

# K-NN como estrategia general en Machine Learning

1. ¿Qué es la técnica de K-NN?
2. ¿Cómo funciona?
3. Aplicaciones.

¿Qué es la  
técnica de K-NN?

K-NN significa **K-Nearest Neighbors**, lo que en español es **K vecinos más cercanos**.

K-NN es un método ampliamente utilizado en la literatura para diversas tareas como por ejemplo:

- » Imputación de datos.
- » Clasificación (lo veremos más adelante)
- » Regresión (lo veremos más adelante)
- » Clustering (también lo veremos más adelante)

Veremos como funciona de manera general y luego veremos su aplicación en imputación.

¿Cómo funciona?

# Funcionamiento de K-NN

K-NN tiene 2 parámetros importantes:

- » La medida de distancia o similaridad.
- » El parámetro “K” o “e” según como se decida crear los vecindarios.

Primera veremos la distancia o similaridad.

# Funcionamiento de K-NN

El concepto de similaridad o distancia son parecidos, ya que tratan de cuantificar lo mismo, pero de puntos de vistas distinto.

La distancia entre más crece, *más alejados o distintos son los objetos a medir.*

La similaridad entre más crece, *más parecidos o cercanos son los objetos a medir.*

Existen diversas formas de calcular estas métricas. Aca veremos algunas.

# Funcionamiento de K-NN

## Distancias:

$$D(x, y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

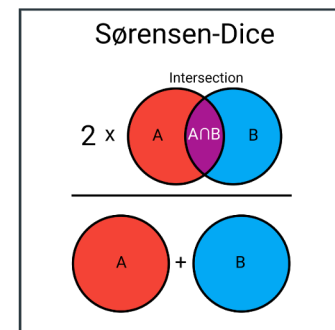
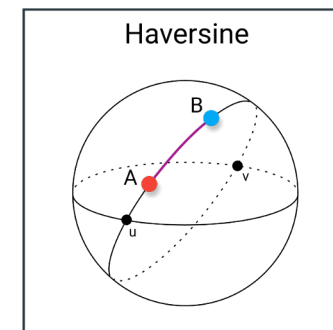
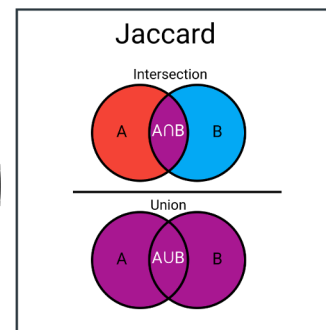
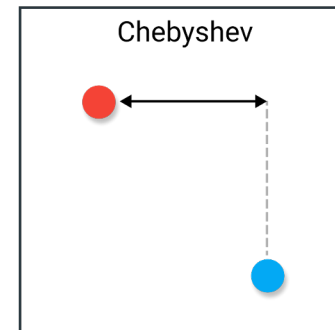
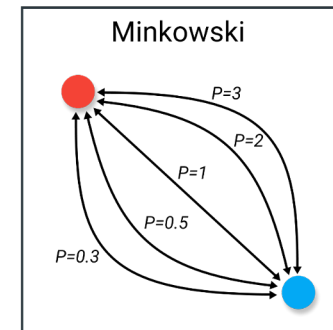
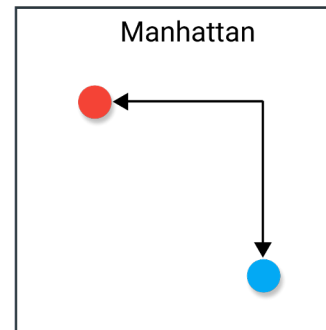
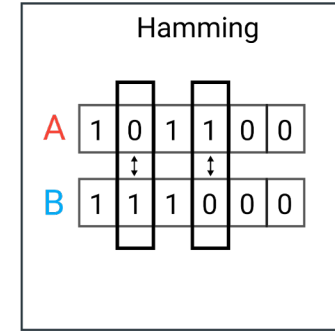
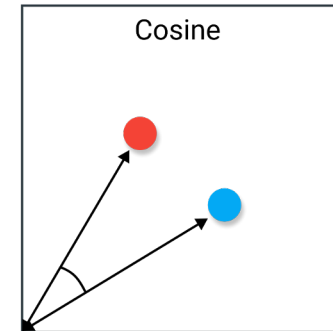
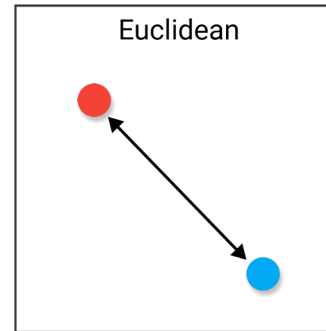
Minkowski

