

HDBSCAN

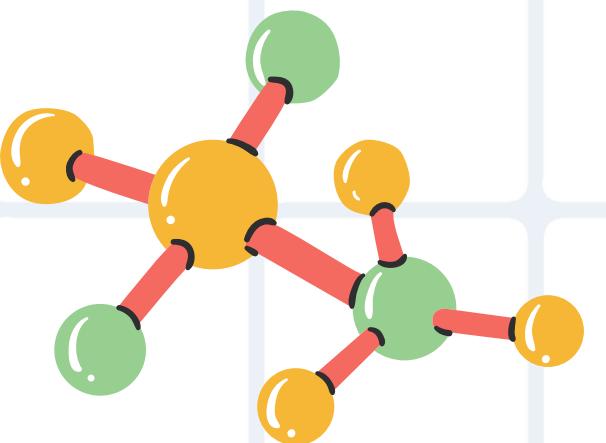
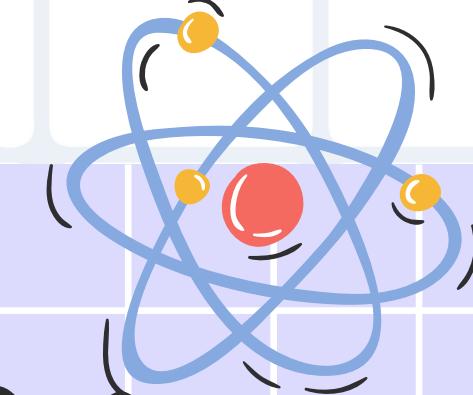
AGRUPAMIENTO JERÁRQICO
BASADO EN DENSIDADES

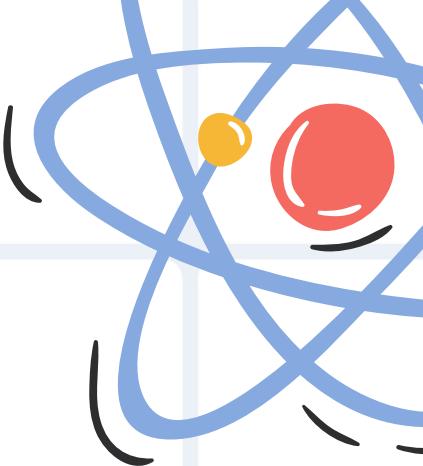
INTEGRANTES:

JOAN SEBASTIÁN FRANCO

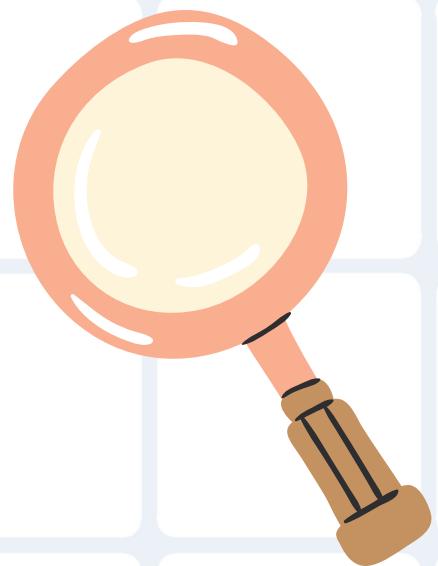
DAVID FERNANDOPARADA

ROSMER MANUEL VARGAS





SELECCIÓN DEL TEMA



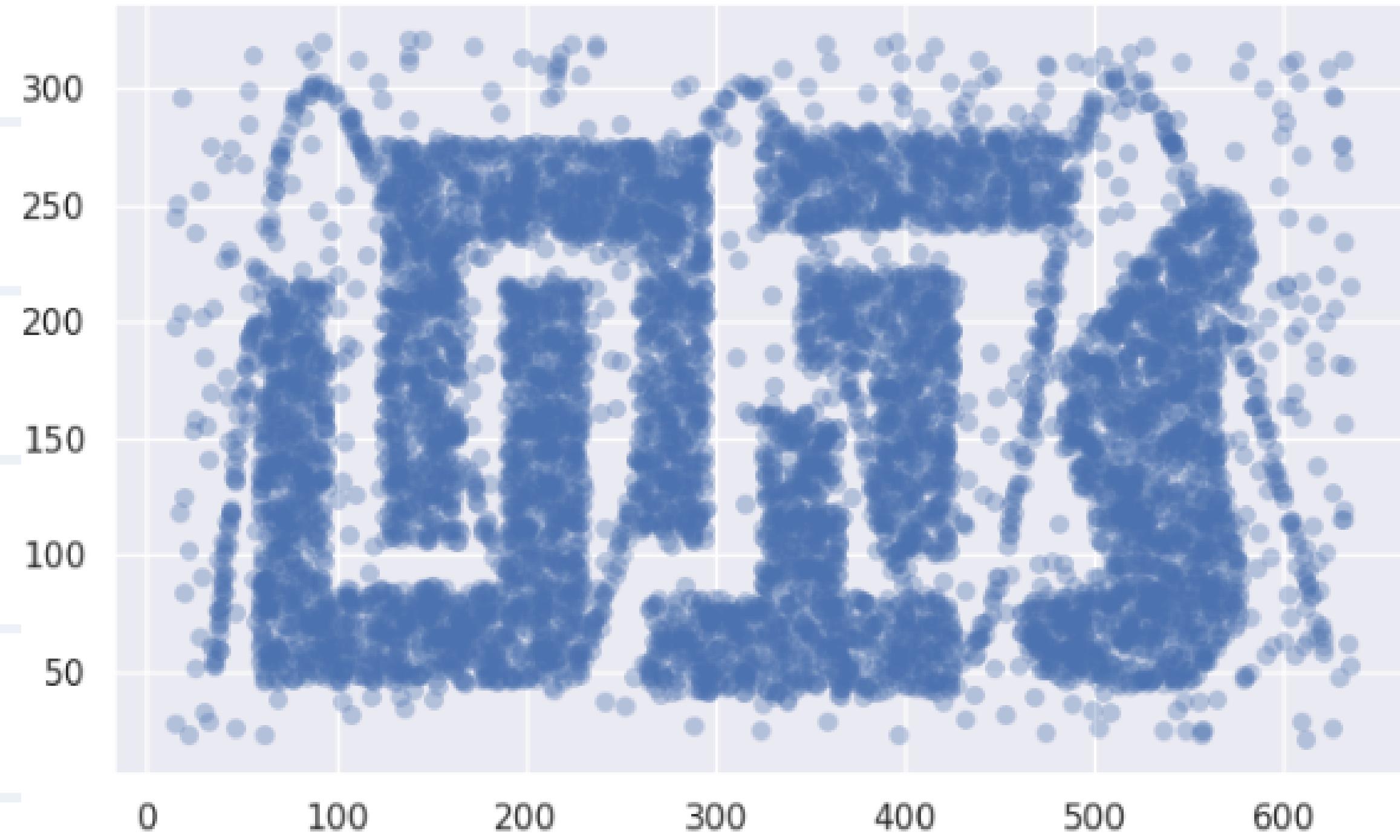
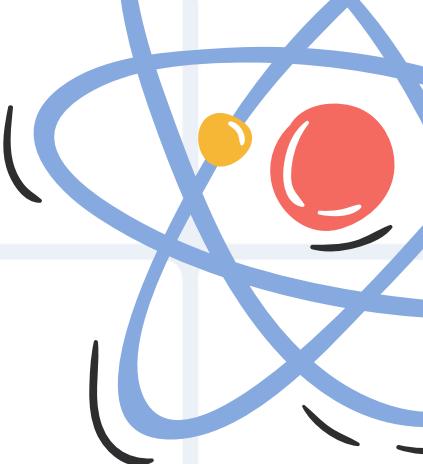
● HDBSCAN

- BREVE INTRODUCCIÓN
- EXPLICACIÓN DEL ALGORÍTMO
- VENTAJAS Y DESVENTAJAS
- EJEMPLO MNIST

● CASO APLICADO REAL: ICFES

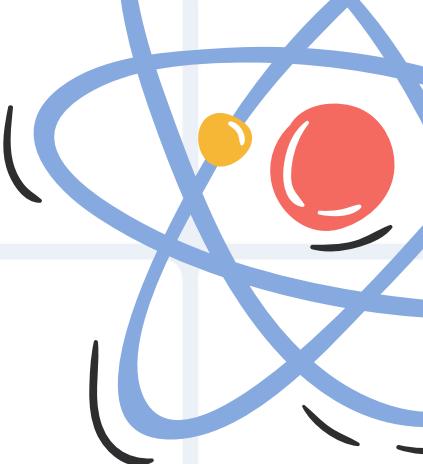


MOTIVACIÓN: CONSIDEREMOS EL SIGUIENTE CONJUNTO DE DATOS. ¿QUÉ SE PUEDE OBSERVAR?

