ACÁMICA

# ¡Bienvenidos/as a Data Science!

# Agenda

¿Cómo anduvieron?

Repaso

Actividad

Break

Aplicaciones Prácticas

Explicación: ¿Qué es aprender en Machine Learning?

Cierre

# ¿Cómo anduvieron?

# Repaso

# Matriz de confusión

| | PREDIJO SI | PREDIJO NO |
|---|---|---|
| REAL SI | | |
| REAL NO | | |

# Matriz de confusión

| | PREDIJO SI | PREDIJO NO |
|---|---|---|
| REAL SI | VERDADERO POSITIVO | FALSO POSITIVO |
| REAL NO | FALSO NEGATIVO | VERDADERO NEGATIVO |

# Matriz de confusión

- Tenemos un dataset con datos de salud de 200 personas (saludables y enfermos).
- Analizamos el resultado corriendo el modelo con una muestra de 200 individuos, de los cuales 90 están enfermos y 110 están sanos.
- El modelo acertó en 80 enfermos y en 100 sanos.
- La matriz de confusión quedaría de la siguiente manera:

| | PREDIJO SI | PREDIJO NO |
|---|---|---|
| REAL SI | 80 | 10 |
| REAL NO | 10 | 100 |

# Matriz de confusión

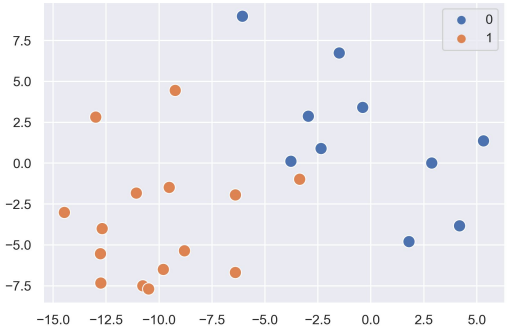
| | PREDIJO SI | PREDIJO NO |
|---|---|---|
| REAL SI | 80 | 10 |
| REAL NO | 10 | 100 |

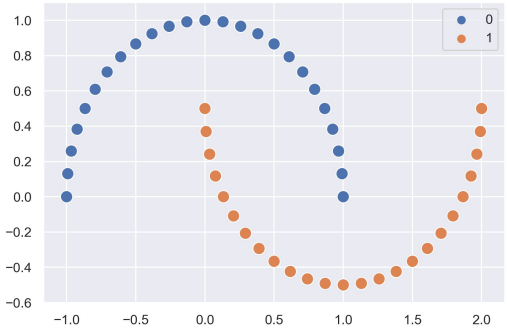
Por ejemplo, usando estos datos podemos calcular la exactitud del modelo (en inglés Accuracy). Eso lo hacemos dividiendo el total de aciertos sobre el total de ejemplos. En nuestro caso sería 180 sobre 200. Eso nos dá 0,9.

Podemos calcular también la precisión que se hace dividiendo la cantidad de verdaderos positivos sobre la suma de verdaderos positivos y falsos positivos. Esto en nuestro caso sería 80 sobre 80 mas 10. Y nos dá 0,88.

# Actividad en equipos (de 4 o 5 personas cada uno)

1. ¿Cuál es la diferencia entre aprendizaje supervisado y aprendizaje no-supervisado?
2. ¿Cuál es la diferencia entre un problema de clasificación y uno de regresión?
3. ¿Qué es el sobreajuste?
4. ¿Por qué es importante separar una porción del dataset antes de entrenar un modelo?
5. Describir el proceso por el cual un árbol de decisión *aprende* de los datos usando impureza Gini.
6. Definir Falsos Positivos, Falsos Negativos, Precisión y Exhaustividad.
7. Dibujar, aproximadamente, las fronteras de decisión que obtendrían con un árbol de decisión de profundidad uno, un árbol de decisión de profundidad dos, KNN con *k=1* y KNN con *k=número de muestras* en los siguientes casos:




*Además, elegir un caso y uno de los modelos y calcular cómo quedaría la matriz de confusión.

