

CLASIFICACIÓN Y REGRESIÓN USANDO K-NEAREST NEIGHBORS

José Miguel Llanos Mosquera

**Asignatura: Minería de datos
Universidad del Valle
2020**

CONTENIDO

- Definición de K-NN
- Objetivo de K-NN
- Cómo hacer clasificación y regresión con K-NN
- Distancia Euclidiana
- Distancia Manhattan
- Distancia de Hamming (binarios, vectores y matriz de conteo)
- Distancia en variables categóricas
- Distancia en variables ordinales
- Conclusiones
- Ejercicio K-NN en Python con sklearn

CONTEXTO

Clasificación

- La clasificación consiste en identificar características de un objeto o registro, con el fin de asignar una clase o categoría definida.

Regresión

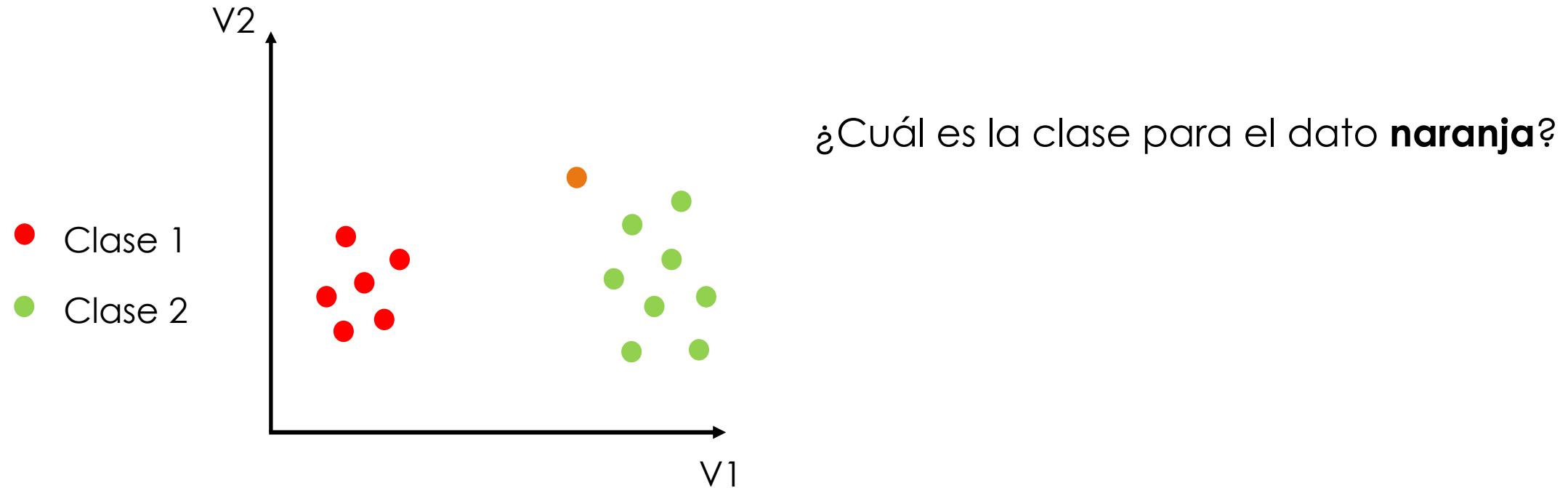
- La regresión permite predecir un valor ausente de una variable, basándose en la relación con otras variables que pertenecen al mismo grupo de datos.

DEFINICIÓN DE K-NN

- Es un algoritmo supervisado (el conjunto de datos de entrenamiento está etiquetado con la clase).
- Puede usarse para clasificar valores **discretos** o para predecir valores **continuos** (regresión).
- Clasifica los valores buscando los datos “más similares” por cercanía, aprendidos en la etapa de entrenamiento.
- La “K” significa la cantidad de “puntos vecinos” que se tienen en cuenta para clasificar.
- Se utiliza en la resolución de problemas relacionados a: **Sistemas de recomendación, búsqueda semántica y detección de anomalías.**

OBJETIVO DE K-NN

- Clasificar un dato nuevo, a partir de casos más cercanos o similares a la clase.



¿CÓMO HACER CLASIFICACIÓN CON K-NN?