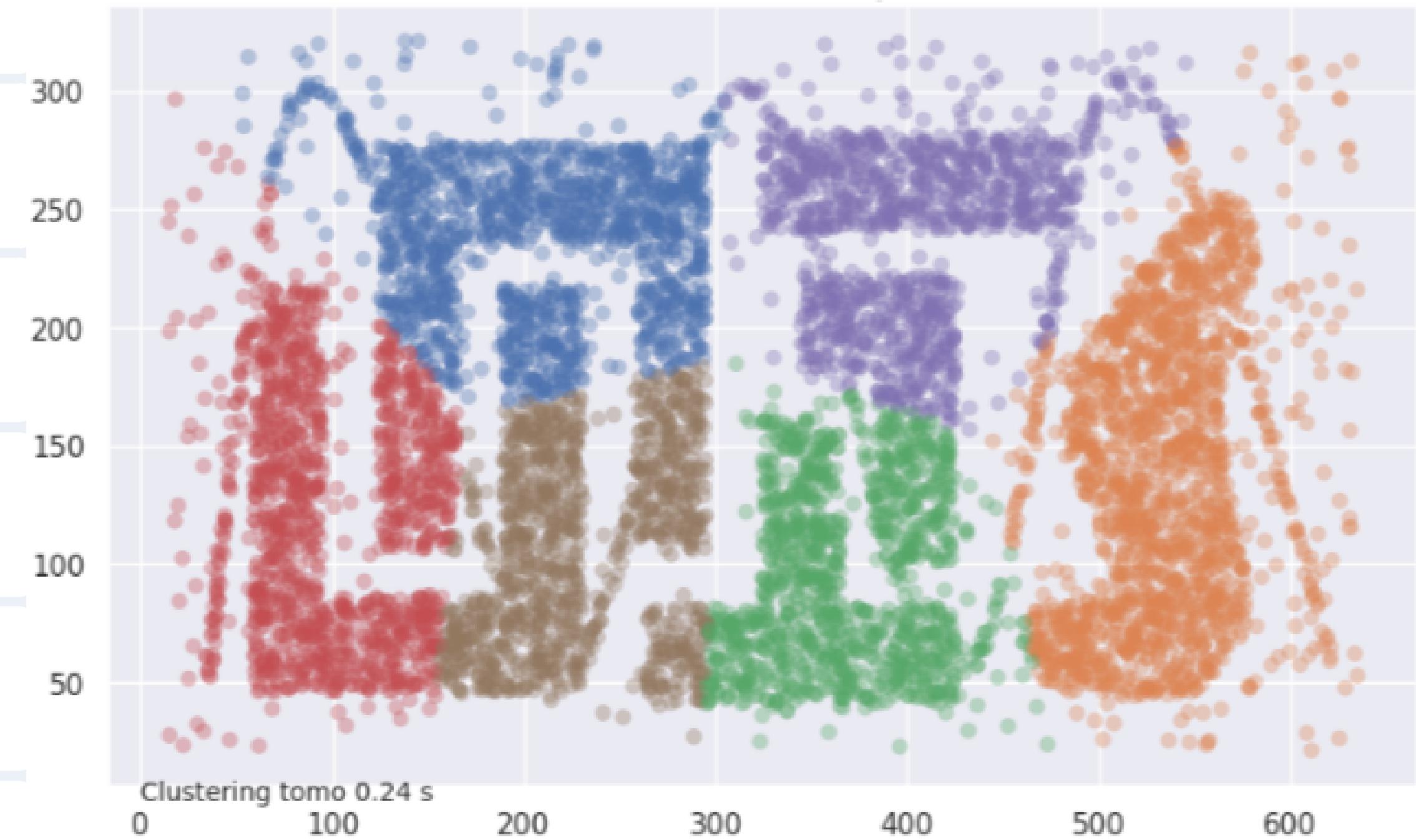




MIREMOS LOS GRUPOS QUE ENCUENTRA EL MÉTODO K-MEANS

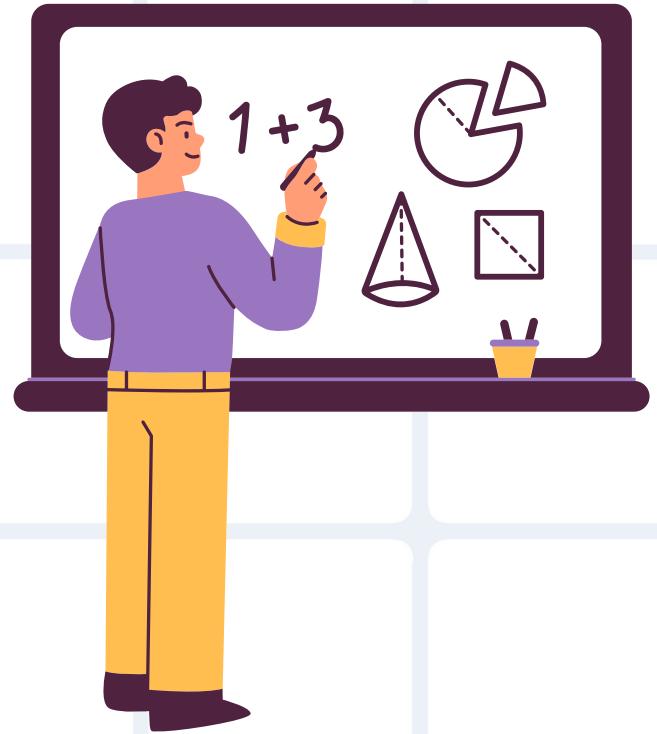
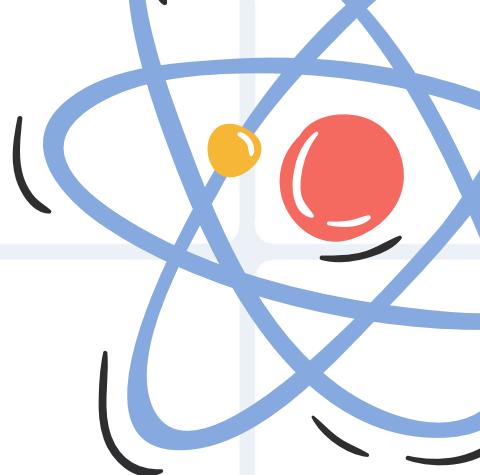


Clusters encontrados por KMeans





MIREMOS LOS GRUPOS QUE ENCUENTRA EL MÉTODO K-MEANS



CLARAMENTE, K-MEANS FALLA PARA DETECTAR LOS PATRONES PRESENTES EN EL CONJUNTO DE DATOS. SE HACE NECESARIA UNA TÉCNICA QUE PUEDA ENCONTRAR DICHOS PATRONES

HDBSCAN

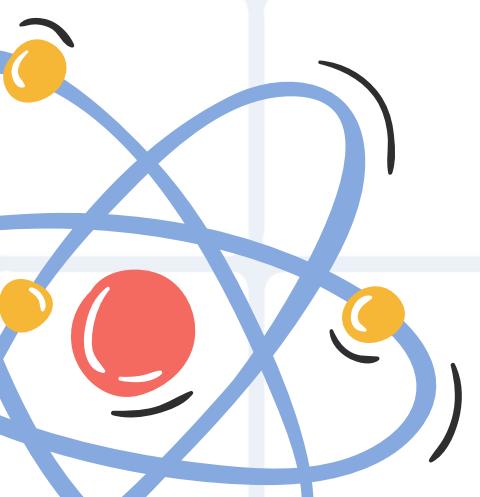
ES UNO DE LOS ALGORÍTROS
DE CLUSTERING MÁS
AVANZADOS, ESTA BASADO
EN DENSIDAD, ESTO SIGNIFICA
QUE NO USAN LAS DISTANCIAS
ENTRE PUNTOS A LA HORA DE
REALIZAR LOS CLUSTERS
COMO POR EJEMPLO PASA
CON K-MEANS



DENSIDAD

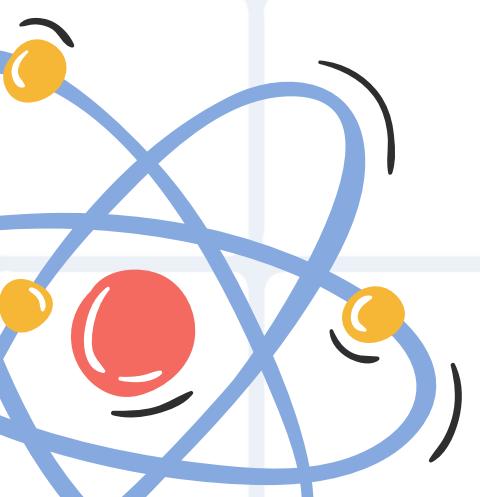
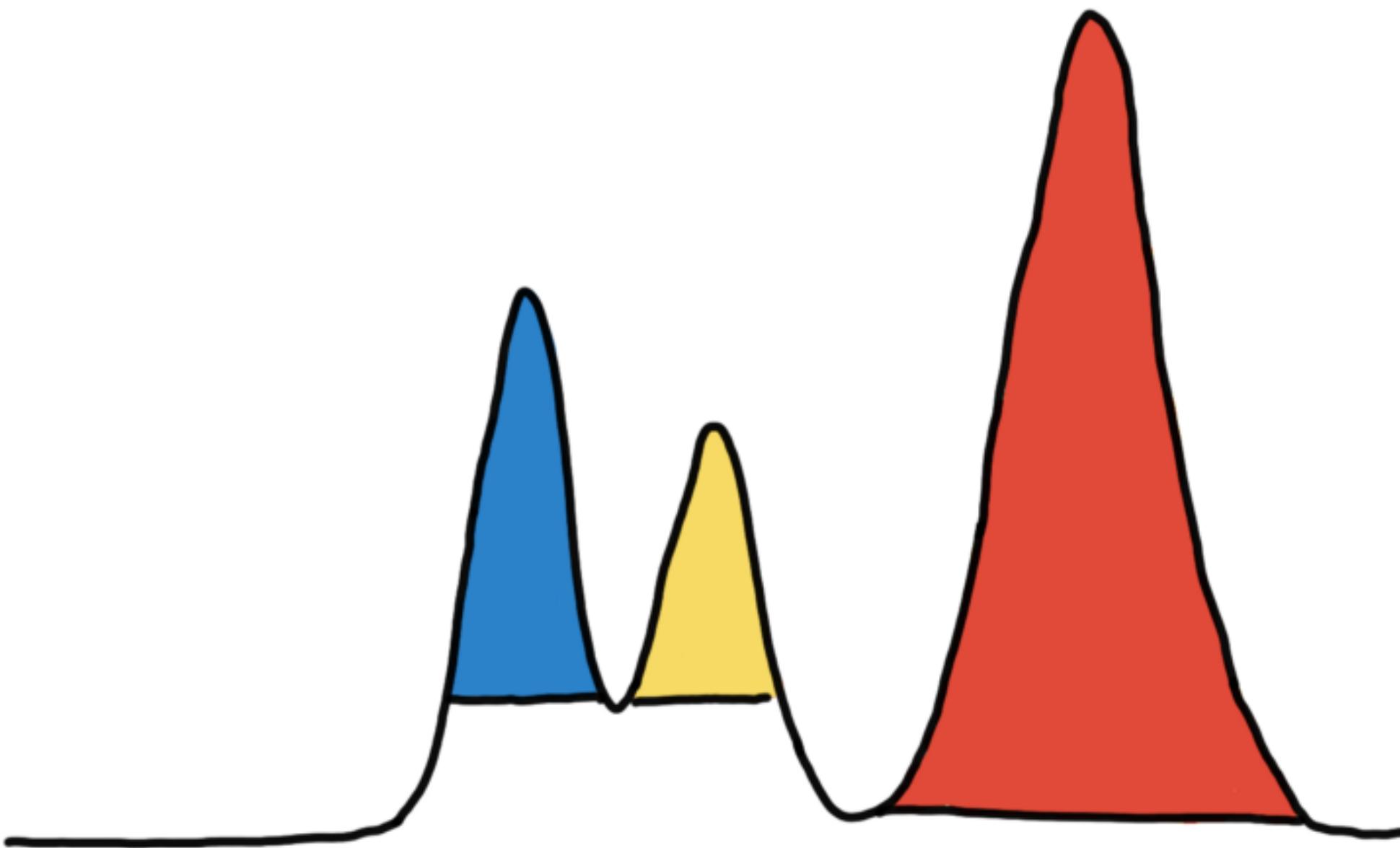
ROBUSTEZ