# Actividad integradora

# Actividad

Apliquen lo aprendido al dataset que eligieron. Reflexionen sobre:

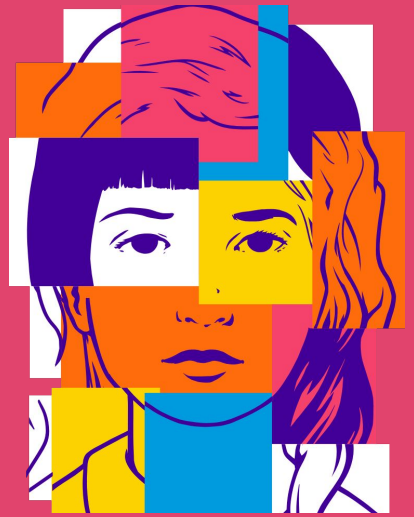
1. ¿Qué métrica usarían para evaluar su desempeño?
2. ¿Cuál modelo se desempeña mejor?
3. ¿Están sobreajustando? Si es el caso, ¿cómo lo evitarían?

# ¡BREAK!

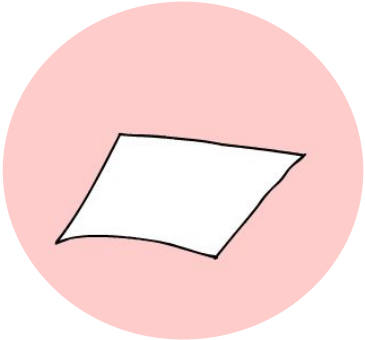
Ph. Credit: Drew Coffmann

# Aplicaciones prácticas

# ¿Qué es *aprender* en Machine Learning?



# Ahora que ya conocemos algunos modelos, tratemos de darles un marco más general.[1]

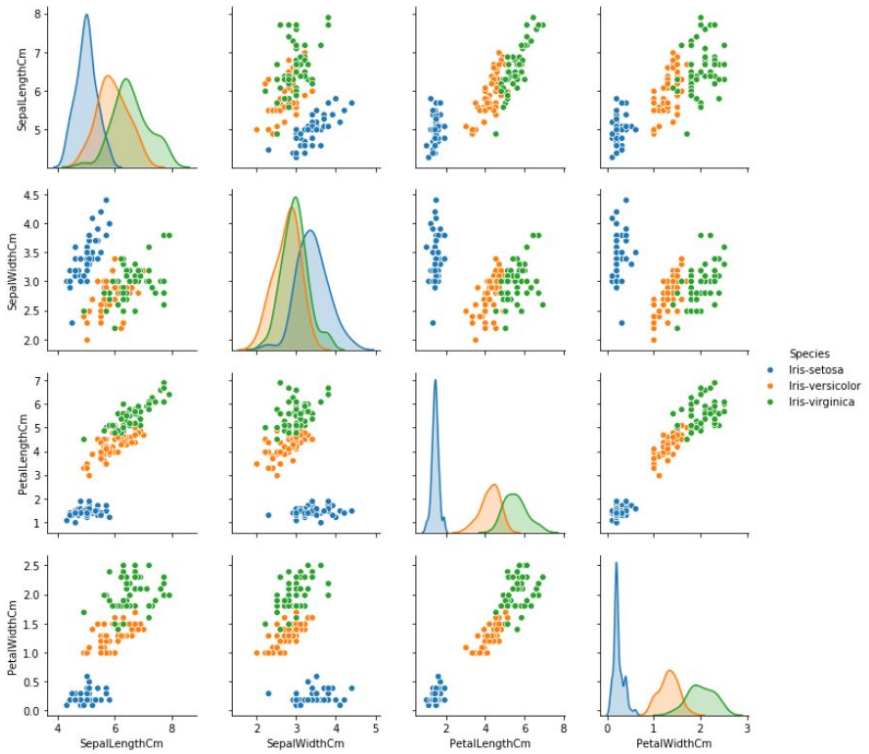


[1] Algunas ideas para esta clase fueron tomadas de la materia Aprendizaje Automático, dictada por Agustín Gravano y Pablo Brusco en FCEN, UBA.