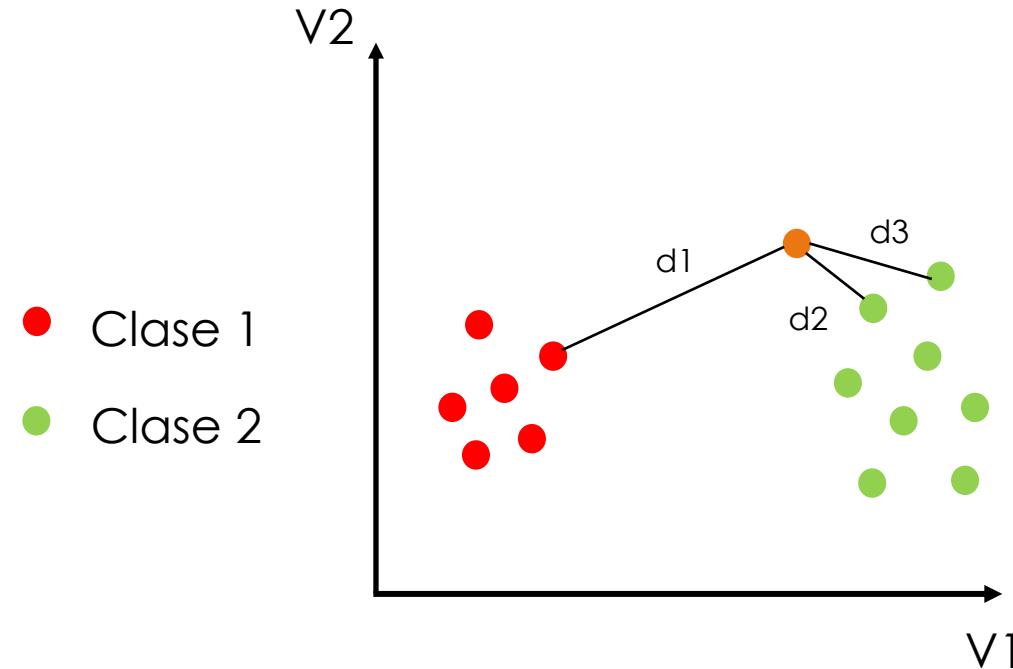
1. Se debe calcular la distancia entre el item y el resto de items del dataset de entrenamiento.
2. Se debe seleccionar los **K** elementos más cercanos (con menor distancia).
3. Finalmente se debe realizar una “votación de mayoría” entre los **K** puntos, donde los elementos de una clase que dominen deciden la clasificación final.

¿CÓMO HACER REGRESIÓN CON K-NN?

1. Primero se identifican los **K** vecinos para cada punto.
2. Luego, en lugar de considerar su clase y establecer un sistema de votación, se considera el valor que toma las etiquetas para cada uno de los **K** vecinos.
3. Finalmente se devuelve como predicción el valor de la media de dichos valores (**K** vecinos).

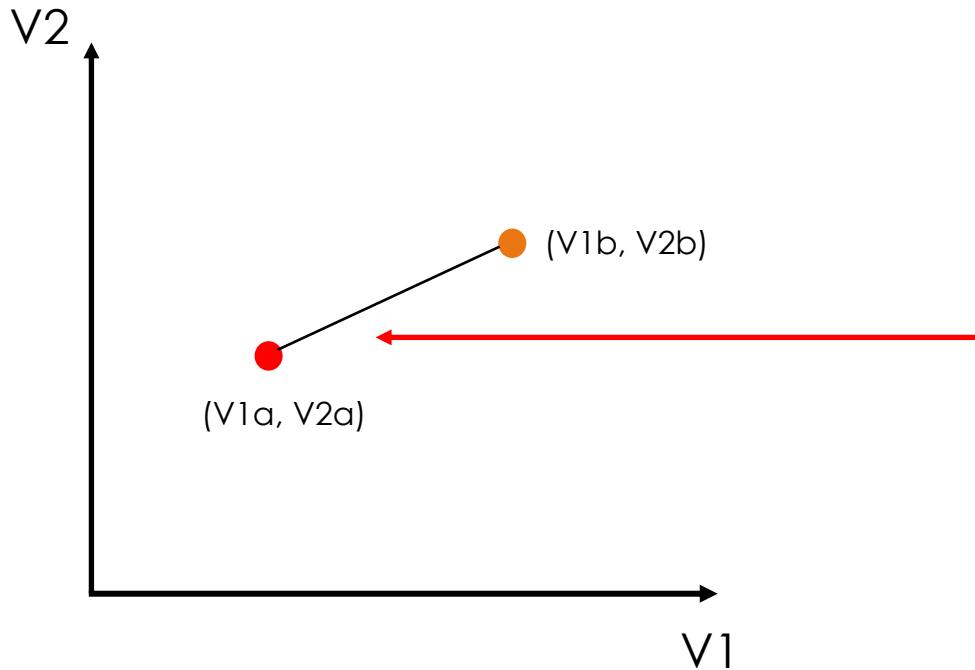
DISTANCIA

- Para encontrar el dato más parecido, se necesita medir la distancia entre los datos.
- Se asume que los datos más parecidos son los que se encuentran más cerca en el espacio.