HDD



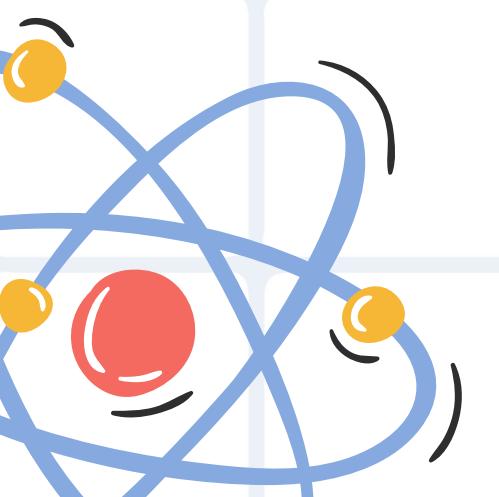
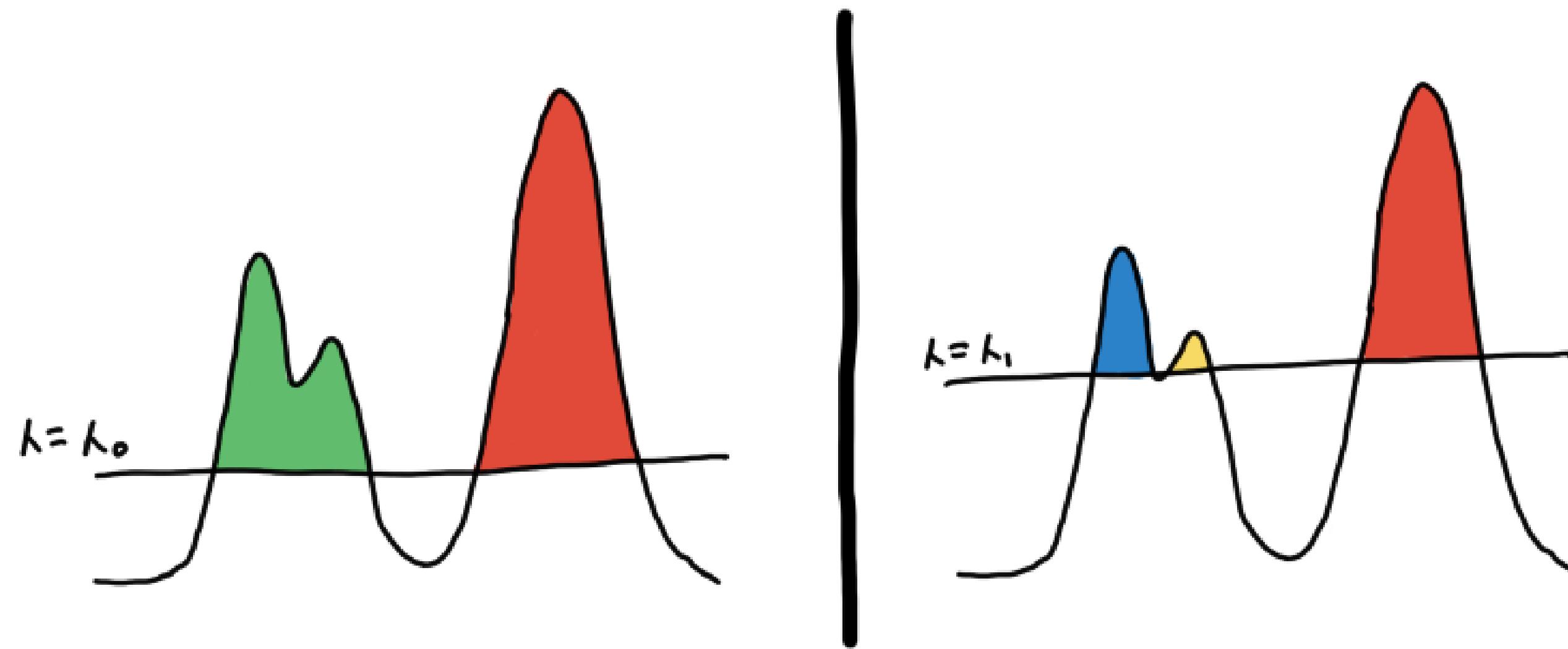
HDBSCAN

BUSQUEDA DE LA DISTRIBUCIÓN SUBYACENTE



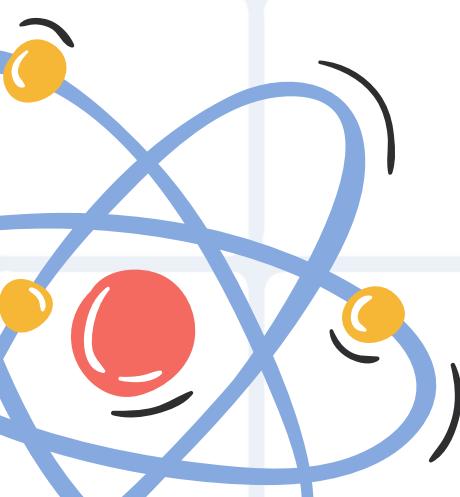
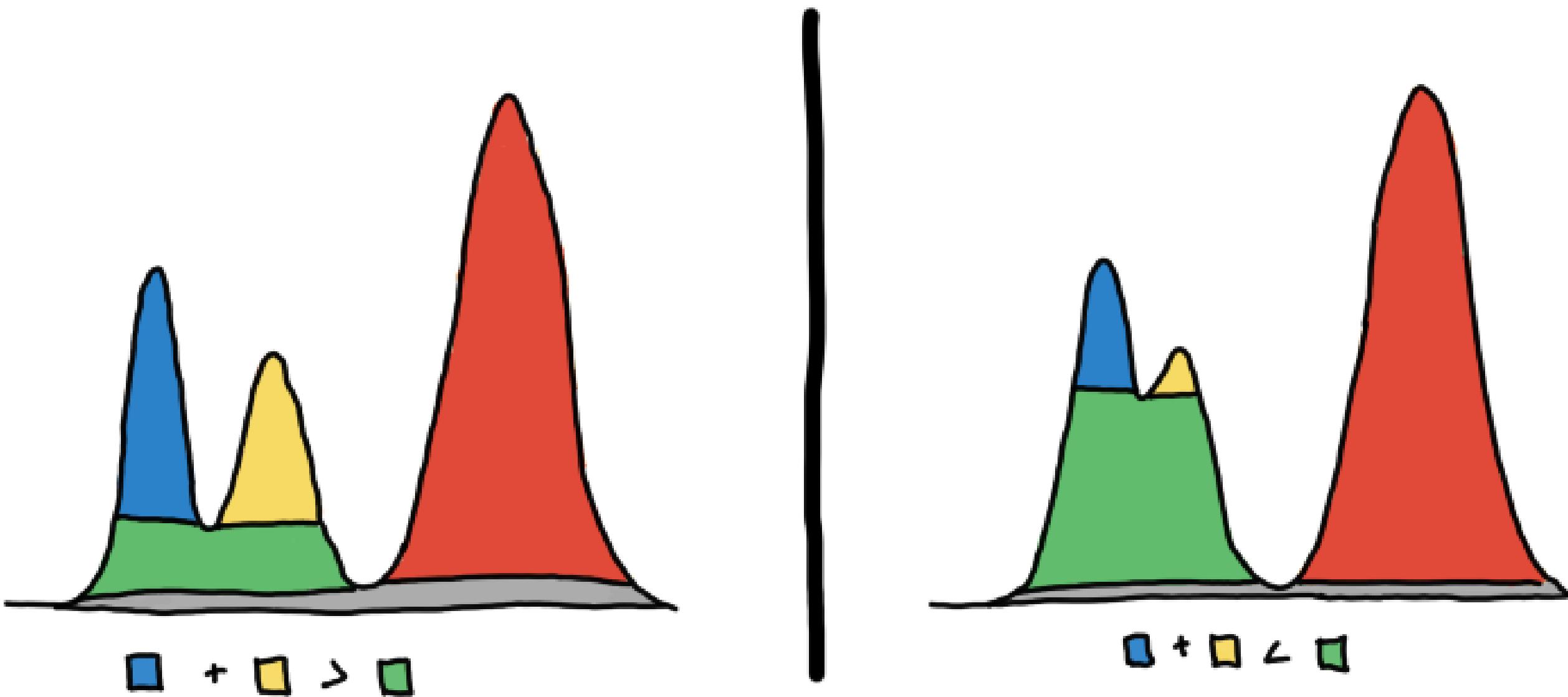
HDBSCAN