# Aprendizaje inductivo

En el Iris Dataset, vemos que algunas características - atributos - de las flores sirven para diferenciar las especies.

Si consideramos sólo dos atributos, podemos ver que existen regiones en el plano que, aproximadamente, corresponden a cada especie.



# Aprendizaje inductivo

La idea subyacente es que podemos aprender a diferenciar cada especie a partir de medir algunas de sus características[1].

Más general, podemos aprender conceptos a partir de un conjunto de ejemplos y sus características.

[1] Desde otro punto de vista, si esas características que medimos nos sirven para diferenciar, tal vez sean buenas características para definir esas especies.

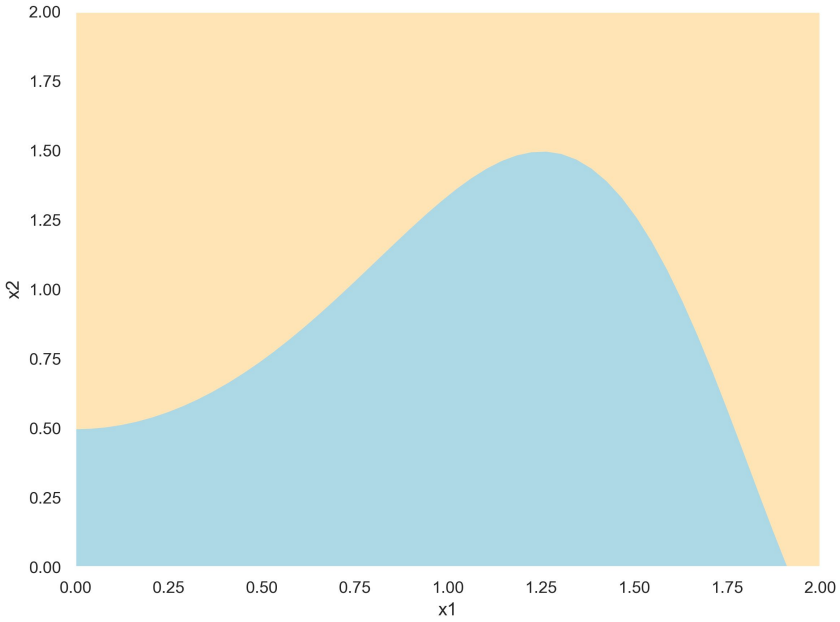
# Aprendizaje inductivo

Usando esas características (o propiedades) que medimos, asumimos que existe una frontera que separa las clases que queremos aprender.

# Aprendizaje inductivo

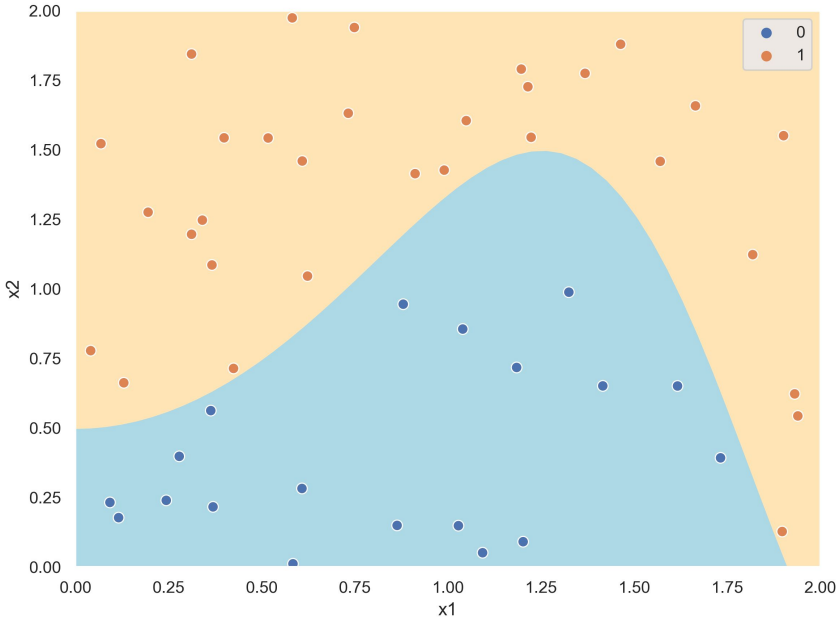
Veamos un ejemplo de juguete.

