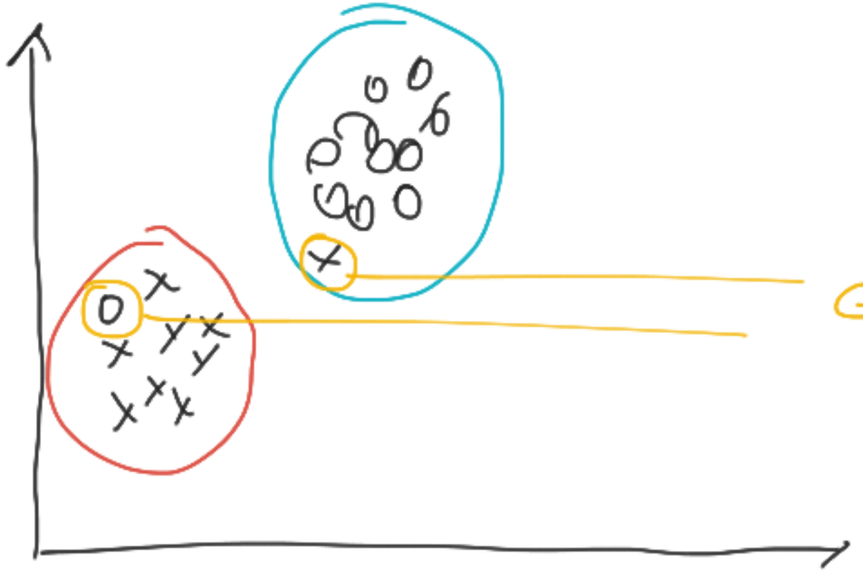
EVALUACIÓN ALGORITMOS DE AGRUPAMIENTO



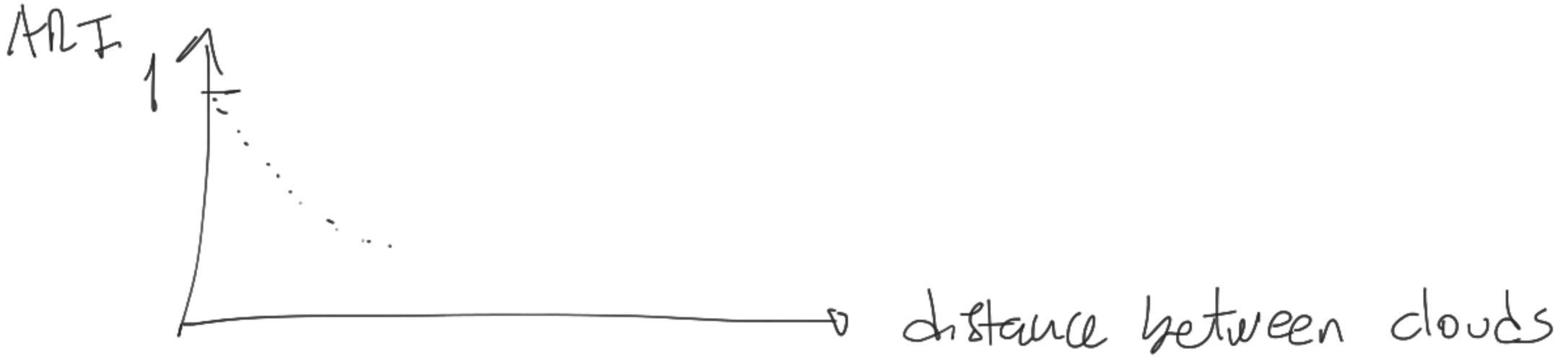
errores de agrupamiento

↳ Adjusted Rand Score

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

Ejercicio: Generar 1000 observaciones 2D
en dos clusters utilizando make_blobs
↓ sklearn

Representar el Adjusted Rand Index
a medida que los centroides se acercan



kj: titanic

's' - survived

'd' - dead

edd dan fee . . .