$$D(x, y) = 1 - \frac{|x \cap y|}{|y \cup x|}$$

Jaccard

$$d = 2 \arcsin \left( \sqrt{\sin^2 \left( \frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left( \frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

Haversine





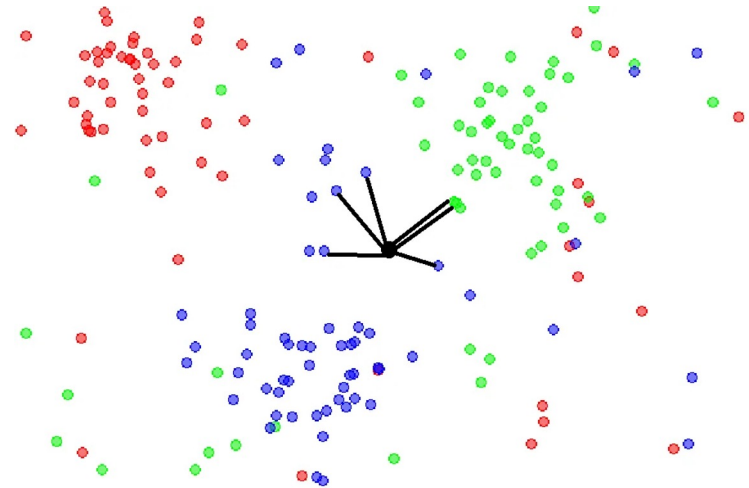
# Funcionamiento de K-NN

Estrategia para crear vecindarios.

**La más tradicional es usar el parámetro K:** Esto se basa en elegir los K vecinos más cercanos según una métrica de distancia o similaridad.

En este ejemplo, el punto negro central busca los  $K=6$  vecinos más cercanos en un espacio de 2 dimensiones.

Esos puntos serían su vecindario.



# Funcionamiento de K-NN

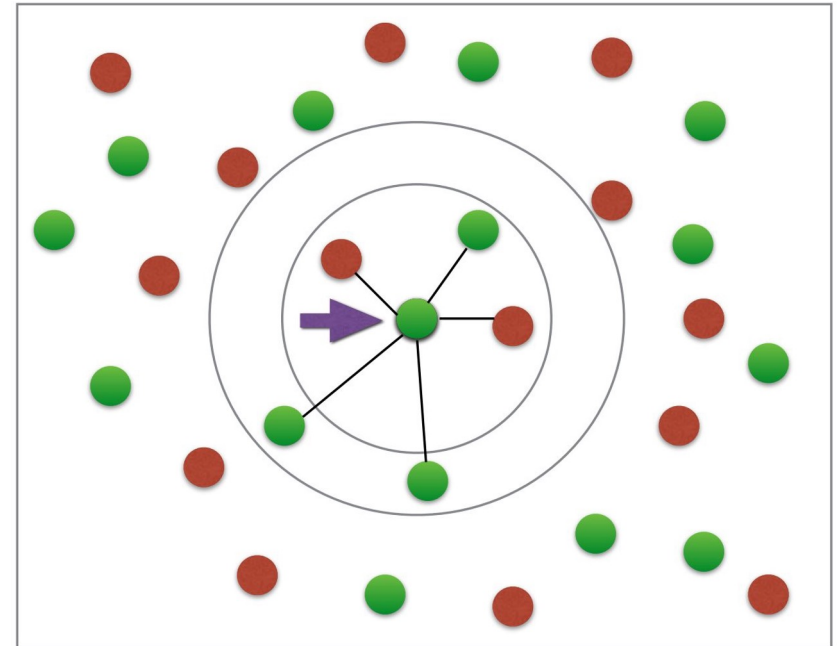
Estrategia para crear vecindarios.