Queremos aprender a qué regiones corresponde el color amarillo y a qué regiones corresponde el color celeste. Es decir, **f(x1, x2) —> color**



# Aprendizaje inductivo

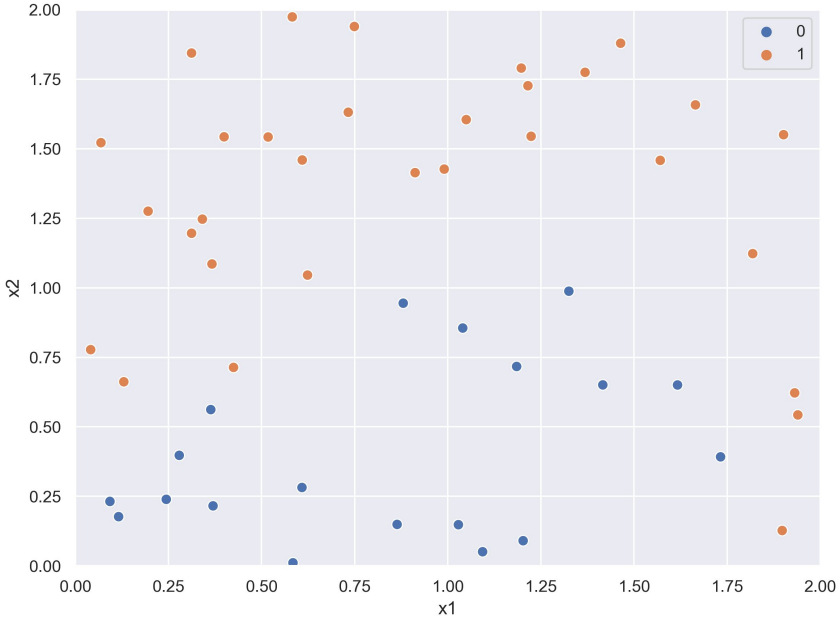
Pero no tenemos las regiones pintadas, sino muestras.



# Aprendizaje inductivo

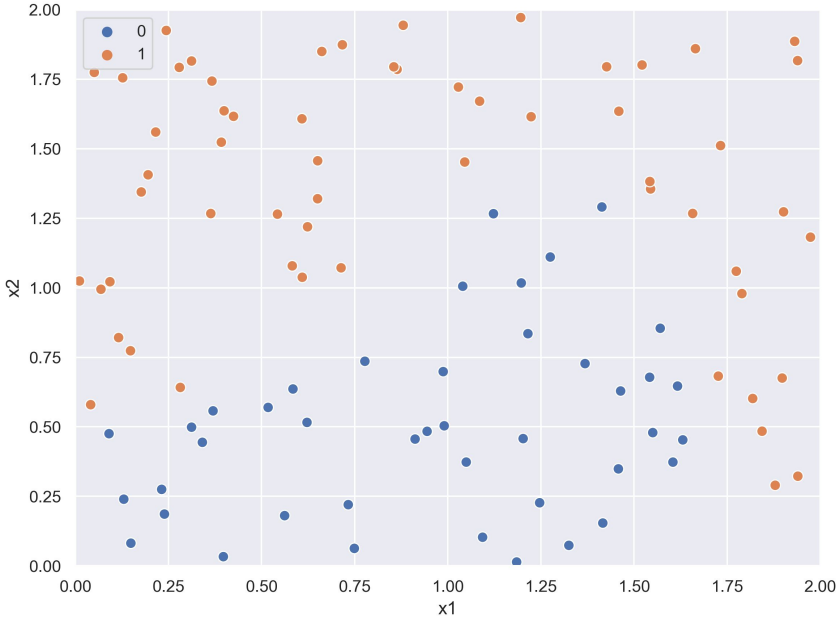
Pero no tenemos las regiones pintadas, sino muestras.

