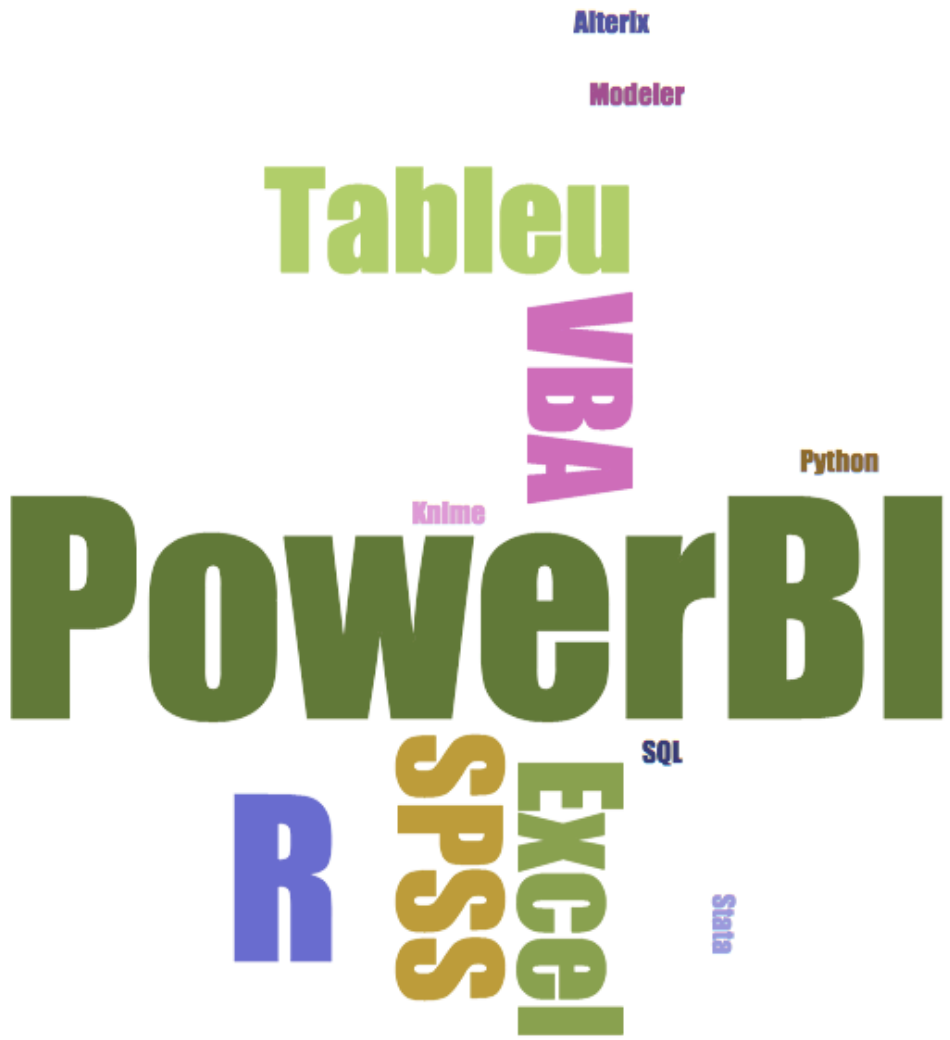
Corriendo un modelo en R

De principio a fin

De qué formas podemos trabajar con datos



SECTOR PRIVADO



SECTOR PÚBLICO

De qué formas podemos trabajar con datos

- Para análisis descriptivos es común usar directamente programas de análisis (Stata, Excel, PowerBI…)
- Para modelos orientados al aprendizaje de máquinas solemos usar herramientas programables (R, Python, etc.)