DISTANCIA EUCLIDIANA

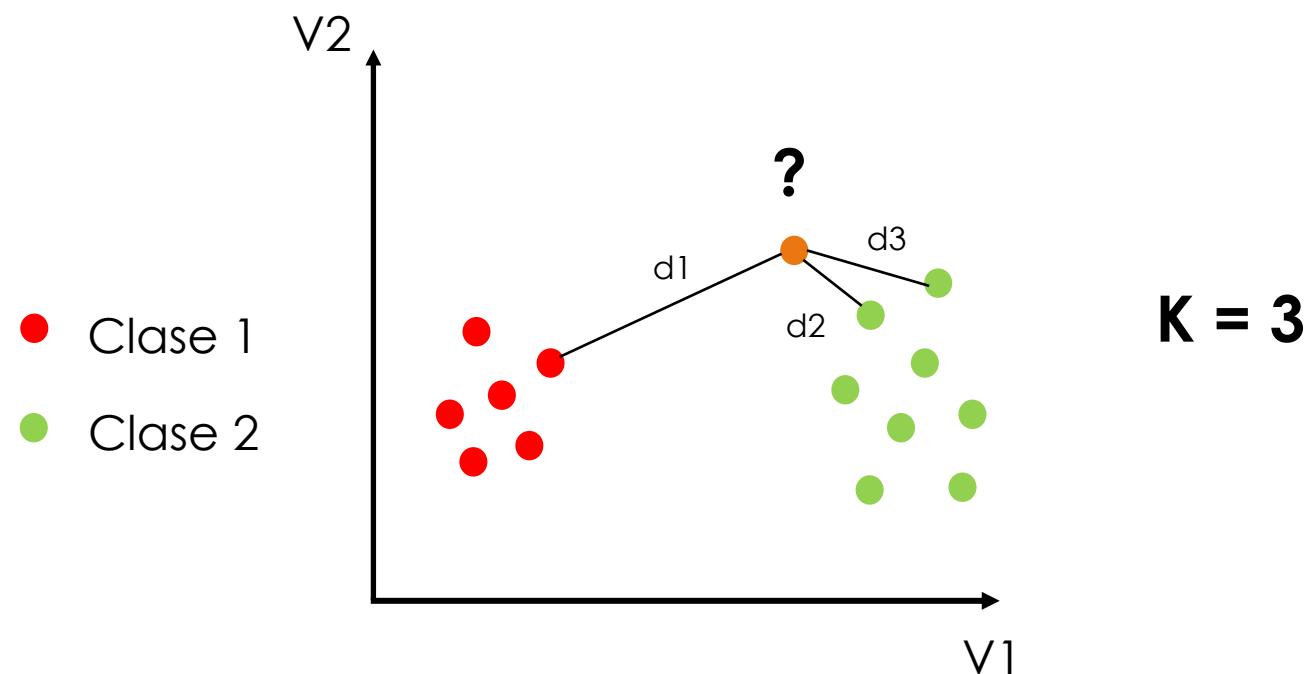
- Para medir esta distancia se utiliza la métrica **distancia Euclídea**.



$$d = \sqrt{(V1a - V1b)^2 + (V2a - V2b)^2}$$

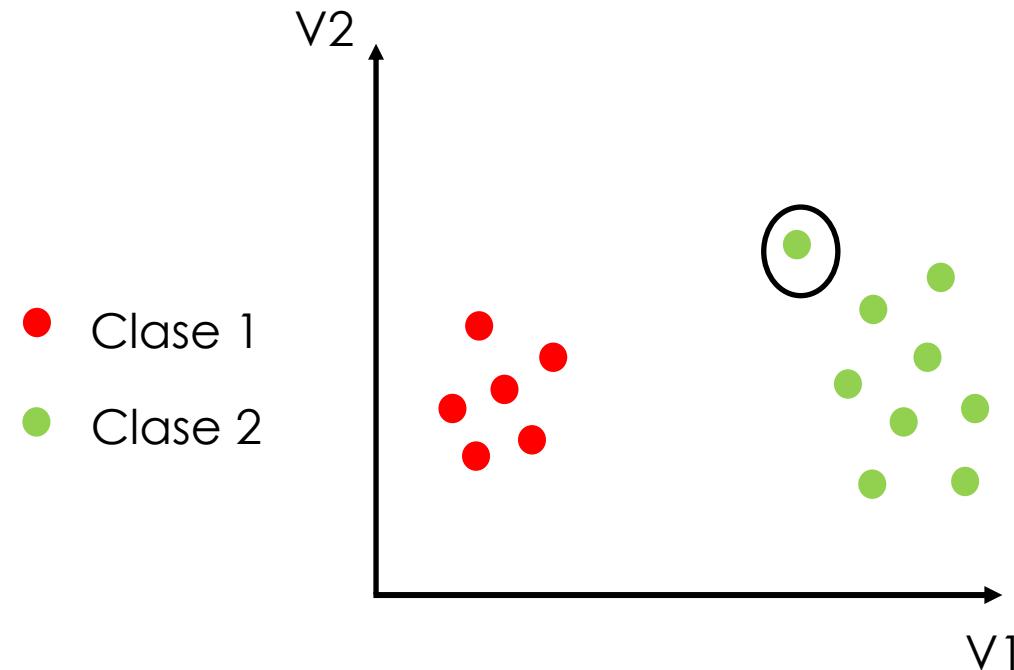
DISTANCIA EUCLIDIANA

- La distancia se debe calcular para cada una de las variables que se están utilizando, la métrica también funciona para un número mayor de variables.
- La distancia más pequeña corresponde a **d2**, entonces ese dato es el vecino más cercano.



