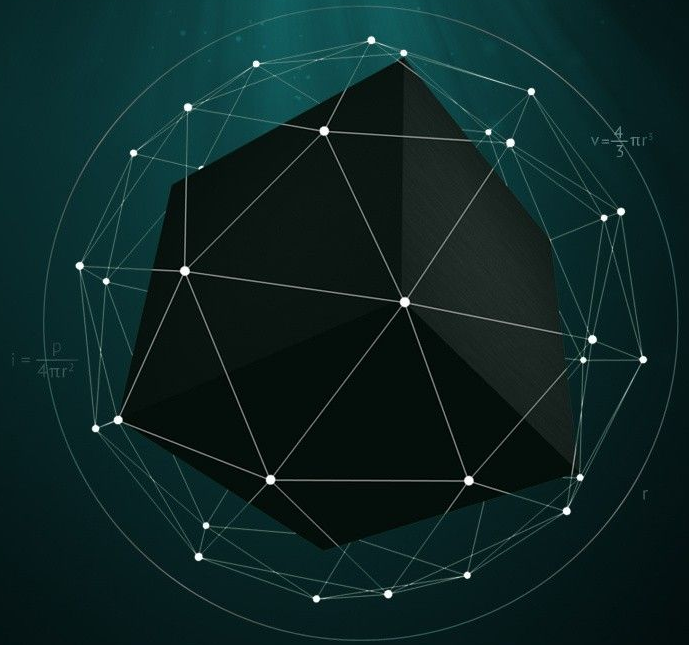
Ciencia de Datos con Python

Una breve introducción



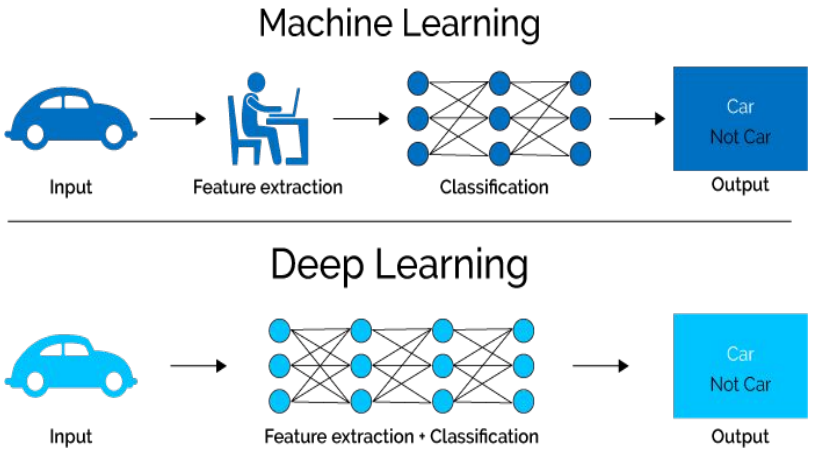
by Walter Casas

Qué es Machine Learning?

“Dar la habilidad a las computadoras de aprender a tomar decisiones en base a la data, sin estar explícitamente programadas.”

Ejemplos:

- Aprender a predecir si un email es spam o no
- Agrupar las entradas de Wikipedia en diferentes categorías