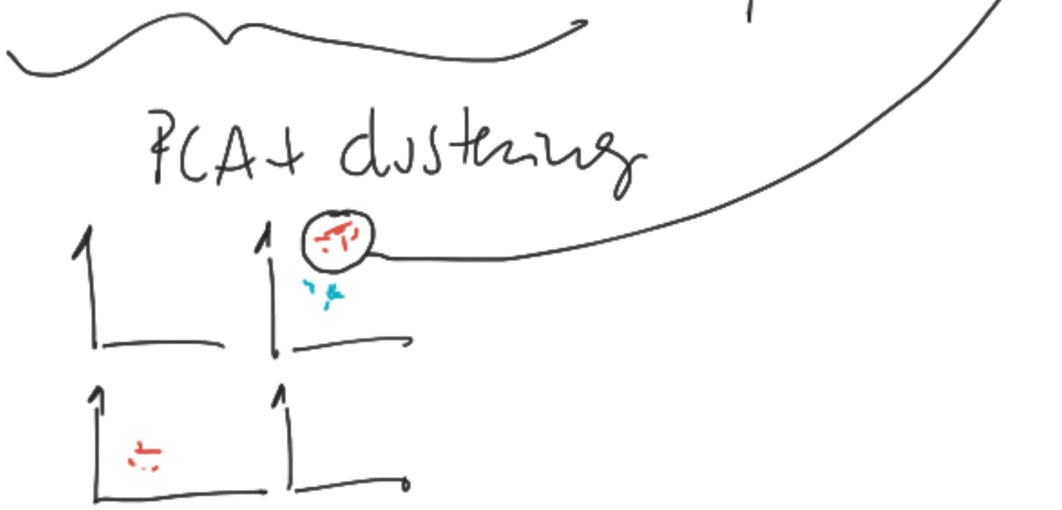
x_1 x_2 . . . x_n | label

's'

'd'

Adjusted Rand Score

post-hoc



SISTEMAS de RECOMENDACIÓN (RECOMMENDERS)

usuarios items valoraciones

Ej: valoración películas

valoración del usuario 1 de la película 2

(rango 1-5)

← valoraciones comunes

Movielens data

100.000 valoraciones

1000 usuarios

1700 películas

https://grouplens.org/datasets/movielens/100k/

Recomendadores basados en distancia entre usuarios (KNN)

K - primeros vecinos de un usuario para recomendar

↳ distancia entre usuarios → valoraciones comunes

→ similaridad Pearson (correlación)

→ similaridad euclídea (a, b):

$d(a,b)$ = distancia euclídea entre las valoraciones comunes

Ex: similaridad entre usuarios 2 y 3

$$s(a,b) = \frac{1}{1+d(a,b)}$$

Usuario 2 → $a = [4,1,2]$

Usuario 3 → $b = [2,1,1]$

$$d(2,3) = \sqrt{(4-2)^2 + (1-1)^2 + (2-1)^2} = \sqrt{4+1} = \sqrt{5} \rightarrow s(2,3) = \frac{1}{1+\sqrt{5}} = 0.31$$

Matriz distancias a pares entre todos los usuarios

d_{ij}

Recomendar una película a usuario r →

buscar los k usuarios más cercanos a r

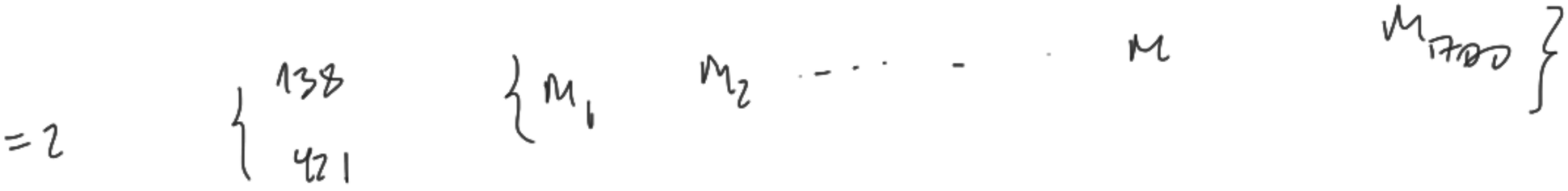
→ RECOMENDACIÓN!