BUSQUEDA DE LA DISTRIBUCIÓN SUBYACENTE



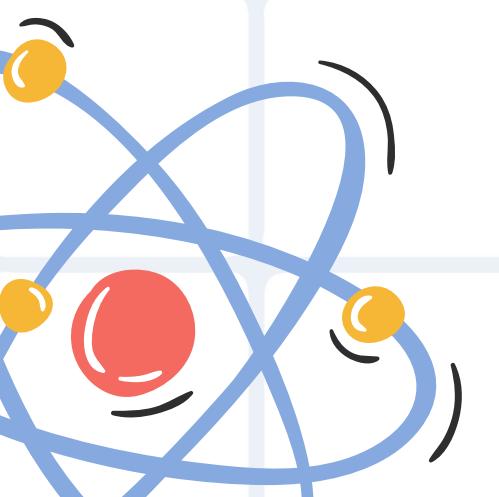
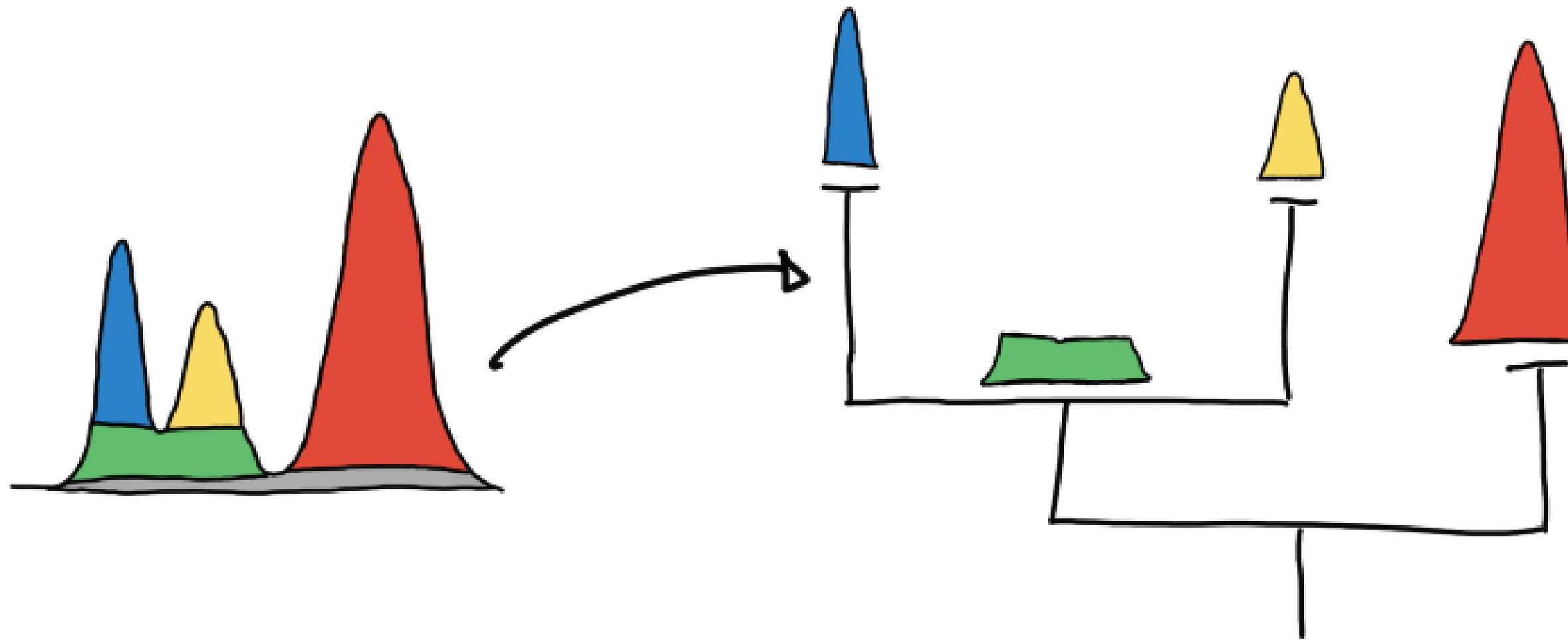
HDBSCAN

BUSQUEDA DE LA DISTRIBUCIÓN SUBYACENTE



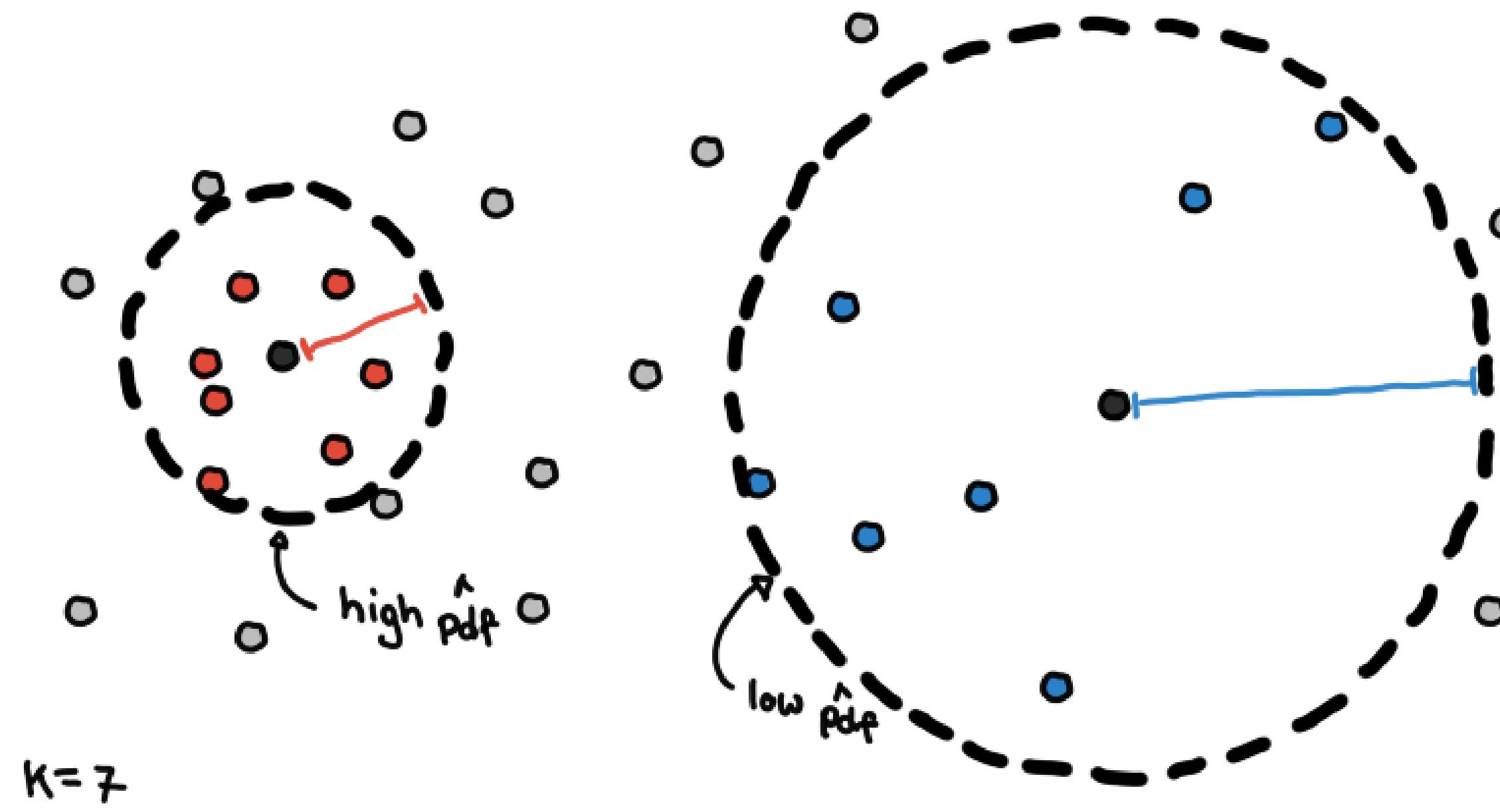
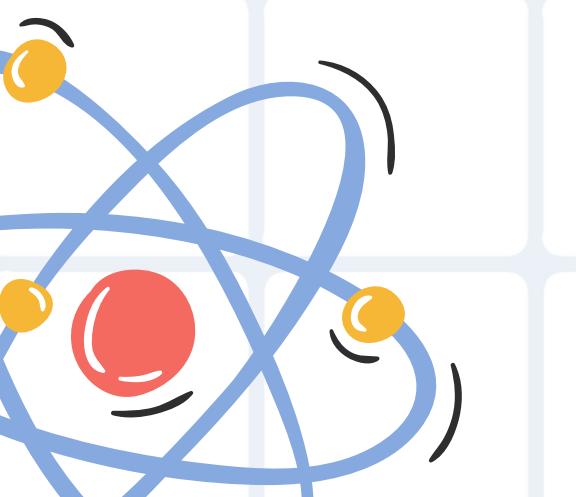
HDBSCAN