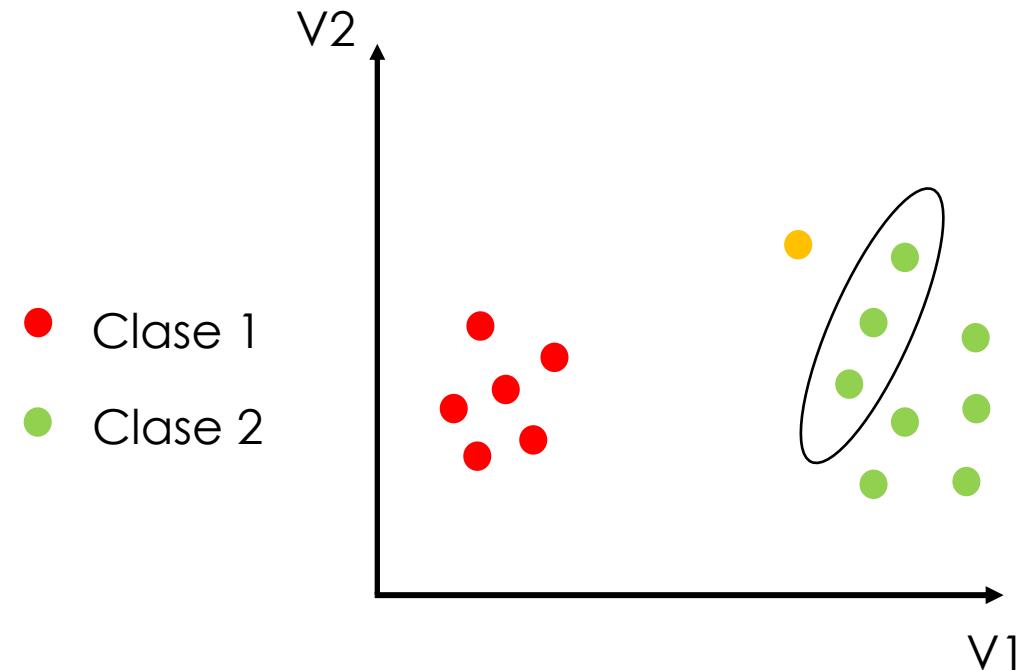
DISTANCIA EUCLIDIANA

- La clasificación que hace el modelo corresponde a la misma clase del vecino más cercano, es decir (**clase 1**: color verde).
- En este caso se considera el voto de la mayoría de las **K** instancias que más se parece.



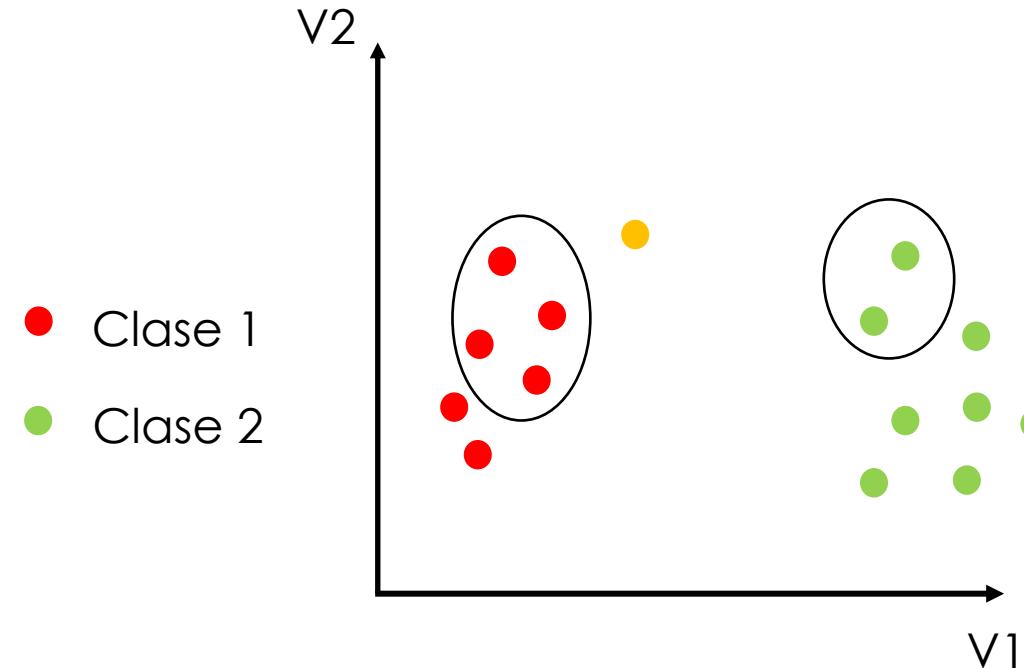
DISTANCIA EUCLIDIANA (EJEMPLO 2)