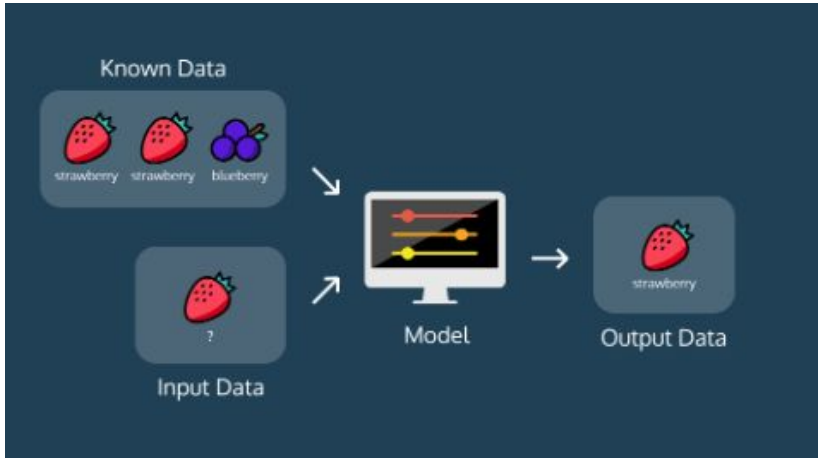


Aprendizaje Supervisado

La data es etiquetada y el programa aprende a predecir el output desde el input

Regresión: predecir valores continuos
- Precio de una casa en Rio
- Valor de las criptomonedas

Clasificación: predice valores discretos
- Esta pintura es de humano o de un AI?
- Este email es SPAM?

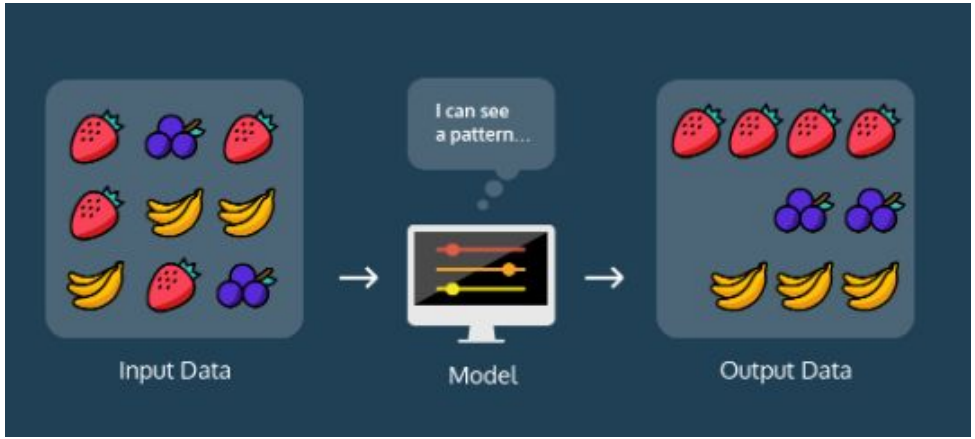


Aprendizaje No Supervisado

Descubre patrones ocultos en data no etiquetada.

Clustering: encuentra patrones y estructuras en data no etiquetada agrupando en clusters.

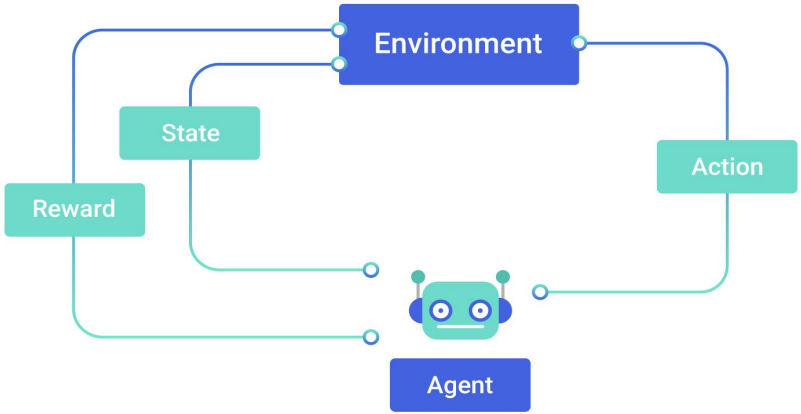
- Agrupa nuevos tópico en Redes Sociales (Twitter)
- Clusters de clientes para recomendación
- Search engine agrupan objetos similares en un cluster



Aprendizaje Reforzado

Software que interactúa con su entorno y aprende a cómo mejorarse. Tiene un sistema de premios y castigos.