CONSTRUYENDO LA JERARQUÍA



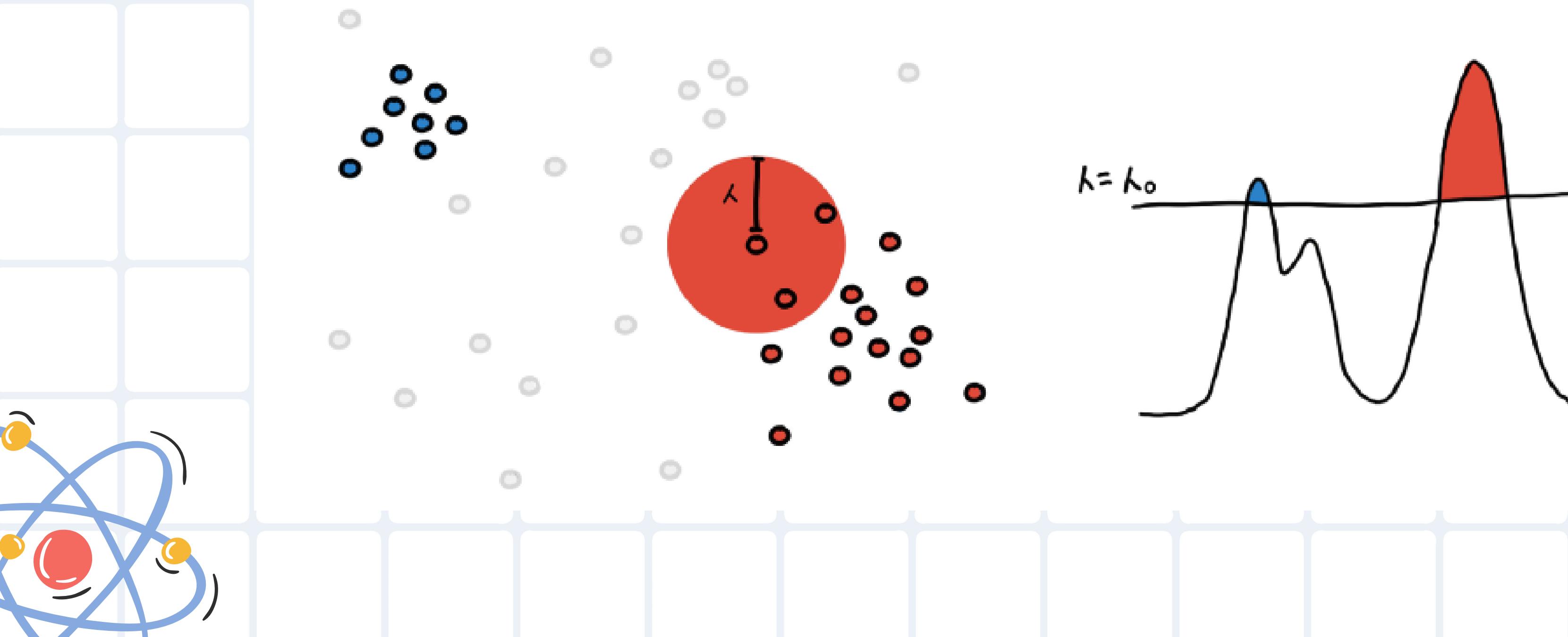
HDBSCAN

DISTANCIA AL K-ÉSIMO VECINO MÁS CERCANO



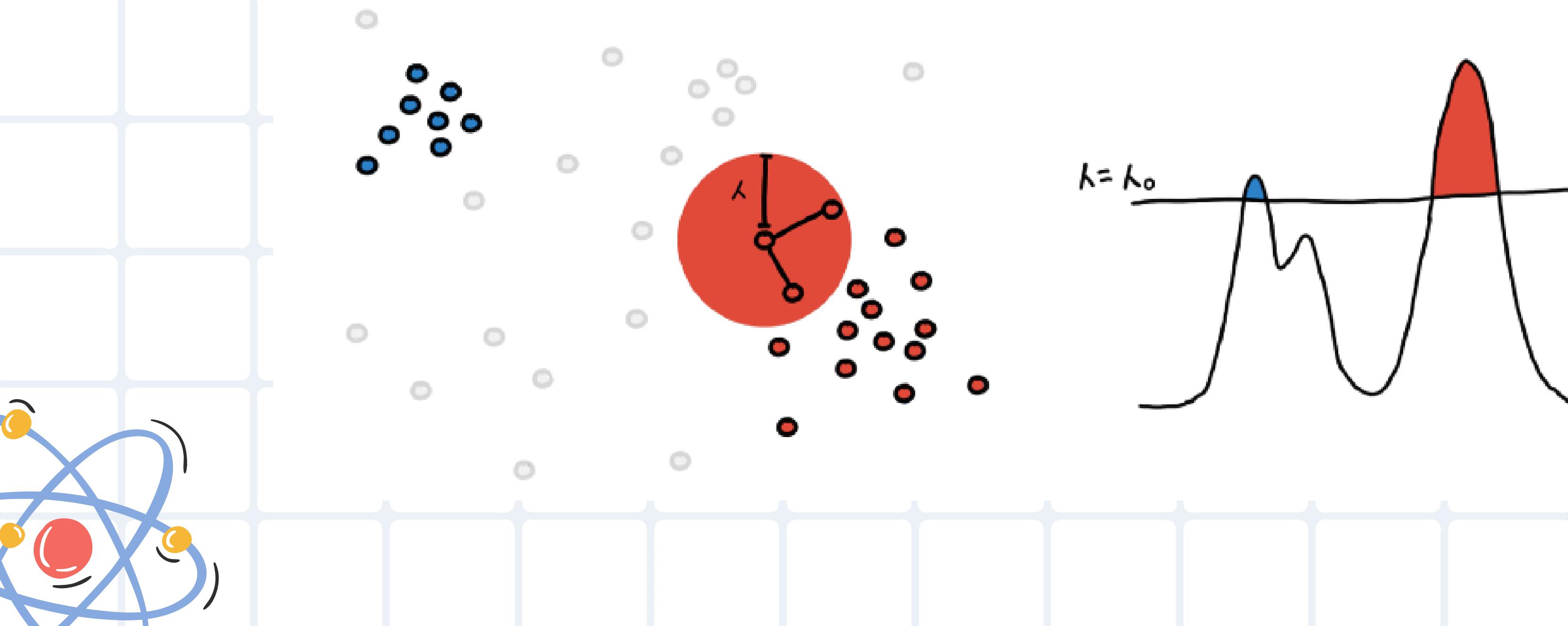
HDBSCAN

DISTANCIA AL K-ÉSIMO VECINO MÁS CERCANO



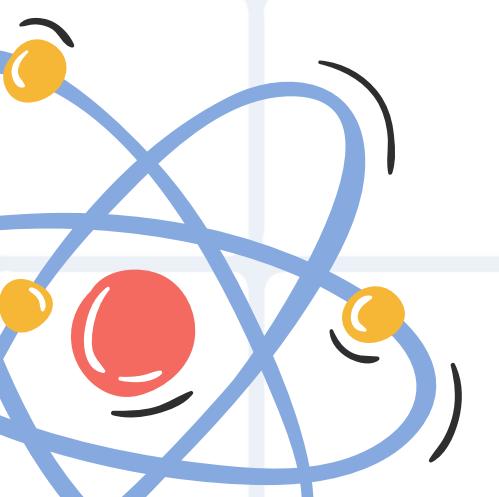
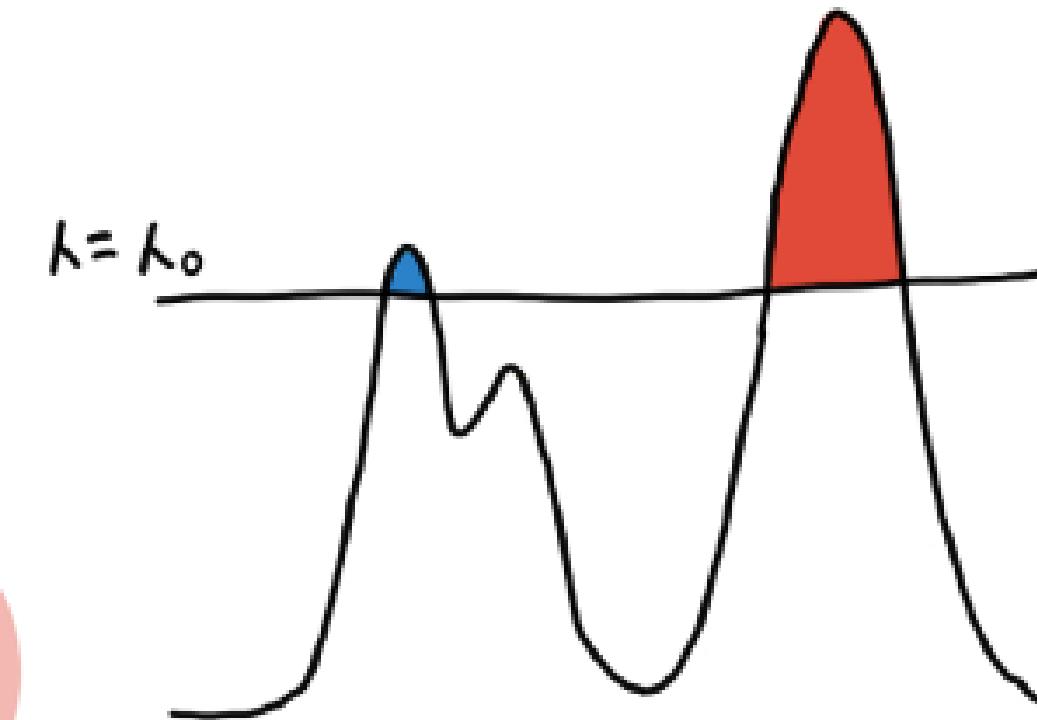
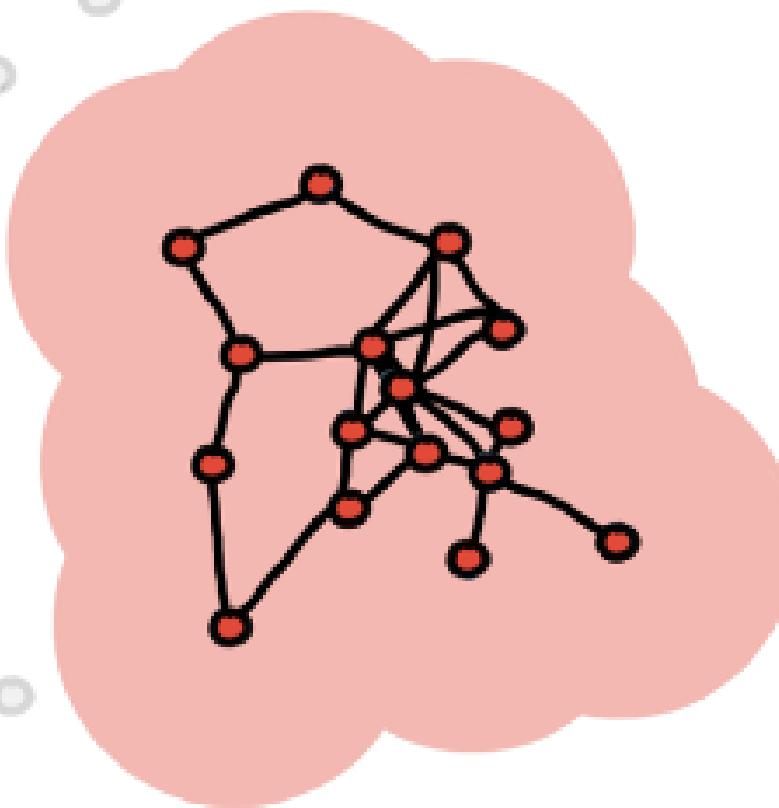
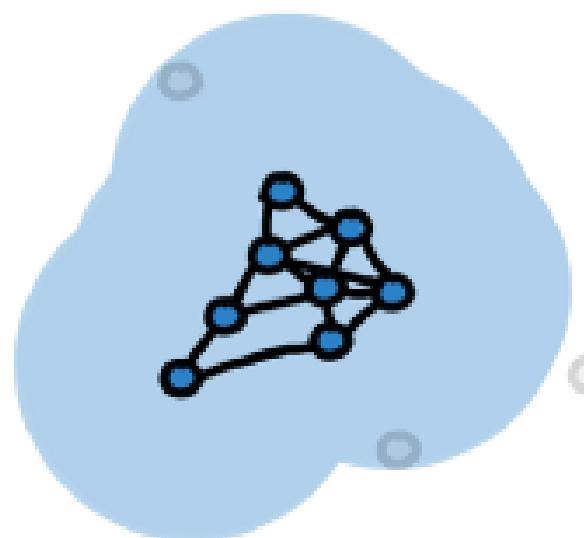
HDBSCAN