- $K = 3$, por unanimidad se encuentra que los tres vecinos más cercanos se encuentran en la misma clase (**clase 2**: color verde)
- El modelo de clasificación define que el dato naranja corresponde a la **clase 2**



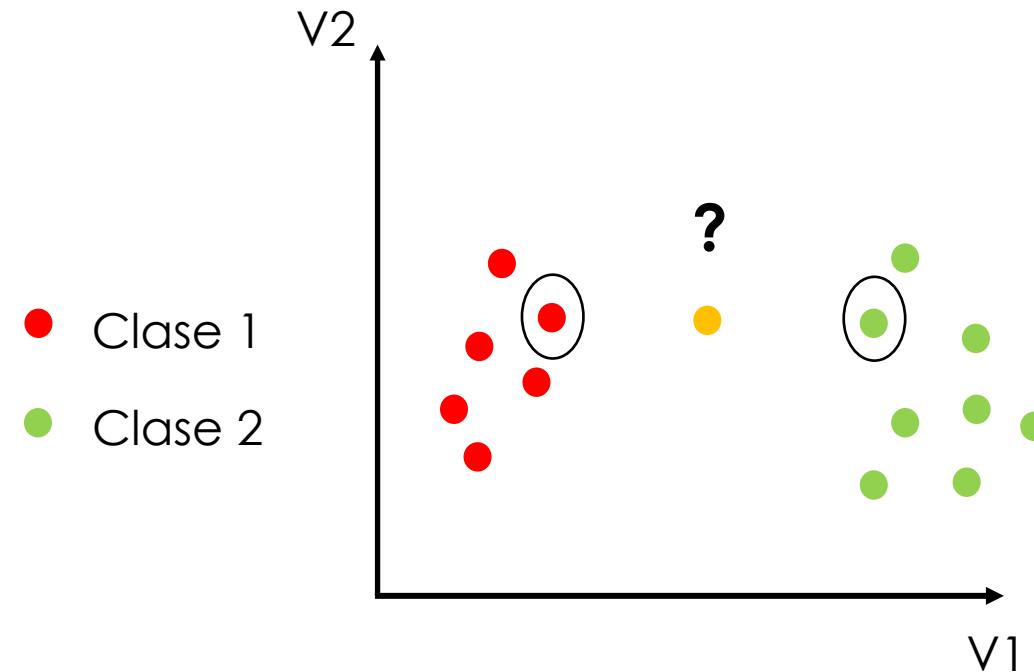
DISTANCIA EUCLIDIANA (EJEMPLO 3)

- $K = 6$
- **Clasificación:** El dato corresponde a la clase 1



DISTANCIA EUCLIDIANA (EJEMPLO 4)

- $K = 2$
- **Clasificación:** Muy difícil, porque no es claro a que clase pertenece el dato.



VARIANTES DEL K-NN

- Los valores que pueden tomar los datos, pueden ser diferentes (categóricos, numéricos, etc).
- Dependiendo el tipo de variable para los datos, se debe seleccionar la métrica.
- Por este motivo existen diferentes medidas para la distancia.

VARIANTES DEL K-NN (EJEMPLO 1)

Ejemplo de clientes: ¿Cuál es la distancia de Maria a Juan?

Nombre	Profesión	Rango Sueldo	Género	Gastos Mensuales	Ubicación (latitud,longitud)
Maria	Enfermera	500 – 1000	Femenino	30	(-27.98, 33.45)
Juan	Músico	100 – 500	Masculino	10	(-35.65, 22.16)

$$d(\text{Maria}, \text{Juan}) = \text{dif}(\text{Enfermera}; \text{Músico}) + \text{dif}(500/1000; 100/500) + \text{dif}(\text{Femenino}; \text{Masculino}) + \text{dif}(30; 10) + \text{dif}((-27.98, 33.45); (-35.65, 22.16))$$

VARIANTES DEL K-NN (EJEMPLO 1)

- Los cálculos de la distancia comparan variables correspondientes (del mismo tipo) no variables diferentes.
- La función de distancia debe ir acumulando las diferencias entre cada una de las variables.

VARIANTES DEL K-NN

¿Cómo calcular la distancia usando la variable ubicación?

- La ubicación corresponde a latitud y longitud.
- Para calcular la distancia, se puede utilizar distancia **Euclíadiana** o de **Manhattan**.

VARIANTES DEL K-NN

(-23.99, 24.25)

(32.51, 91.82)

Distancia Euclíadiana:

$$d(\text{UbicaciónCliente1}, \text{UbicaciónCliente2}) \\ d((-23.99, 24.25), (32.51, 91.82))$$

$$\sqrt{(32.51 - (-23.99))^2 + (91.82 - 24.25)^2}$$

$$d(\text{UbicaciónCliente1}, \text{UbicaciónCliente2}) \approx \mathbf{88.079}$$

VARIANTES DEL K-NN

Distancia Euclíadiana:

- Se puede extender a ***n*** dimensiones

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

DISTANCIA MANHATTAN

(-23.99, 24.25)

(32.51, 91.82)

Distancia Manhattan:

$$dM(P, Q) = |x_2 - x_1| + |y_2 - y_1|$$

$d(\text{UbicaciónCliente1}, \text{UbicaciónCliente2})$

$d((-23.99, 24.25), (32.51, 91.82))$

$|32.51 - (-23.99)| + |91.82 - 24.25|$

$d(\text{UbicaciónCliente1}, \text{UbicaciónCliente2}) \approx \mathbf{124.07}$

$$d(P, Q) = \sum_{i=1}^D |P_i - Q_i|$$

DISTANCIA HAMMING

¿Cómo calcular la distancia usando la variable Género?