Lenguaje + Entorno de desarrollo

Etapas de todo proceso de ML (de la clase pasada)

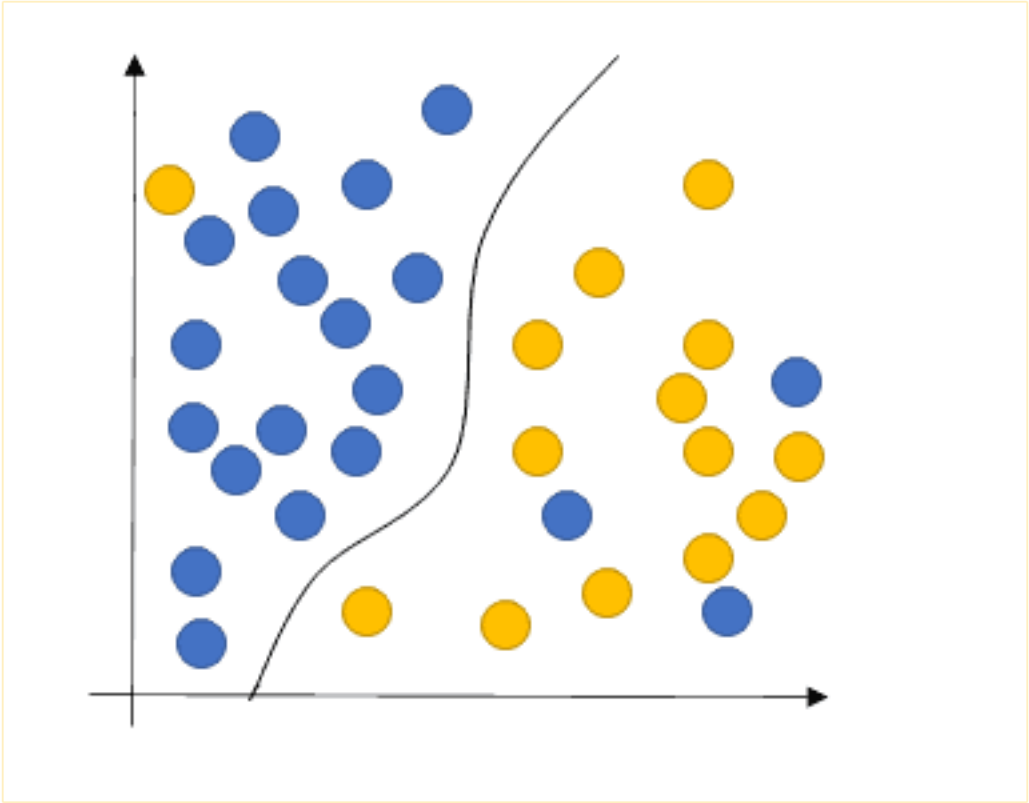
1. Limpieza OK
2. Exploración descriptiva OK
3. Agrupación: datos de entrenamiento y datos de prueba* OK
4. Entrenamiento
5. Evaluación*

* Veremos en más detalle en un momento o al final del curso

Photo by Christina Morillo from Pexels

Agrupación: el problema del sobre-ajuste

- Siempre hay que distribuir nuestros datos en grupos: (ej.) *train, test*
- Entrenamos nuestros modelos con los datos de entrenamiento (*train*)
- Luego revisamos que los errores de predicción no incrementen mucho en datos nuevos (*test*).



Imágen: Just into Data

Entrenamiento

- En la etapa de entrenamiento, nuestro modelo prueba diferentes parámetros a partir de los datos que le damos.

$$y = b + mx$$

¡Recuerden usar sólo los datos *train*!

...

Entrenamiento

- En la etapa de entrenamiento, nuestro modelo prueba diferentes parámetros a partir de los datos que le damos.

¡Recuerden usar sólo los datos *train*!

$$y = 5 + 6x$$

$$y = 5 + 6.1x$$

$$y = 4 + 6.6x \quad \ldots$$

$$y = -4 + 2.3x$$

Entrenamiento

- Entonces… cuando llega un nuevo dato con valor de x=2

La máquina nos da una predicción:

$$y = -4 + 2.3x$$

$$1 \longleftarrow 0.6 = -4 + 2.3(2)$$

Evaluación

- Al evaluar queremos ver qué tanto nuestro modelo se equivoca al predecir.
- Con datos conocidos…
- Con datos nuevos…
- Qué tipo de errores comete más…

Veremos una clase sobre esto.