La otra estrategia es usar el parámetro “e”: este parámetro define un radio “e” y todo dato que esté dentro de ese radio “e” es un dato perteneciente al vecindario.

En este ejemplo, hay un punto central indicado con la flecha. Este punto define su vecindario según el valor del parámetro “e”.

En el ejemplo hay 2 circunferencias que son el resultado de 2 valores de “e” distintos.

Según el valor de “e”, es el tamaño del vecindario (3 y 5).



# Funcionamiento de K-NN

Para usar el algoritmo K-NN se deben definir al menos estos 2 parámetros.

K-NN es extremadamente versátil como técnica, pero eso produce pagar un precio:

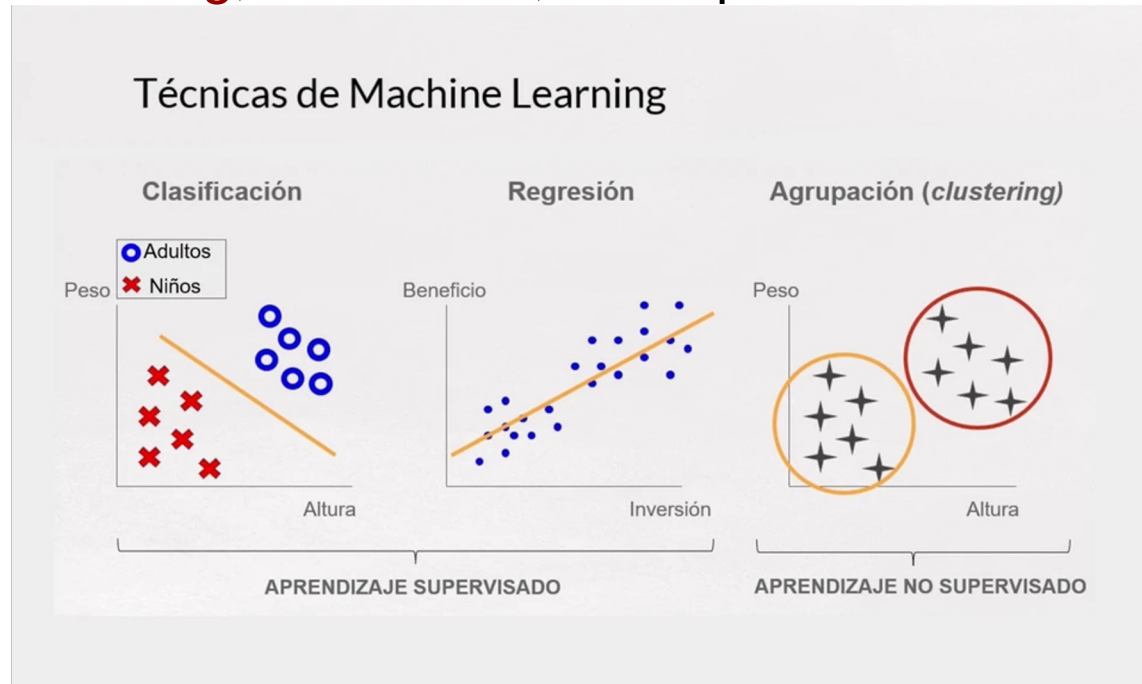
- » El algoritmo es muy sensible a los valores de sus parámetros (especialmente a K o “e”).
- » No es una técnica que pueda resolver problemas muy complicados. Por eso, muchas veces sirve de apoyo para una solución más integral a un problema de ML complicado.

# Aplicaciones de KNN

# Aplicaciones de K-NN

Como se mencionó antes, existe varias aplicaciones de algoritmo K-NN.

K-NN se puede usar para resolver problemas de **clasificación**, **regresión** y **clustering**, entre otros, todos problemas de Machine learning.



Fuente: <https://openwebinars.net/blog/modelos-de-machine-learning/>

Aquí mostraremos K-NN como técnica de Imputación.

# Aplicaciones de K-NN

## **K-NN para imputación.**

La idea de usar K-NN para imputación es usar los vecindarios para obtener el valor que reemplazaremos por el MV.

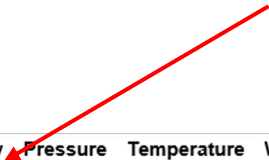
Definidos los parámetros del algoritmo (distancia y valor de K o “e”) buscaremos cuáles son los datos que conformarán el vecindario.