Muestras sobre un fondo sin color.



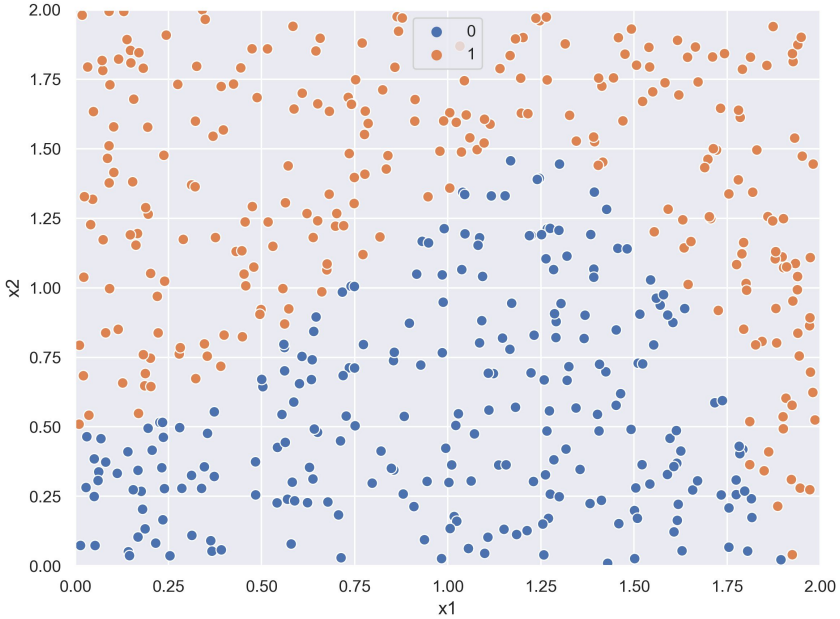
# Aprendizaje inductivo

Cuantas más muestras,
probablemente sea más fácil la tarea.



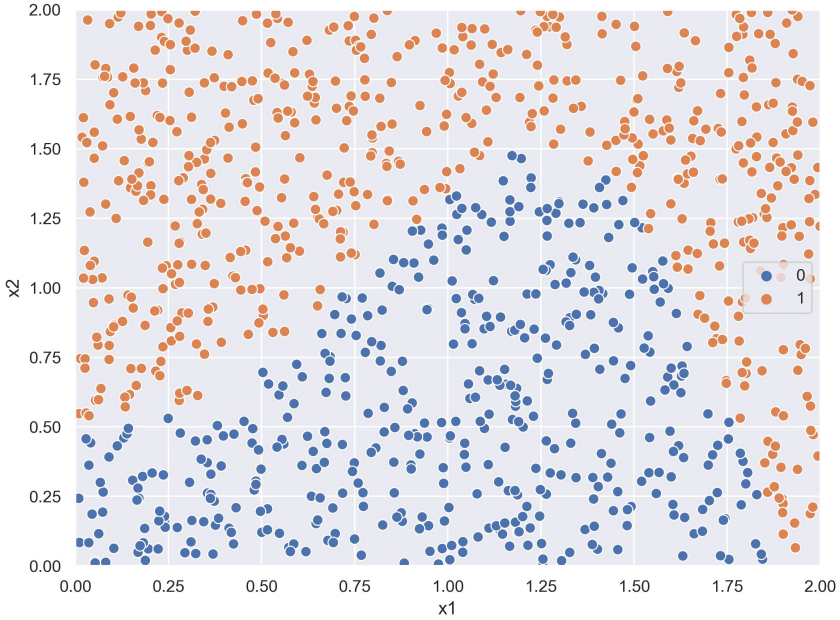
# Aprendizaje inductivo

Cuantas más muestras,
probablemente sea más fácil la tarea.



# Aprendizaje inductivo

Cuantas más muestras,
probablemente sea más fácil la tarea.