- Economía
- Juegos (AlphaGo)



Pregunta

Cuál de estos problemas es un problema de aprendizaje supervisado de clasificación:

1. Usar data financiera etiquetada para predecir si el valor de un bien crecerá o disminuirá.
2. Usar data etiquetada de precios de vivienda para predecir el precio de la vivienda basado en sus características.
3. Usar data no etiquetada para agrupar a los estudiantes en diversas categorías para ofrecerles cursos.
4. Usar data financiera etiquetada para predecir el valor de un bien la próxima semana.

Aprendizaje Supervisado

Aprendizaje Supervisado

La meta del aprendizaje supervisado es:

- Automatizar el tiempo consumido o gasto por una tarea manual.
  - Ejemplo: Diagnósticos de doctores.
- Hacer predicciones del futuro
  - EJemplo: El cliente le dará clic al anuncio o no?
- Necesitas data etiquetada
  - Data histórica etiquetada
  - Experimentos para conseguir data etiquetada
  - Data etiquetada por Crowd-sourcing (ReCAPTCHA)

Aprendizaje Supervisado

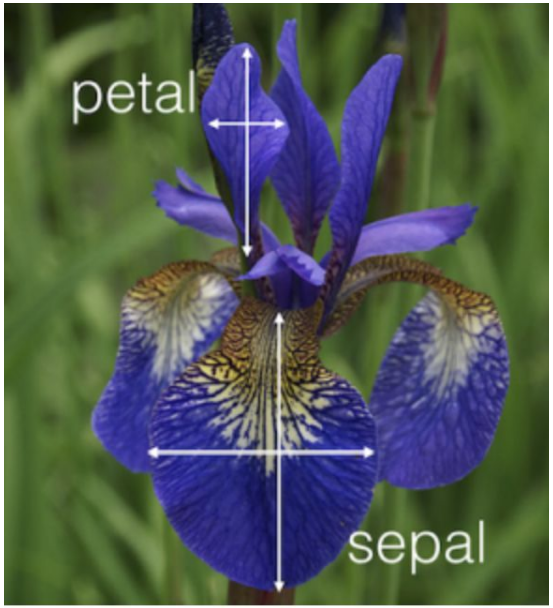
- Características = variables predictoras = variables independientes (Features)
- Variable objetivo = variable dependiente = variable respuesta (Target)

Predictor variables

Target variable

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

El dataset Iris



Características:

- Longitud del pétalo
- Ancho del pétalo
- Longitud del sépalo
- Ancho del sépalo

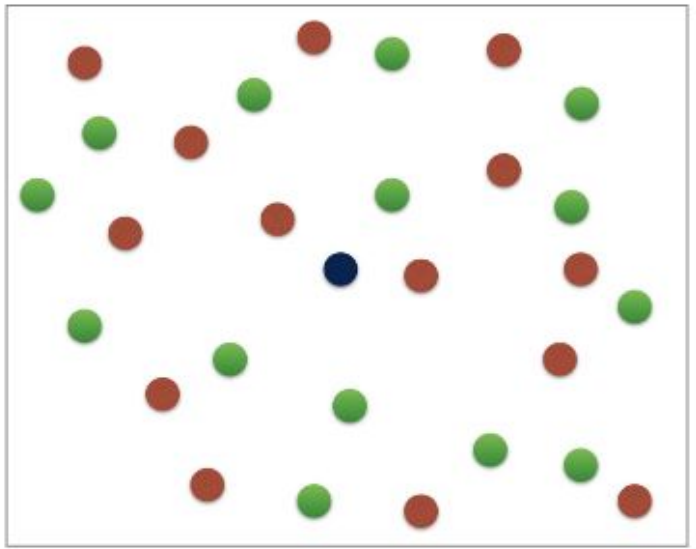
Variable objetivo: Especie

- Versicolor
- Virginica
- Setosa

k-Nearest Neighbors

Idea básica: predecir la etiqueta de un punto basado en:

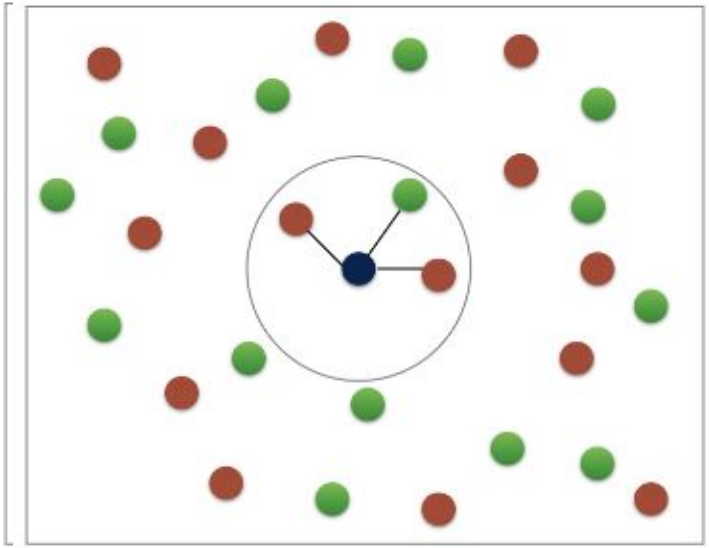
- Observación de los “k” puntos etiquetados más cercanos
- Tomar lo que dice la mayoría



k-Nearest Neighbors

Idea básica: predecir la etiqueta de un punto basado en:

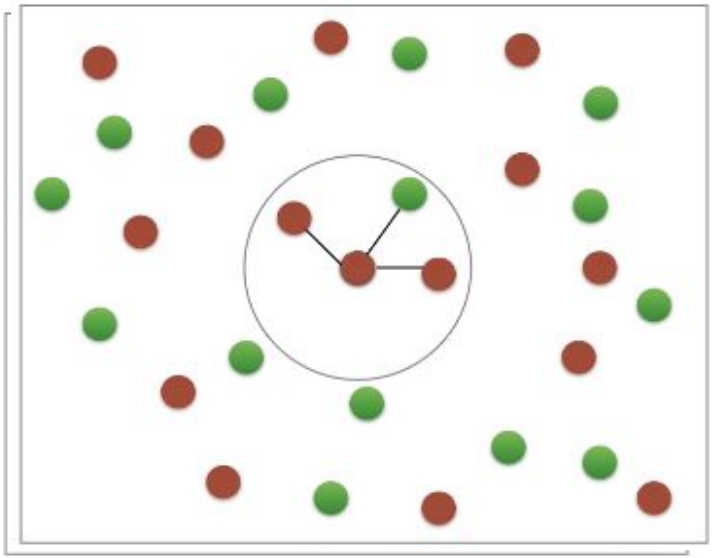
- Observación de los “k” puntos etiquetados más cercanos
- Tomar lo que dice la mayoría



k-Nearest Neighbors

Idea básica: predecir la etiqueta de un punto basado en:

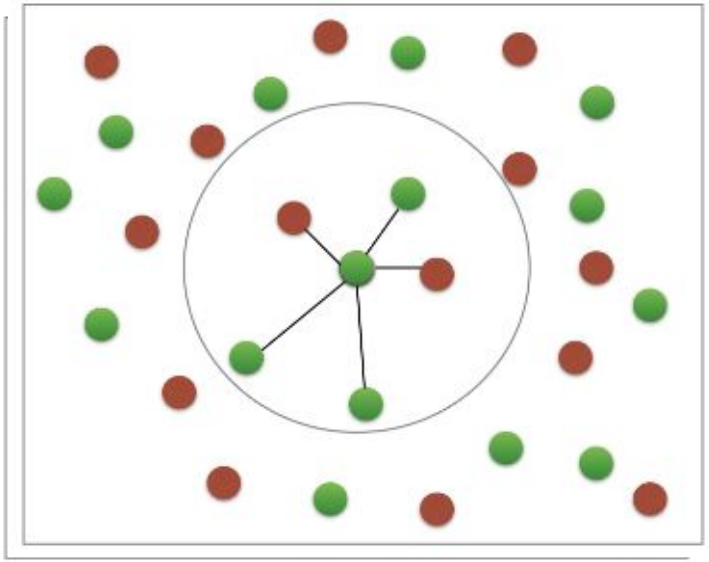
- Observación de los “k” puntos etiquetados más cercanos
- Tomar lo que dice la mayoría