La distancia se debe calcular con las variables observables y, además, se deben elegir datos que tengan valores que sirvan para realizar la imputación, es decir, valores en la variable con MV.
















Veamos un ejemplo:

# Aplicaciones de K-NN

1.- Supongamos que queremos imputar este MV



	Site Num	Latitude	Longitude	Sample Measurement	Date	Humidity	Pressure	Temperature	Weather_descr	Wind_dir	Wind_speed
0	1	41.670992	-87.732457	17.7	2013-01-01 00:00:00	NaN	1024.0	-0.19	16.0	200.0	4.0
1	1	41.670992	-87.732457	14.6	2013-01-01 01:00:00	64.0	1022.0	0.28	0.0	180.0	3.0
2	1	41.670992	-87.732457	13.5	2013-01-01 02:00:00	69.0	1022.0	0.33	16.0	190.0	6.0
3	1	41.670992	-87.732457	11.9	2013-01-01 03:00:00	NaN	1021.0	0.12	16.0	190.0	7.0
4	1	41.670992	-87.732457	10.3	2013-01-01 04:00:00	68.0	1021.0	0.04	0.0	210.0	7.0
5	1	41.670992	-87.732457	8.4	2013-01-01 05:00:00	68.0	1020.0	-0.04	16.0	200.0	7.0
6	1	41.670992	-87.732457	4.9	2013-01-01 06:00:00	NaN	NaN	0.00	22.0	63.0	3.0
7	1	41.670992	-87.732457	3.7	2013-01-01 07:00:00	68.0	1018.0	-0.15	16.0	200.0	4.0
8	1	41.670992	-87.732457	5.7	2013-01-01 08:00:00	NaN	1018.0	0.60	0.0	210.0	6.0
9	1	41.670992	-87.732457	7.2	2013-01-01 09:00:00	NaN	1018.0	0.52	0.0	210.0	5.0
10	1	41.670992	-87.732457	6.9	2013-01-01 10:00:00	69.0	1017.0	0.83	16.0	220.0	6.0
11	1	41.670992	-87.732457	5.7	2013-01-01 11:00:00	64.0	1017.0	1.01	16.0	230.0	6.0
12	1	41.670992	-87.732457	5.7	2013-01-01 12:00:00	55.0	1017.0	1.29	16.0	240.0	7.0
13	1	41.670992	-87.732457	6.8	2013-01-01 13:00:00	59.0	1017.0	1.20	0.0	240.0	5.0
14	1	41.670992	-87.732457	6.9	2013-01-01 14:00:00	59.0	1017.0	1.34	16.0	230.0	5.0



2.- Entonces, podemos usar las variables numéricas y observables para calcular la distancia.

# Aplicaciones de K-NN

3.- La distancia se calcula entre el vector que tiene el MV que queremos imputar.


	Site Num	Latitude	Longitude	Sample Measurement	Date	Humidity	Pressure	Temperature	Weather_descr	Wind_dir	Wind_speed
0	1	41.670992	-87.732457	17.7	2013-01-01 00:00:00	NaN	1024.0	-0.19	16.0	200.0	4.0
1	1	41.670992	-87.732457	14.6	2013-01-01 01:00:00	64.0	1022.0	0.28	0.0	180.0	3.0
2	1	41.670992	-87.732457	13.5	2013-01-01 02:00:00	69.0	1022.0	0.33	16.0	190.0	6.0
3	1	41.670992	-87.732457	11.9	2013-01-01 03:00:00	NaN	1021.0	0.12	16.0	190.0	7.0
4	1	41.670992	-87.732457	10.3	2013-01-01 04:00:00	68.0	1021.0	0.04	0.0	210.0	7.0
5	1	41.670992	-87.732457	8.4	2013-01-01 05:00:00	68.0	1020.0	-0.04	16.0	200.0	7.0
6	1	41.670992	-87.732457	4.0	2013-01-01 06:00:00	NaN	NaN	0.00	22.0	03.0	3.0
7	1	41.670992	-87.732457	3.7	2013-01-01 07:00:00	68.0	1018.0	-0.15	16.0	200.0	4.0
8	1	41.670992	-87.732457	5.7	2013-01-01 08:00:00	NaN	1018.0	0.60	0.0	210.0	6.0
9	1	41.670992	-87.732457	7.2	2013-01-01 09:00:00	NaN	1018.0	0.52	0.0	210.0	5.0
10	1	41.670992	-87.732457	6.9	2013-01-01 10:00:00	69.0	1017.0	0.83	16.0	220.0	6.0
11	1	41.670992	-87.732457	5.7	2013-01-01 11:00:00	64.0	1017.0	1.01	16.0	230.0	6.0
12	1	41.670992	-87.732457	5.7	2013-01-01 12:00:00	55.0	1017.0	1.29	16.0	240.0	7.0
13	1	41.670992	-87.732457	6.8	2013-01-01 13:00:00	59.0	1017.0	1.20	0.0	240.0	5.0
14	1	41.670992	-87.732457	6.9	2013-01-01 14:00:00	59.0	1017.0	1.34	16.0	230.0	5.0