El aprendizaje inductivo consiste en construir un modelo general (color del fondo) a partir de información específica (instancias).

# Principio de Aprendizaje Inductivo

Cualquier modelo que aproxime bien a una función objetivo sobre un conjunto suficientemente grande de datos también aproximará bien a la función objetivo sobre datos no observados.

# Aprendizaje supervisado

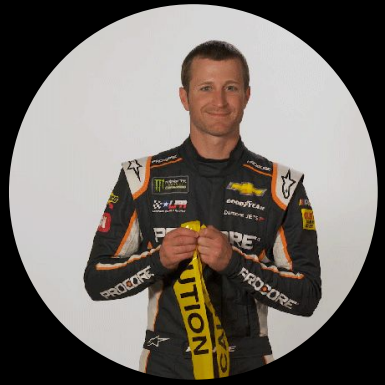
Dada una función objetivo ***f*** desconocida, queremos aproximarla mediante un modelo, que se suele notar **f^**[1].

**Entrenar un modelo** —*consiste en*→ ajustar sus parámetros (encontrar valores óptimos) dado un conjunto de datos.

Los algoritmos de aprendizaje automático son procedimientos para entrenar modelos a partir de un conjunto de datos

[1]El sombrerito va a arriba de la **f** pero no encontramos cómo escribirlo.

# Aprendizaje supervisado



**¡Modelo y algoritmo no son lo mismo!**

(aunque a veces los vamos a usar como sinónimos en un abuso de lenguaje).

# Sesgo inductivo

Pero hay muchas formas para construir un modelo. Por ejemplo, árboles de decisión y vecinos más cercanos aprenden de los datos de formas distintas.

# Sesgo inductivo

Pero hay muchas formas para construir un modelo. Por ejemplo, árboles de decisión y vecinos más cercanos aprenden de los datos de formas distintas.

El **sesgo inductivo** de un algoritmo de aprendizaje es el conjunto de afirmaciones que el algoritmo utiliza para construir un modelo

# Sesgo inductivo

Pero hay muchas formas para construir un modelo. Por ejemplo, árboles de decisión y vecinos más cercanos aprenden de los datos de formas distintas.

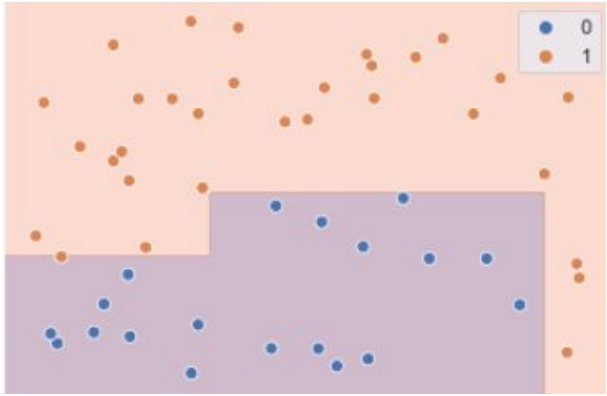
El **sesgo inductivo** de un algoritmo de aprendizaje es el conjunto de afirmaciones que el algoritmo utiliza para construir un modelo