Es una variable Binaria, se utiliza la distancia de **Hamming**

Masculino → 0

Femenino → 1

DISTANCIA HAMMING

Distancia de Hamming

Distancia de Hamming

$$\text{dist}(0; 0) = 0$$

$$\text{dist}(0; 1) = 1$$

$$\text{dist}(1; 0) = 1$$

$$\text{dist}(1; 1) = 0$$

DISTANCIA HAMMING (EJEMPLO)

Distancia Hamming

$$d(\text{Hombre}, \text{Mujer}) \longrightarrow 1$$

$$d(\text{Mujer}, \text{Mujer}) \longrightarrow 0$$

DISTANCIA HAMMING

- La distancia Hamming también se puede aplicar a **vectores**

10**11101**

y

10**00111**

= 3

DISTANCIA HAMMING (EJEMPLO)

Existe una **mujer** la cual **no** es trabajadora independiente y **si** tiene hijos.

101

Existe un **hombre** el cual **si** es trabajador independiente y **no** tiene hijos.

010

Usando la distancia de **Hamming**, temenos:

$d(101; 010)$

= 3

DISTANCIA HAMMING

- La distancia Hamming también se puede **normalizar**

10 $\textcolor{red}{1}1101$

y

10 $\textcolor{red}{0}0111$

= 3/7

- Se utiliza mucho en la práctica para mantener los valores pequeños en las distancias.
- Porque, si un grupo de variables toma valores de distancias muy grandes, gobernarían la distancia total, al momento de unir las diferentes distancias.

DISTANCIA HAMMING

- Para visualizar la distancia entre dos vectores también se puede utilizar la **matriz de conteo**.
- muestra el número de veces que ocurre cada combinación.

C1

1 0 1 1 1 0 1

C2

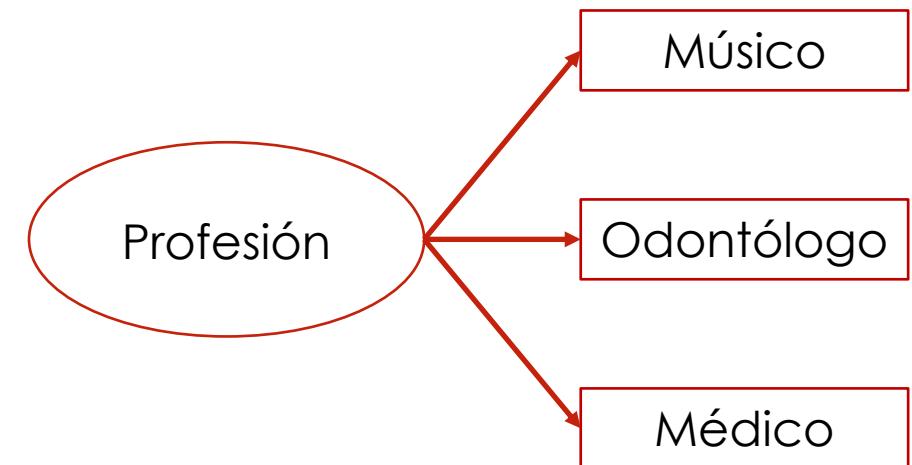
1 0 0 0 1 1 1

C1/C2	0	1
0	1	1
1	2	3

DISTANCIA VARIABLES CATEGÓRICAS

¿Cómo encontrar la distancia entre variables Categóricas?

- Tienen una cantidad de posibles categorías
- No tienen un orden definido entre sus valores



DISTANCIA VARIABLES CATEGÓRICAS

¿Cómo encontrar la distancia entre variables categóricas?

Nombre	Profesión	Ciudad	Deporte
Pedro	Médico	Bogotá	Tenis
Sofia	Odontólogo	Cali	Tenis