4.- Y los vectores que también tienen valores en esas variables y que tengan datos observados en la variable **Humidity**  
Eso hace que los datos: 3, 6, 8 y 9 no sean considerados por tener MV.



# Aplicaciones de K-NN

5. Supongamos que  $K=3$ . Entonces, los vecinos más cercanos serían: 1, 5 y 7.



	Site Num	Latitude	Longitude	Sample Measurement	Date	Humidity	Pressure	Temperature	Weather_descr	Wind_dir	Wind_speed
0	1	41.670992	-87.732457	17.7	2013-01-01 00:00:00	NaN	1024.0	-0.19	16.0	200.0	4.0
1	1	41.670992	-87.732457	14.6	2013-01-01 01:00:00	64.0	1022.0	0.28	0.0	180.0	3.0
2	1	41.670992	-87.732457	13.5	2013-01-01 02:00:00	69.0	1022.0	0.33	16.0	190.0	6.0
3	1	41.670992	-87.732457	11.9	2013-01-01 03:00:00	NaN	1021.0	0.12	16.0	190.0	7.0
4	1	41.670992	-87.732457	10.3	2013-01-01 04:00:00	68.0	1021.0	0.04	0.0	210.0	7.0
5	1	41.670992	-87.732457	8.4	2013-01-01 05:00:00	68.0	1020.0	-0.04	16.0	200.0	7.0
6	1	41.670992	-87.732457	4.9	2013-01-01 06:00:00	NaN	NaN	0.00	22.0	63.0	3.0
7	1	41.670992	-87.732457	3.7	2013-01-01 07:00:00	68.0	1018.0	-0.15	16.0	200.0	4.0
8	1	41.670992	-87.732457	5.7	2013-01-01 08:00:00	NaN	1018.0	0.60	0.0	210.0	6.0
9	1	41.670992	-87.732457	7.2	2013-01-01 09:00:00	NaN	1018.0	0.52	0.0	210.0	5.0
10	1	41.670992	-87.732457	6.9	2013-01-01 10:00:00	69.0	1017.0	0.83	16.0	220.0	6.0
11	1	41.670992	-87.732457	5.7	2013-01-01 11:00:00	64.0	1017.0	1.01	16.0	230.0	6.0
12	1	41.670992	-87.732457	5.7	2013-01-01 12:00:00	55.0	1017.0	1.29	16.0	240.0	7.0
13	1	41.670992	-87.732457	6.8	2013-01-01 13:00:00	59.0	1017.0	1.20	0.0	240.0	5.0
14	1	41.670992	-87.732457	6.9	2013-01-01 14:00:00	59.0	1017.0	1.34	16.0	230.0	5.0

6.- Finalmente, con esos 3 datos estimamos el MV con la técnica que queramos. En este caso, podríamos usar la media:  
 $(64.0 + 68.0 + 68.0)/3$

# Aplicaciones de K-NN

- Aquí vimos un ejemplo de K-NN para imputación con un técnica Univariada, es decir, solo considera el promedio de la variable en cuestión.
- Pero, también podemos usar técnicas más sofisticadas que involucren un proceso de estimación del MV que consideren varias variables al mismo tiempo: Técnica Multivariada.
- Más adelante veremos técnicas de este tipo que podrían aplicarse a esto.
- Spoiler!!!!
  - Regresión