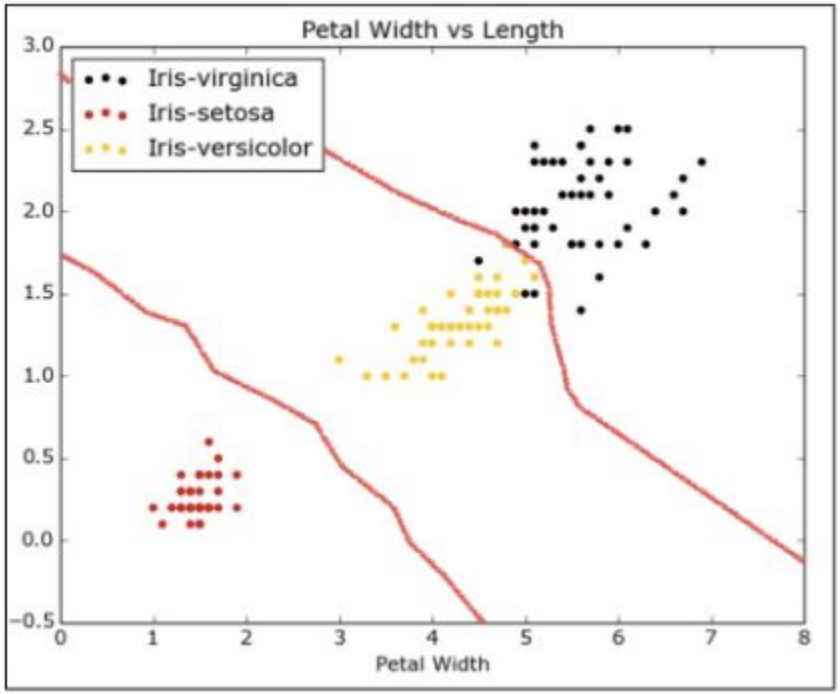
k-Nearest Neighbors

Idea básica: predecir la etiqueta de un punto basado en:

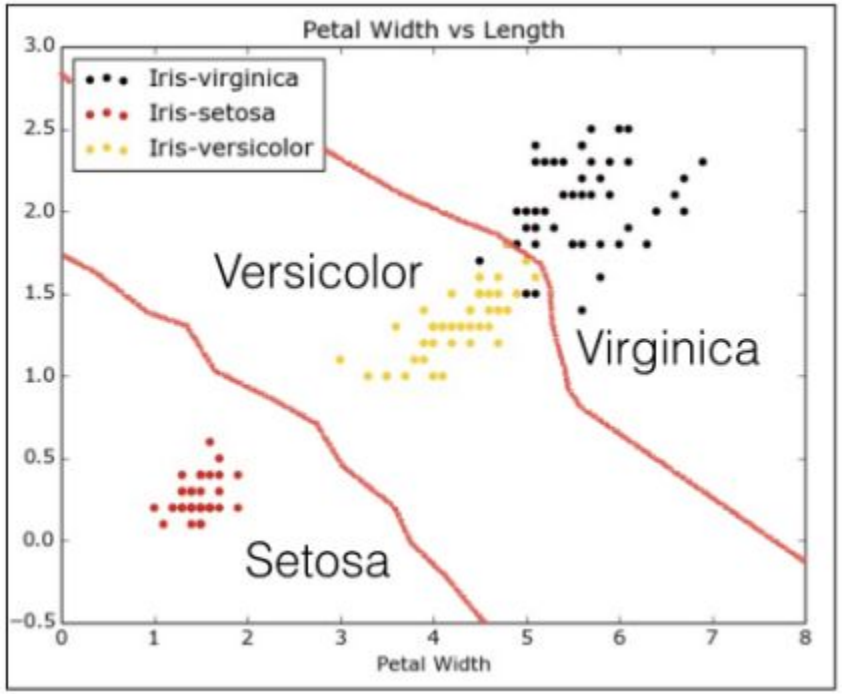
- Observación de los “k” puntos etiquetados más cercanos
- Tomar lo que dice la mayoría