LIBRERÍAS → Surprise

http://surpriselib.com/

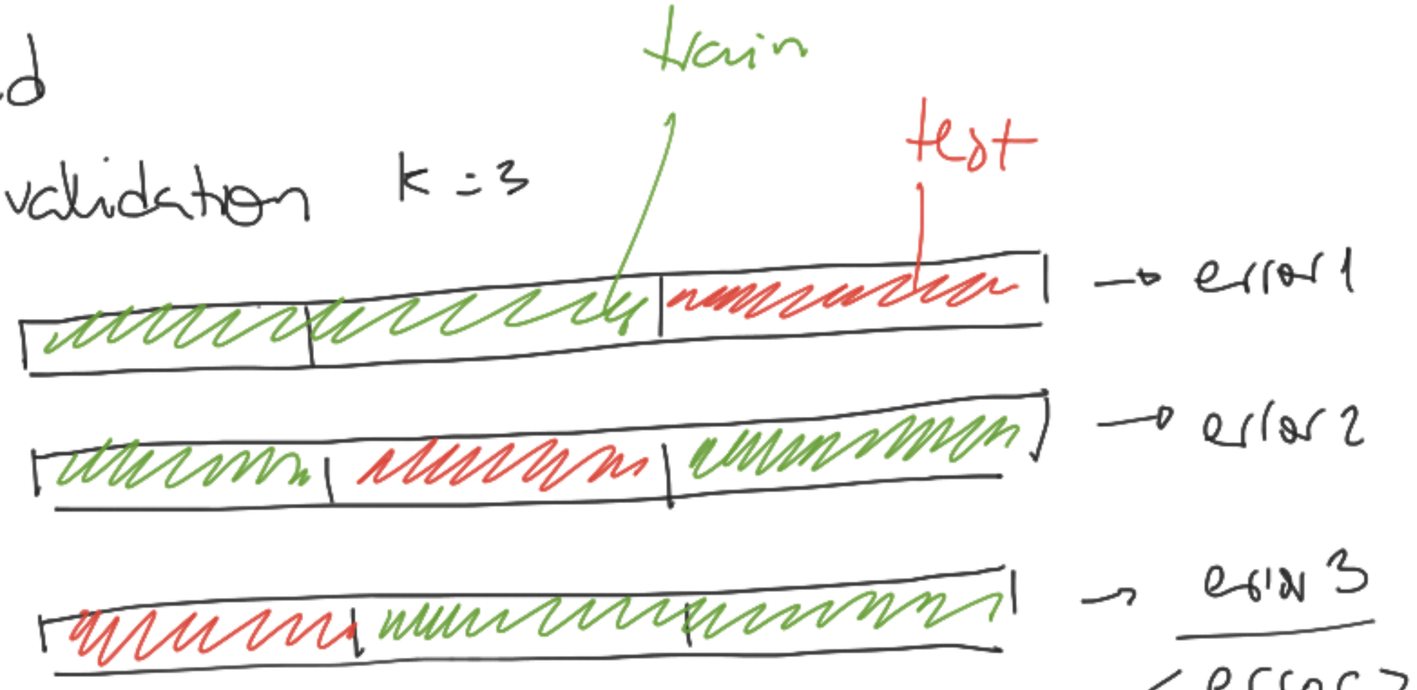
1 2 3 1700

fill with mean of nearest vers



- datos → entrenamiento (≈ 70 - 80 %) → Ajustar el modelo de recomendador → predecir valoraciones
  - (≈ 30 - 20 %) validación | test →
- k-fold

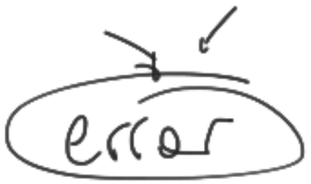
Cross-validation k = 3



$\frac{\text{error 3}}{\langle \text{error} \rangle}$

val (user i, item j)