≠

≠

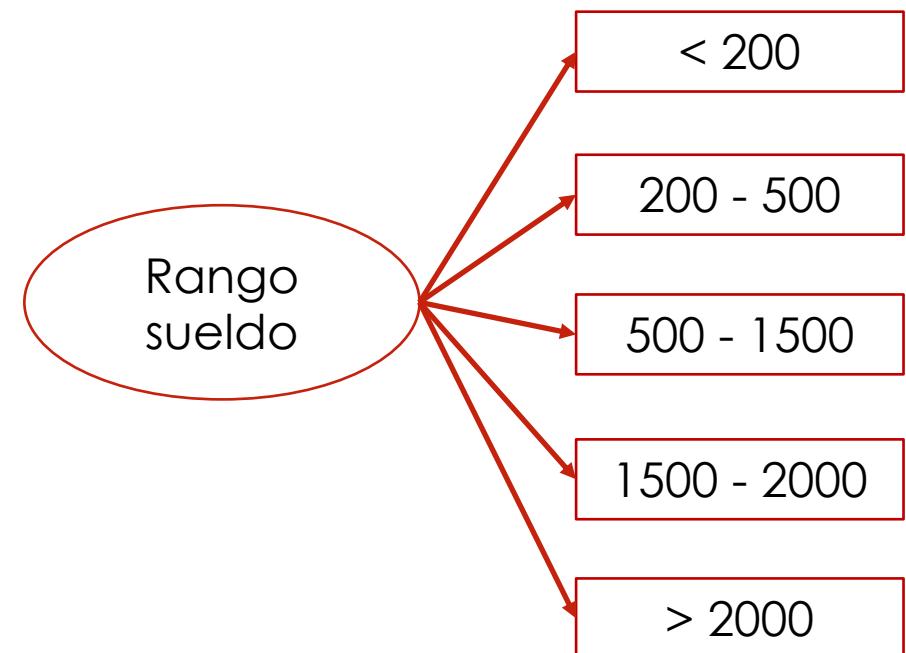
=

$$d(Pedro, Sofia) = 2/3$$

DISTANCIA EN VARIABLES ORDINALES

¿Cómo encontrar la distancia entre variables Ordinales?

- Tienen una cantidad fija de posibles categorías
- Tienen un orden definido entre sus valores



DISTANCIA EN VARIABLES ORDINALES

¿Cómo encontrar la distancia entre variables Ordinales?

- Los valores se mapean a un rango entre 0 y 1
- Luego se aplica la distancia **Euclidian**a para las variables

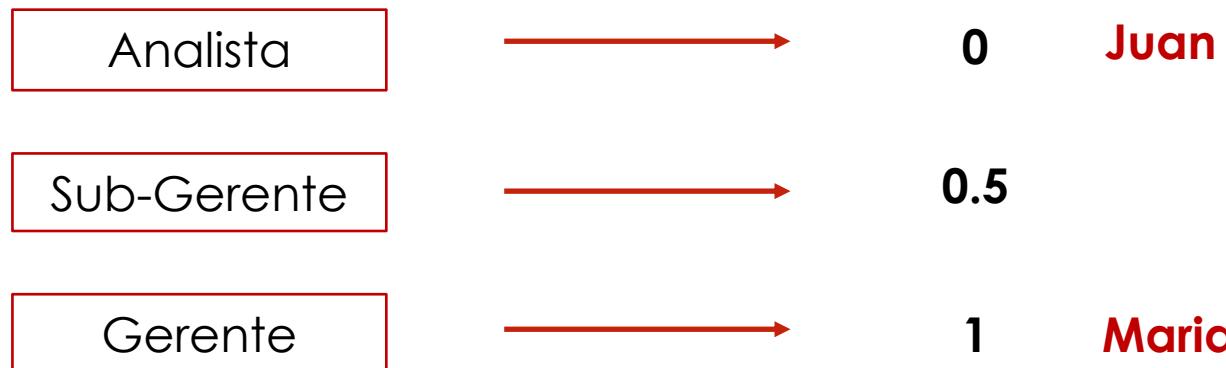
DISTANCIA EN VARIABLES ORDINALES (EJEMPLO)

Ejemplo: ¿Cuál es la distancia entre Juan y María?