k-NN: Resultado



k-NN: Resultado

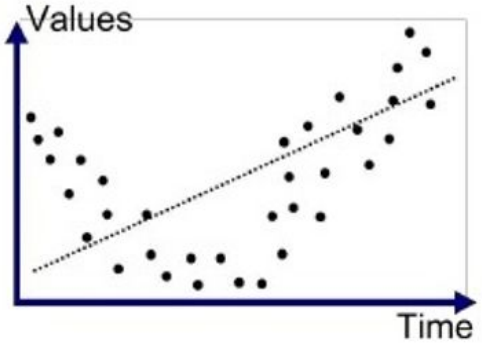


Midiendo el Performance del Modelo

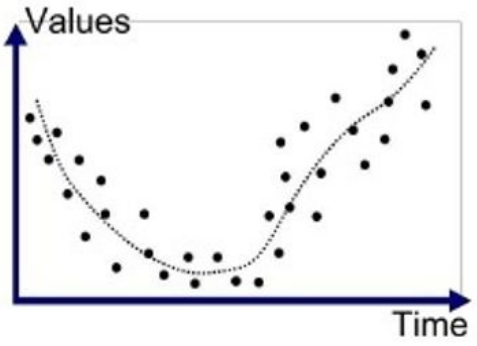
- En clasificación, el accuracy es una métrica comúnmente usada.
- Accuracy = Fracción de predicciones correctas.
- Cuál data debe ser usada para calcular el accuracy?
- Cómo rendirá nuestro modelo en una nueva data?