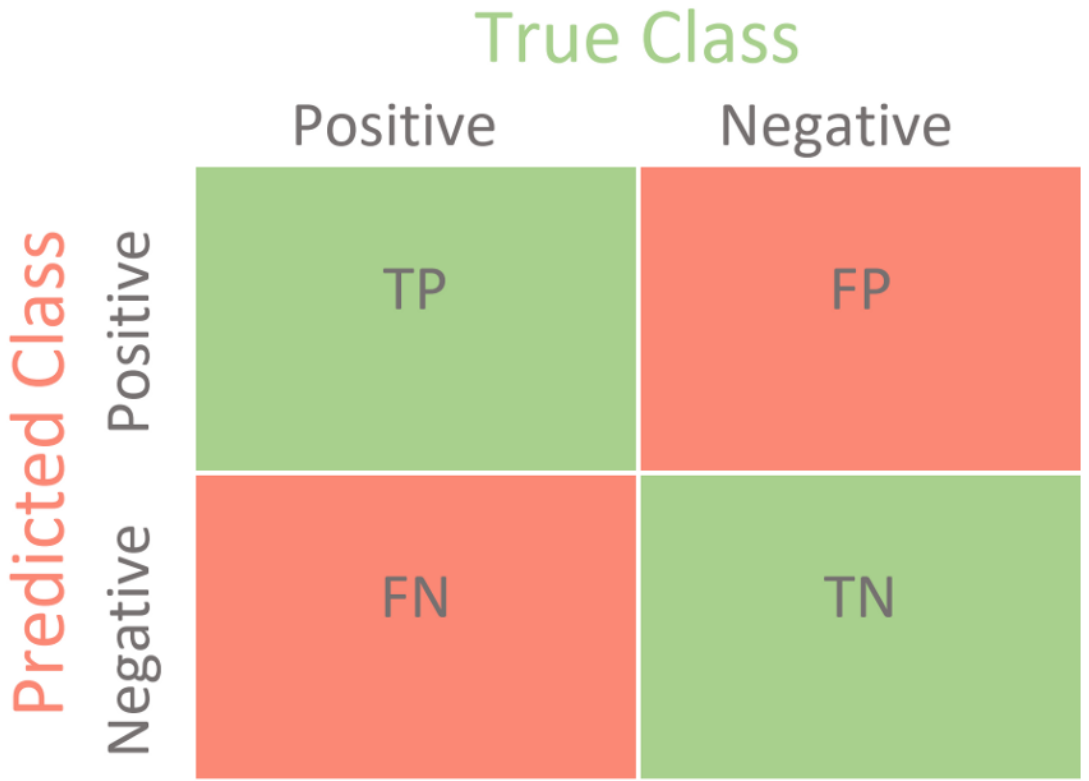


Imagen: Towards Data Science

Manos al código

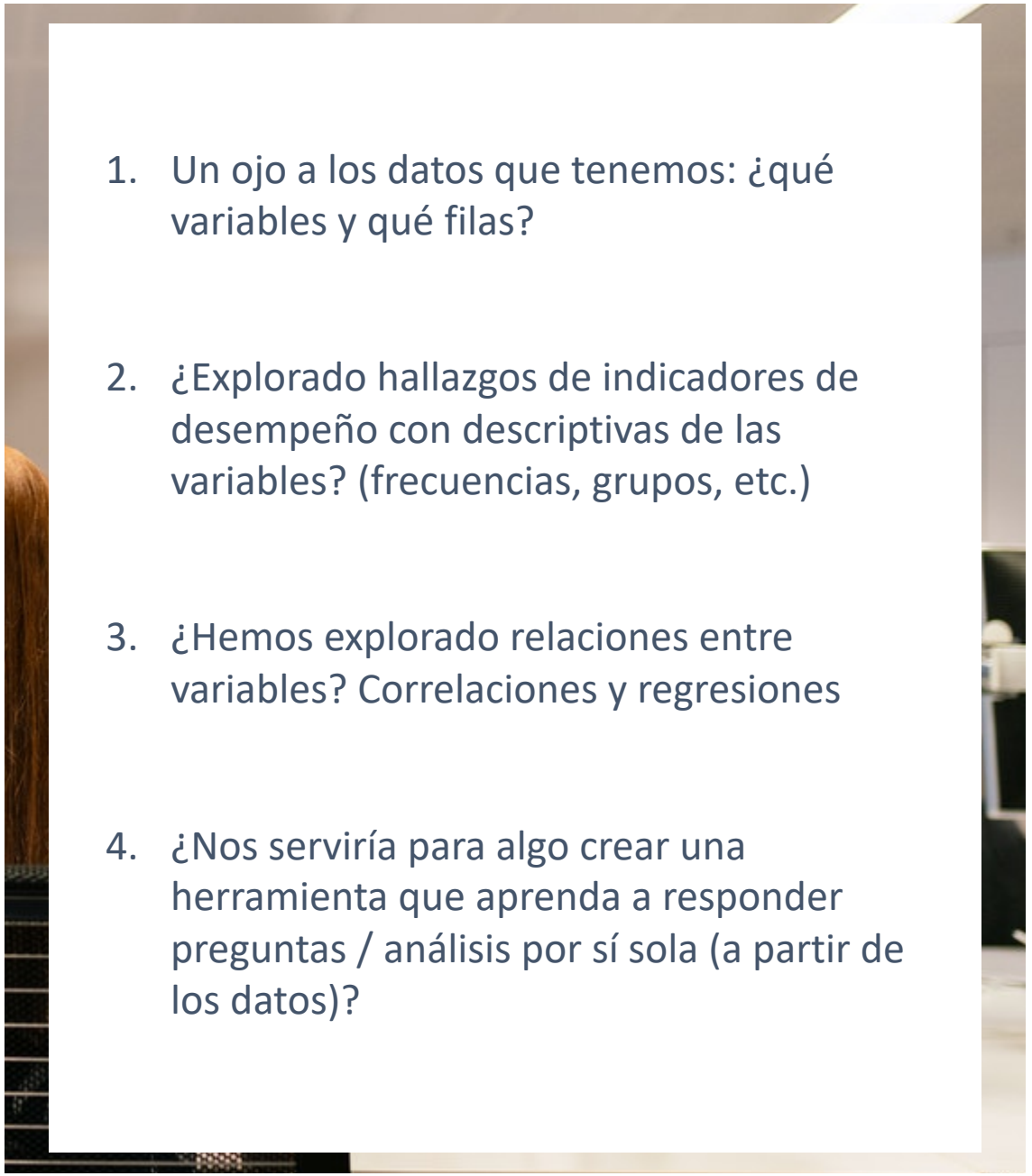
Veamos un código de ML de principio a fin

Taller diferenciado

Por grupos

Una lista de tareas:

-
-
-
-
-
-



1. Un ojo a los datos que tenemos: ¿qué variables y qué filas?
2. ¿Explorado hallazgos de indicadores de desempeño con descriptivas de las variables? (frecuencias, grupos, etc.)
3. ¿Hemos explorado relaciones entre variables? Correlaciones y regresiones
4. ¿Nos serviría para algo crear una herramienta que aprenda a responder preguntas / análisis por sí sola (a partir de los datos)?