DISTANCIA AL K-ÉSIMO VECINO MÁS CERCANO

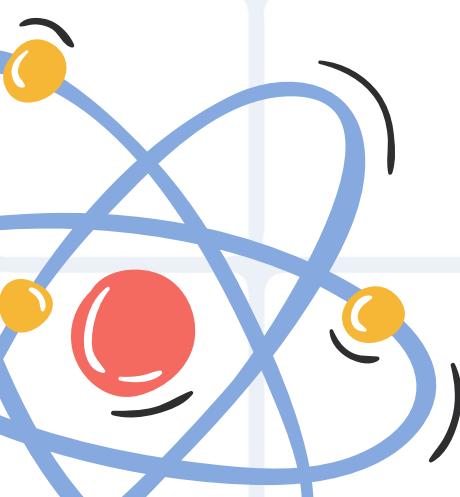
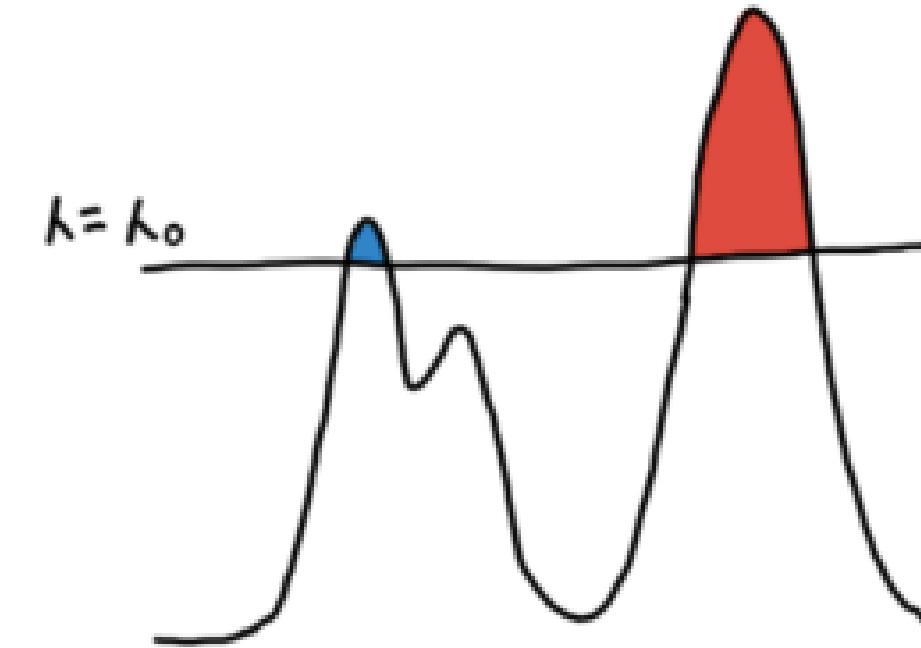
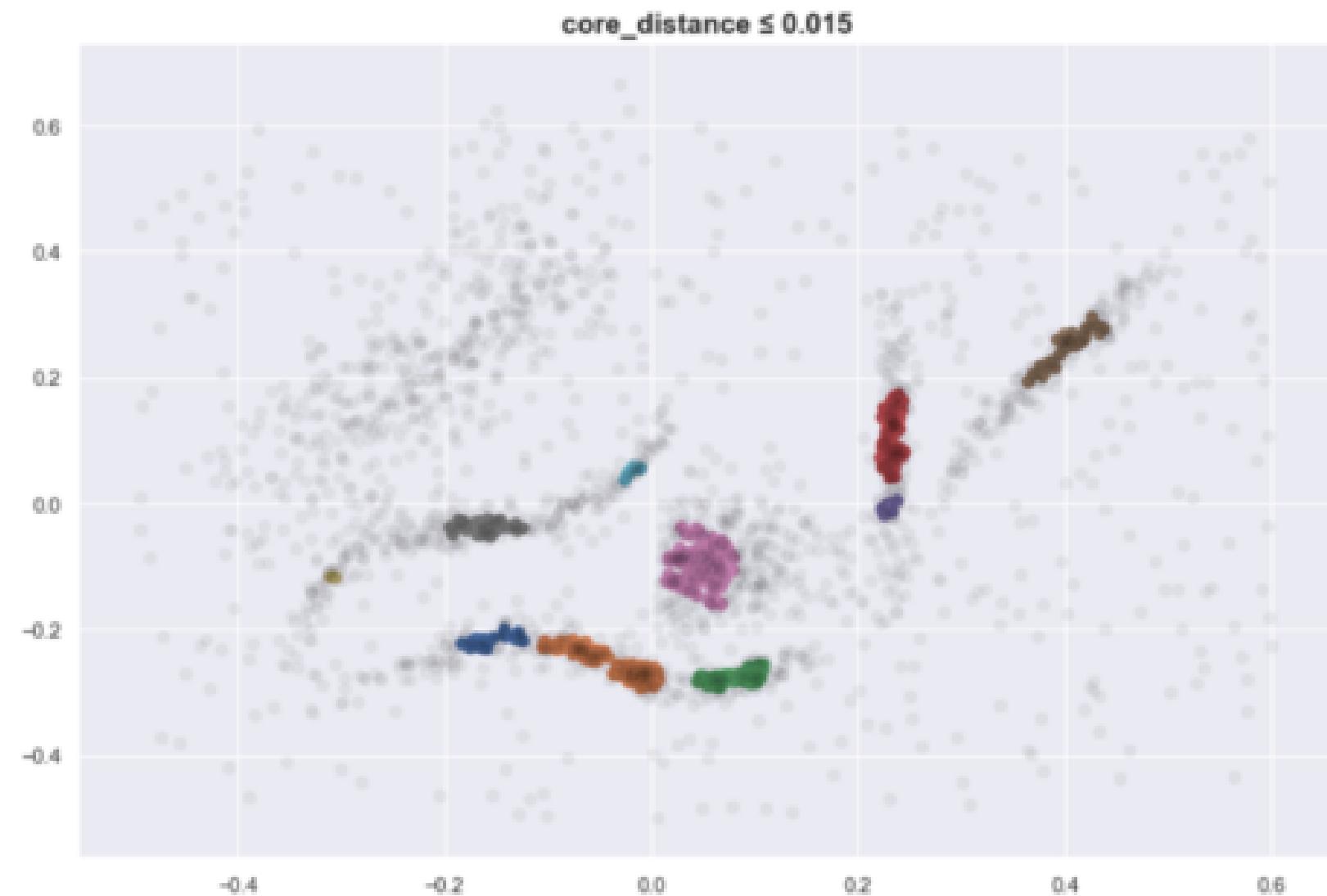