$$accuracy = \frac{correct}{correct + incorrect}$$

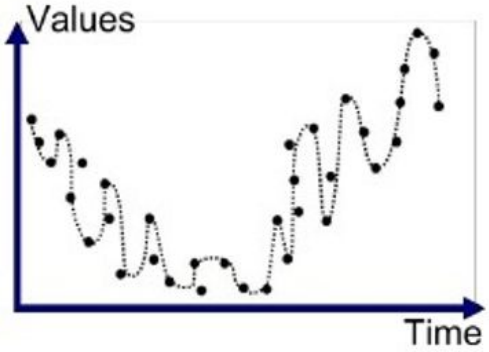
Overfitting & Underfitting



Underfitted



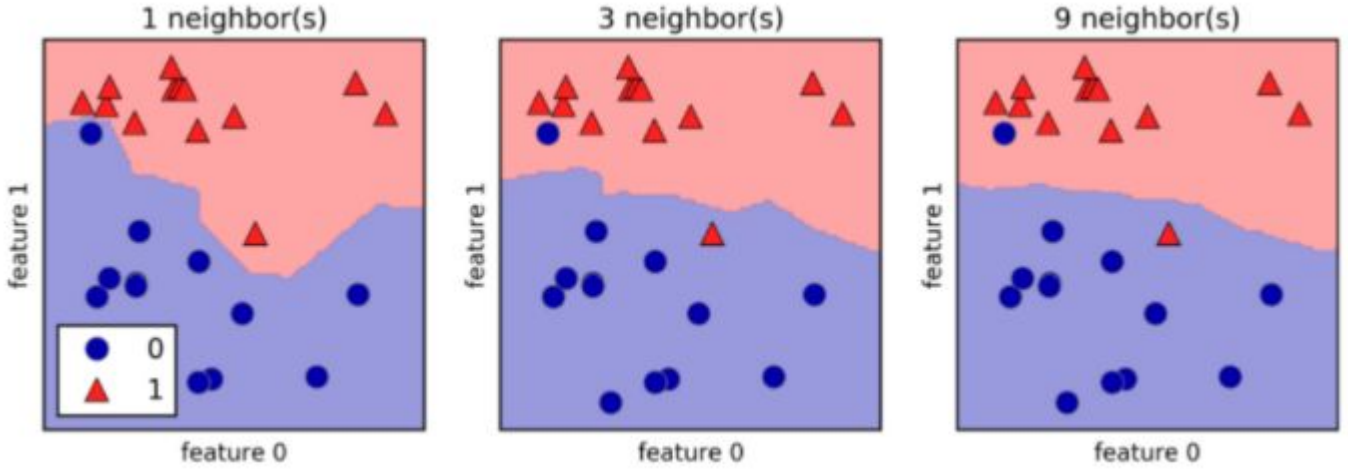
Good Fit/Robust



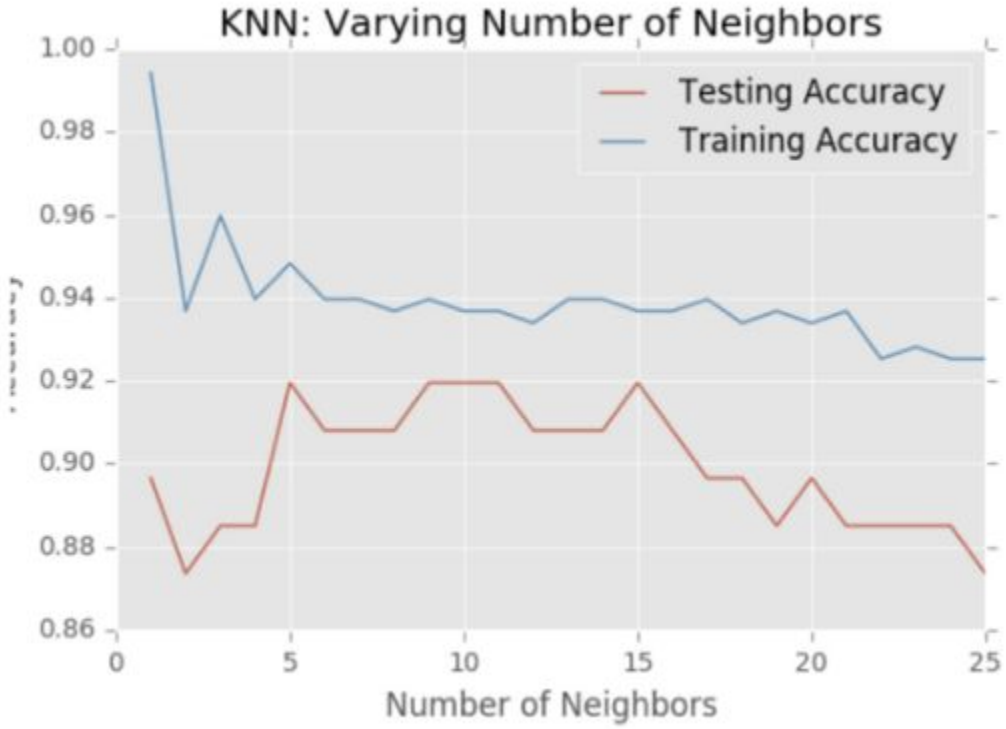
Overfitted

Complejidad del Modelo

- k grande = límites más lineales = modelo menos complejo
- k pequeño = modelo más complejo = puede llevar a overfitting



Complejidad del Modelo y over/underfitting



Complejidad del Modelo y over/underfitting