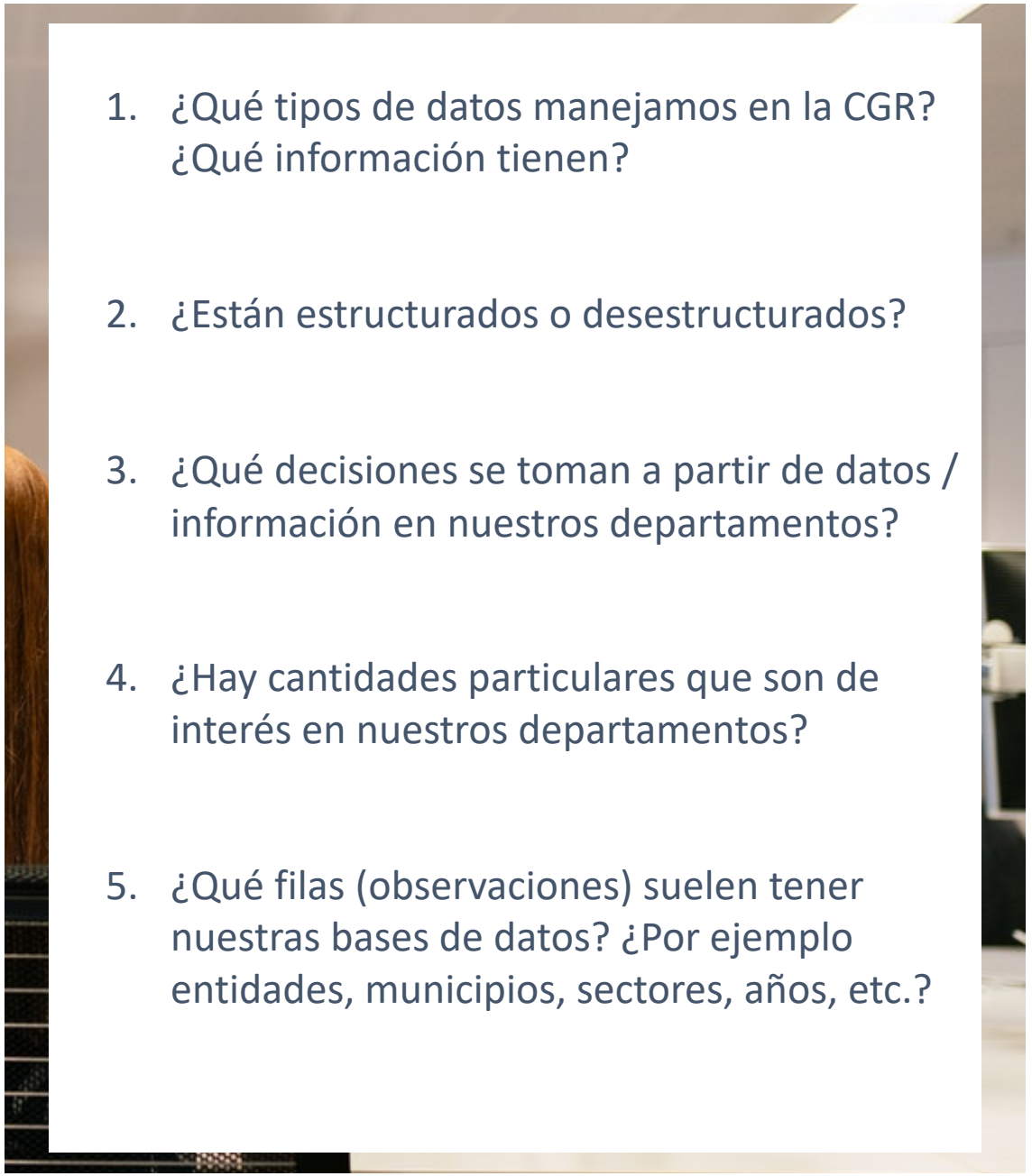
Retos potenciales:

Primera exploración

Clasificación

Regresión

No supervisado



1. ¿Qué tipos de datos manejamos en la CGR? ¿Qué información tienen?
2. ¿Están estructurados o desestructurados?
3. ¿Qué decisiones se toman a partir de datos / información en nuestros departamentos?
4. ¿Hay cantidades particulares que son de interés en nuestros departamentos?
5. ¿Qué filas (observaciones) suelen tener nuestras bases de datos? ¿Por ejemplo entidades, municipios, sectores, años, etc.?