HDBSCAN

DISTANCIA AL K-ÉSIMO VECINO MÁS CERCANO



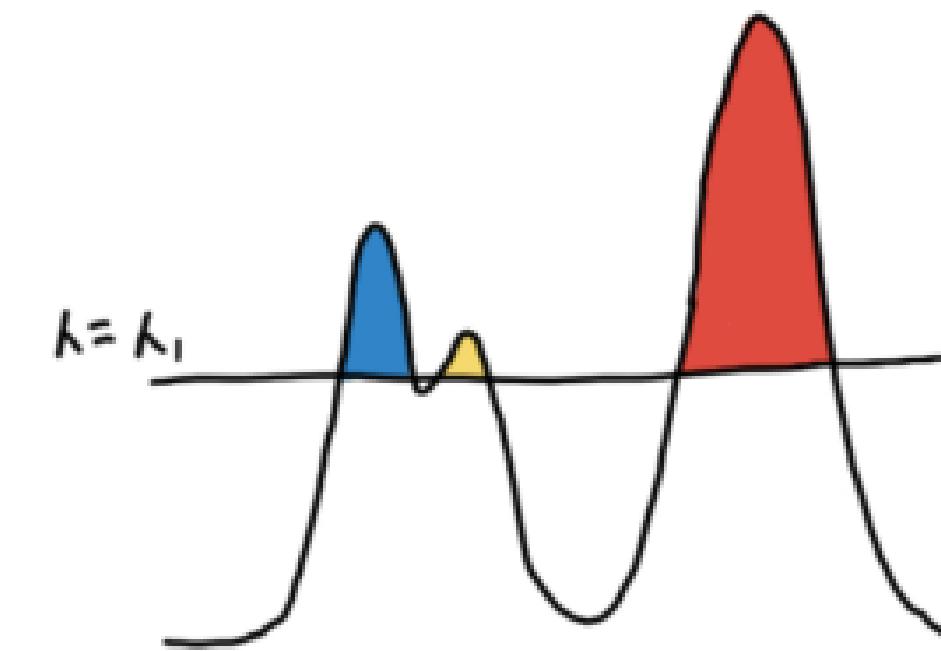
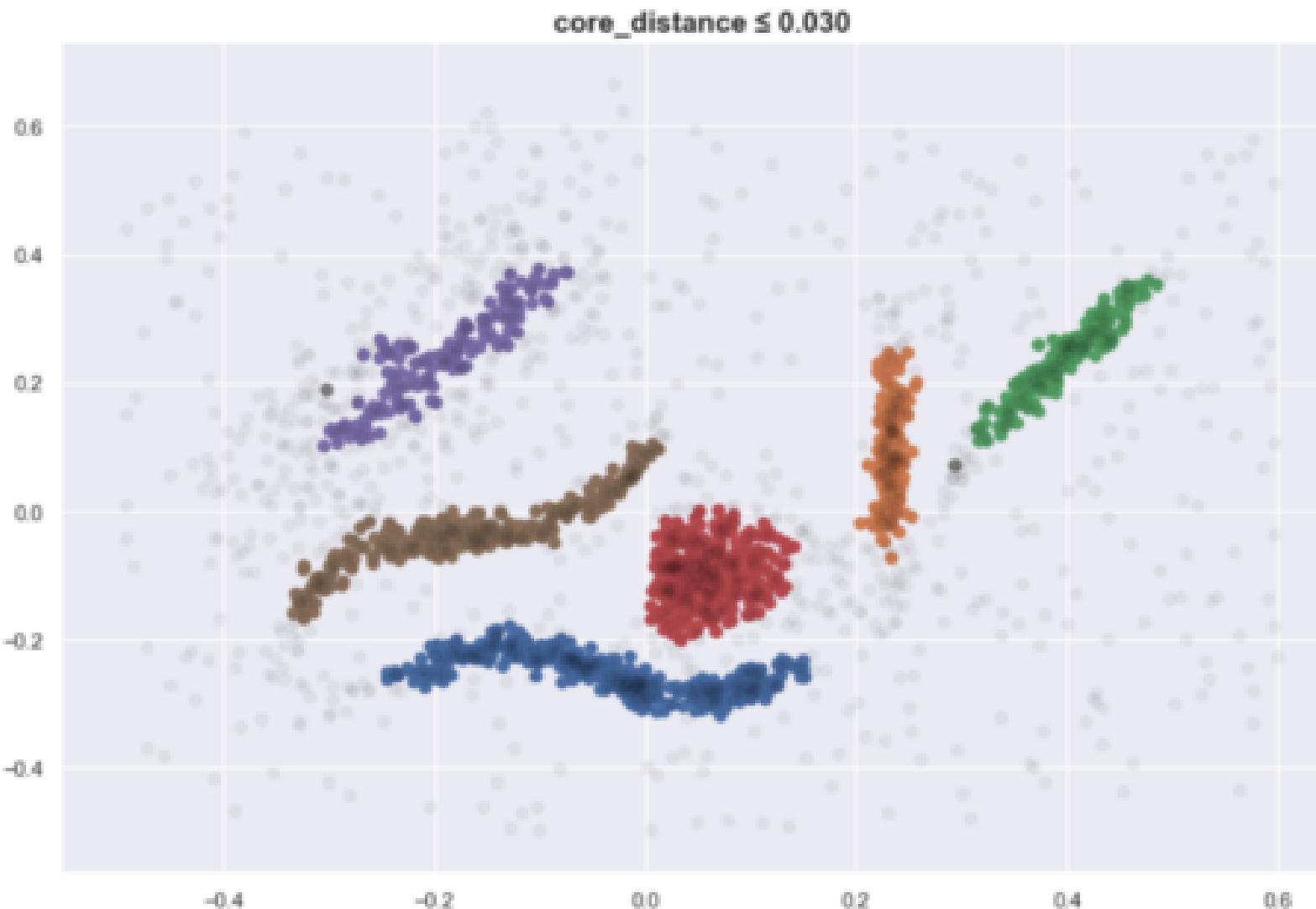
HDBSCAN



ROBUSTEZ

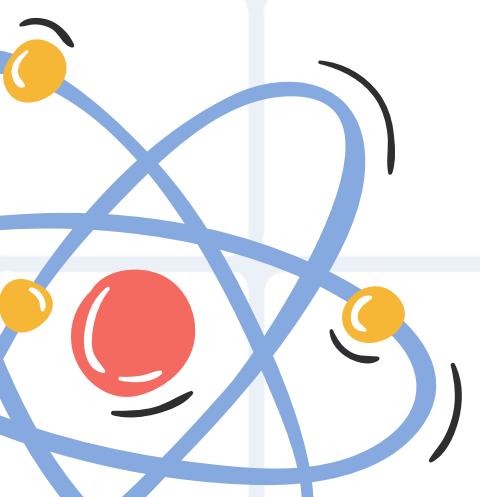
DENSIDAD

HDBSCAN

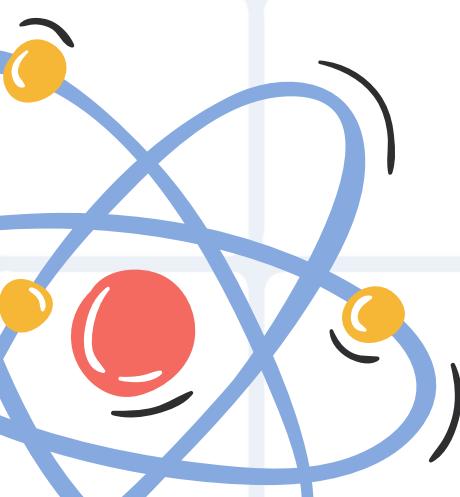
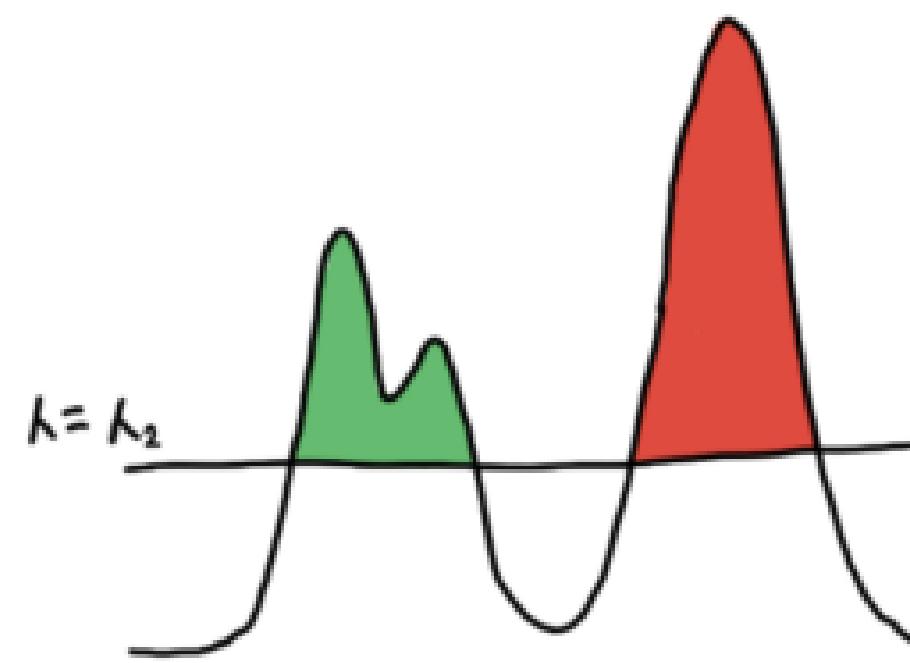
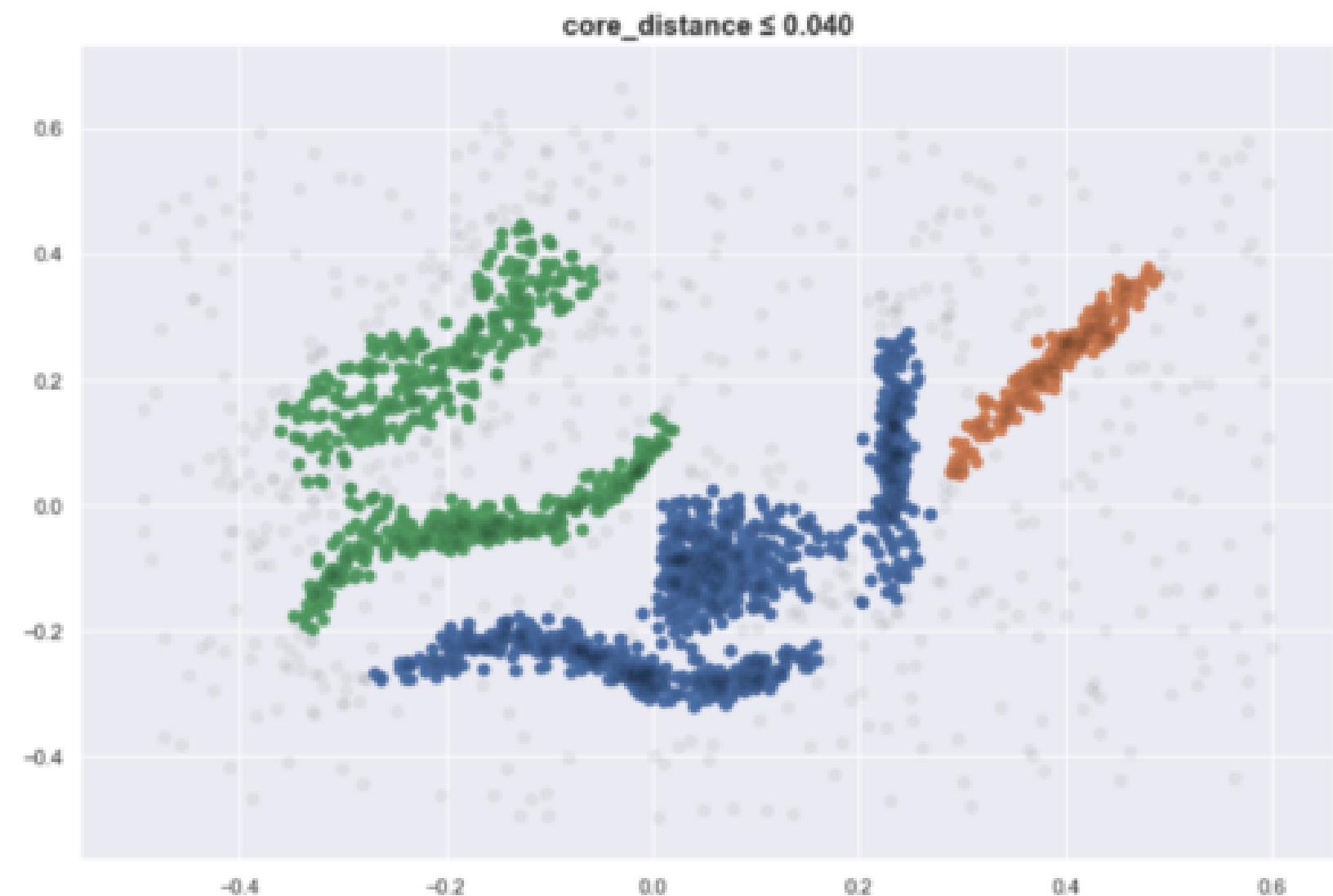