Nombre	Cargo	Rango sueldo
Juan	Analista	400
Maria	Gerente	1700

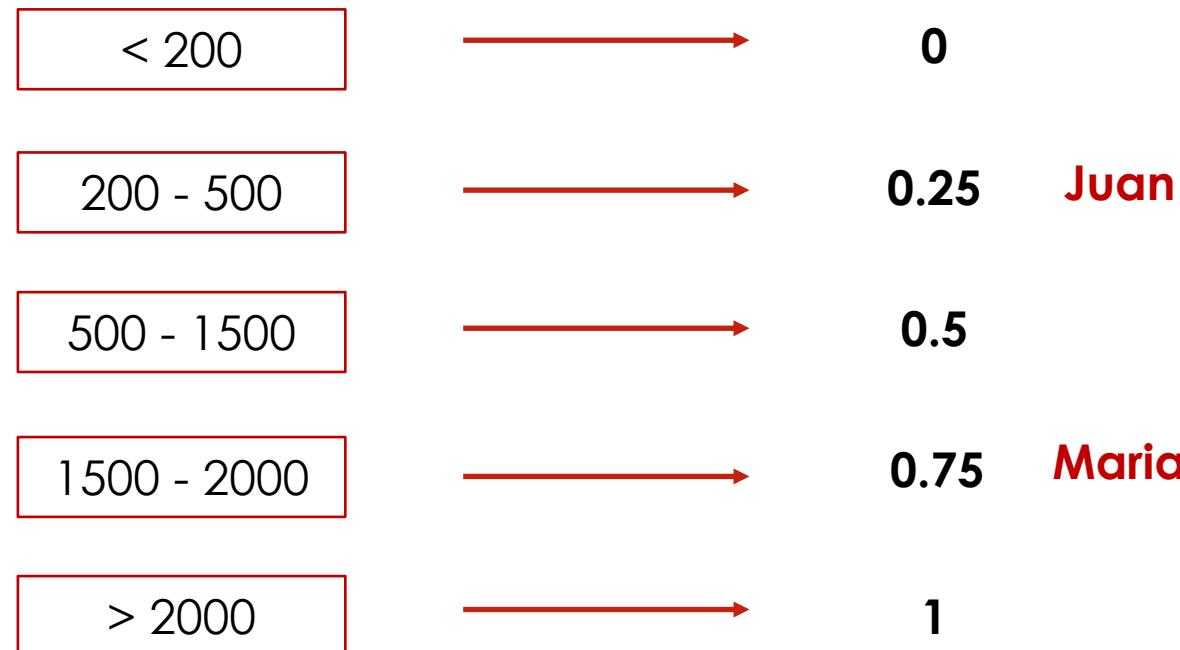
DISTANCIA EN VARIABLES ORDINALES (EJEMPLO)

- Los valores del **Cargo** se mapean (valores entre 0 y 1)



DISTANCIA EN VARIABLES ORDINALES (EJEMPLO)

- Los valores del **Rango Sueldo** se mapean (valores entre 0 y 1)



DISTANCIA EN VARIABLES ORDINALES (EJEMPLO)

Ahora se aplica la distancia **Euclidian**a

Nombre	Cargo	Rango sueldo
Juan	Analista	400
Maria	Gerente	1700



Nombre	Cargo	Rango sueldo
Juan	0	0.25
Maria	1	0.75

$$d(Juan, Maria) = \sqrt{(0 - 1)^2 + (0.25 - 0.75)^2}$$

$$d(Juan, Maria) \approx 1.11$$

CONCLUSIONES

- K-NN es un algoritmo de clasificación.
- Usa los datos más similares para clasificar.
- Requiere de una medida de distancia y un valor **K**.
- Existen diferentes métricas para los tipos de datos.
- Sólo usa tiempo de computación a la hora de realizar la clasificación, pero utiliza todo el dataset para entrenar, entonces se recomienda para pocos datos por el uso de memoria.

REFERENCIAS

1. Curso introducción a la Minería de Datos, Pontificia Universidad Católica de chile (Coursera).
2. Aprende Machine Learning, Juan Ignacio Bagnato,
<https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/>

EJERCICIO PYTHON CON SKLEARN

- Se tiene el archivo **reviews.csv** con diferentes opiniones de usuarios sobre una App (257 registros).
- De los diferentes campos se utilizaran las columnas **wordcount** y **sentimentValue** como características para alimentar el algoritmo K-NN.
- La columna **wordcount** indica la cantidad de palabras utilizadas para la opinión.
- La columna **sentimentValue** indica si el comentario fue valorado como positivo o negativo (-4 hasta 4).
- La etiqueta será **Star Rating** indica las estrellas que dieron los usuarios a la App, son valores discretos del 1 al 5.

<https://github.com/jose-llanos/K-NN>

EJERCICIO PYTHON CON SKLEARN

The diagram consists of a large downward-pointing arrow centered above a table. Three smaller red arrows point downwards from the top of the table towards the column headers: 'Review Title', 'Star Rating', and 'sentimentValue'.

	Review Title	Review Text	wordcount	titleSentiment	textSentiment	Star Rating	sentimentValue
0	Sin conexión	Hola desde hace algo más de un mes me pone sin...	23	negative	negative	1	-0.486389
1	faltan cosas	Han mejorado la apariencia pero no	20	negative	negative	1	-0.586187
2	Es muy buena lo recomiendo	Andres e puta amoooo	4	NaN	negative	1	-0.602240
3	Version antigua	Me gustana mas la version anterior esta es mas...	17	NaN	negative	1	-0.616271
4	Esta bien	Sin ser la biblia.... Esta bien	6	negative	negative	1	-0.651784
...
252	Muy buena aplicación	Muy buena genial	3	positive	positive	5	2.814818
253	Buena	Genial	1	positive	positive	5	2.924393

EJERCICIO PYTHON CON SKLEARN

Pasos a realizar:

1. Importar las librerías (numpy, pandas, matplotlib y sklearn)
2. Cargar los datos
3. Generar las estadísticas de los datos (wordcount, Star Rating, sentimentValue)
4. Visualizar los datos con histogramas
5. Observar los registros de la etiqueta (Star Rating)
6. Crear el set de entrenamiento y prueba
7. Usar K-NN con sklearn
8. Generar la matriz de confusión y la precision del modelo

WORKSHOP



GRACIAS.