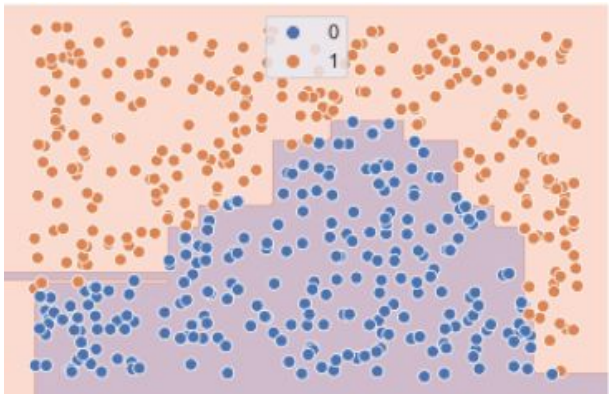
**Incluye:**

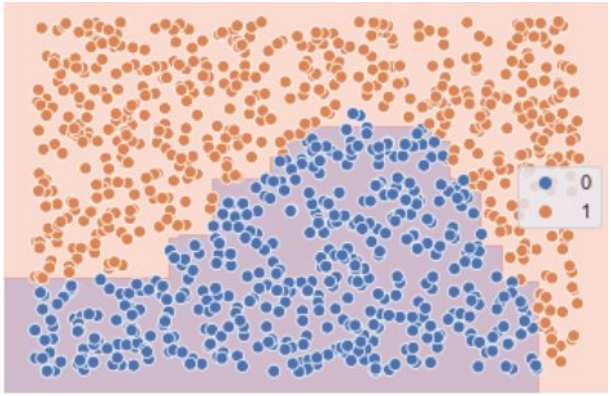
- Forma de las hipótesis (número y tipo de parámetros)
- Características del funcionamiento del algoritmo (cómo recorre el espacio de hipótesis para elegir un único modelo).

# Sesgo inductivo

Veamos algunos ejemplos. ¿Qué ocurre si ajustamos un árbol de decisión al ejemplo que inventamos?



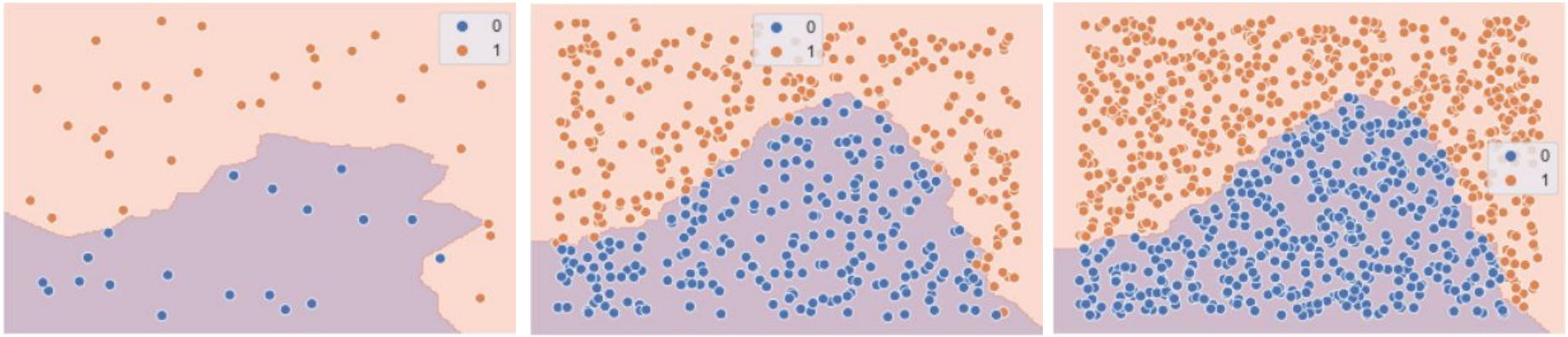




El espacio de hipótesis de un árbol de decisión sólo permite fronteras formadas por rectas horizontales y verticales. Podemos modificar estas fronteras variando la profundidad del árbol o la cantidad de instancias.

# Sesgo inductivo

¿Y si ajustamos un modelo de vecinos más cercanos?




El espacio de hipótesis de vecinos más cercanos permite fronteras más versátiles que los árboles (¡esto no significa que sea un mejor modelo!). Y podemos modificar estas fronteras variando la cantidad de vecinos (k) o la cantidad de instancias.

# Navaja de Ockham

En igualdad de condiciones (por ejemplo, igual desempeño), elegir la explicación (modelo) más simple.

## ¿Por qué? Simplemente porque esperamos que generalice mejor.

Es un principio metodológico.

# Para la próxima

1. Ver los videos de la plataforma “Machine Learning: Algoritmos de Regresión”, “Machine Learning: Regresión lineal” y “Validación y Testeo de modelos: Cross Validation”

2. Leer la consigna de la Entrega 03

3. Trabajar en notebooks atrasados o continuar trabajando en su dataset.

ACÁMICA