predicción modelo — valoración real



PROYECTO

(1) ELECCIÓN EL PROBLEMA – CONJUNTO DATOS

(2) MANIPULACIÓN DATOS, FILTRADO, IMPUTACIÓN Y REPRESENTACIÓN GRÁFICA

+CÓDIGO + TEXTO COMENTARIOS

Python notebook
- autoexplicativo
- ejecutable

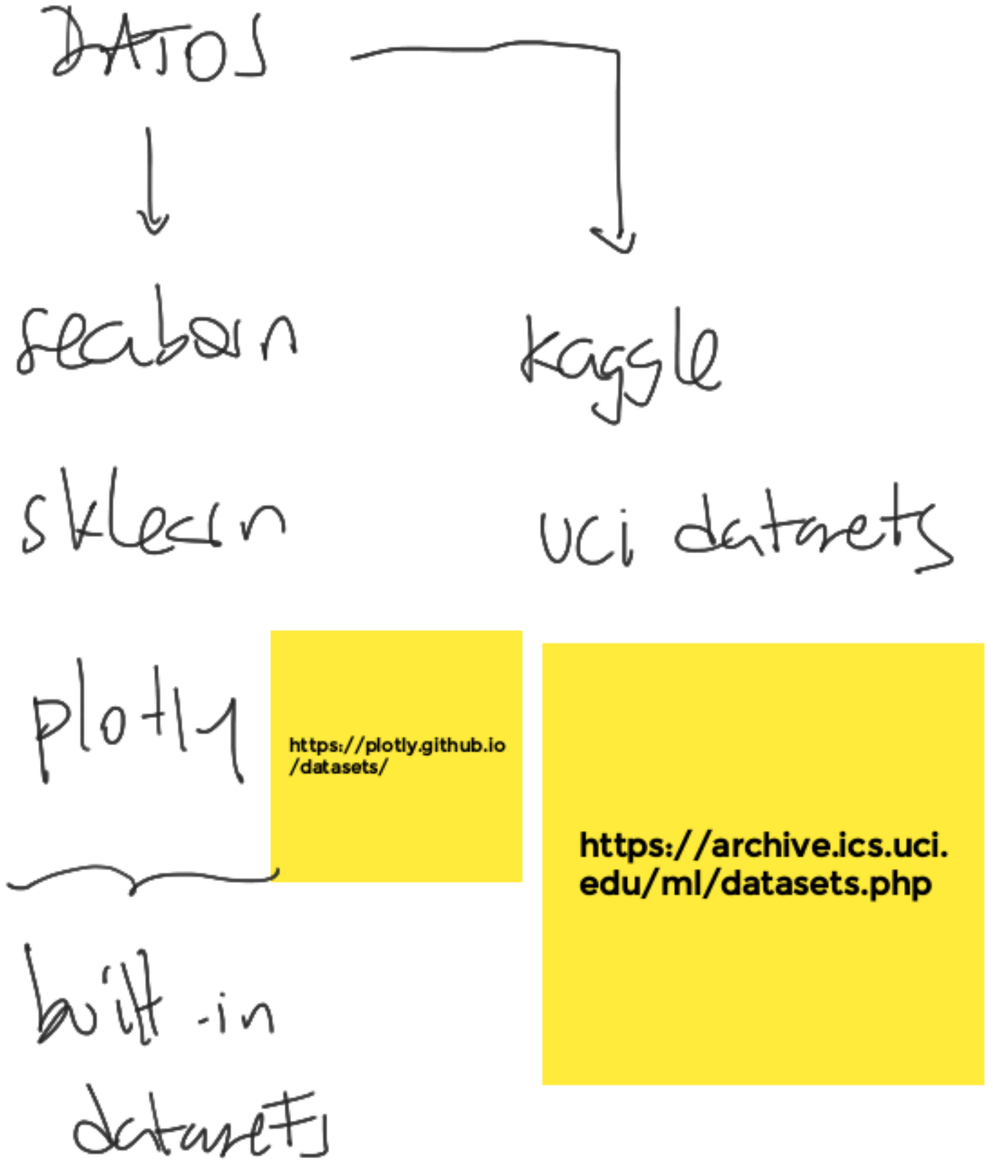
- tamaño
- filtrado - interpretación
- presencia missing data
- tipo de variables
- etiqueta de clase *
- gráficas (scatterplot, histogramas, pairplots ...)

(3) REDUCCIÓN DIMENSIONALIDAD (PCA)

(4) CLUSTERING (K-means, aglomerativo, GMM, GMM+BIC)

(5) SELECCIÓN ATRIBUTOS * (RFE, univariada)

(6) REGRESIÓN UNIVARIADA (LINEAL | ÁRBOLES DECISIÓN)


DATOS
seaborn
sklearn
plotly
kaggle
uci datasets
built-in
datasets
https://plotly.github.io/datasets/
https://archive.ics.uci.edu/ml/datasets.php

EJERCICIOS - MISCELÁNEA:

① ¿Cuántos supervivientes mayores de 40 años hay en el titanic dataset?

② Cuál es el adjusted Rand Score de un clustering k-means (k=3) de los datos de mpg-cars con respecto a la etiqueta de origen de los coches (ground truth).

③ cuál es la variable que mejor predice la supervivencia en el titanic y la que mejor predice el origen en los datos mpg.

④ cuántas observaciones del iris dataset tienen un petal length fuera del rango intercuartil $IDR = [Q_1, Q_3]$

representar el histograma con