ROBUSTEZ

DENSIDAD



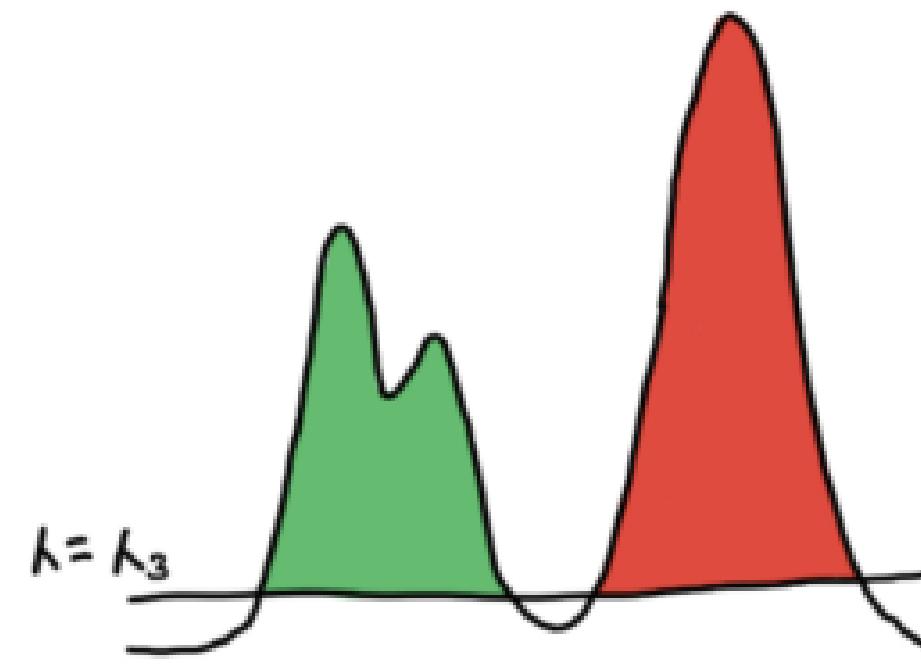
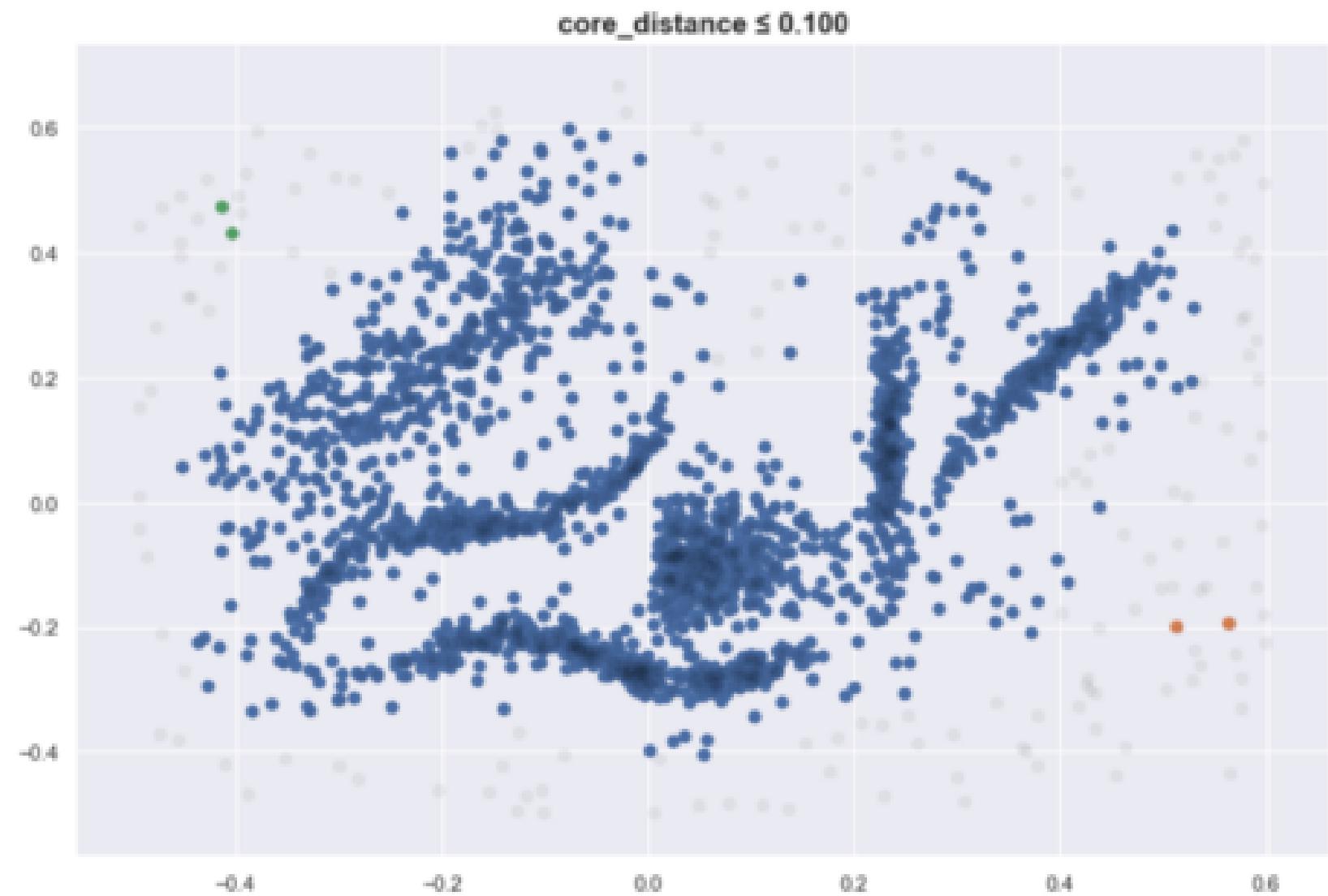
HDBSCAN



ROBUSTEZ

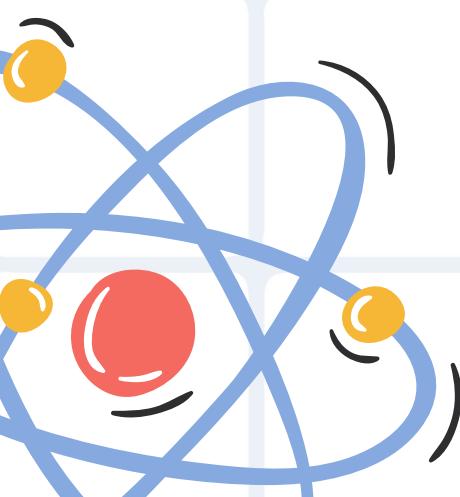
DENSIDAD

HDBSCAN

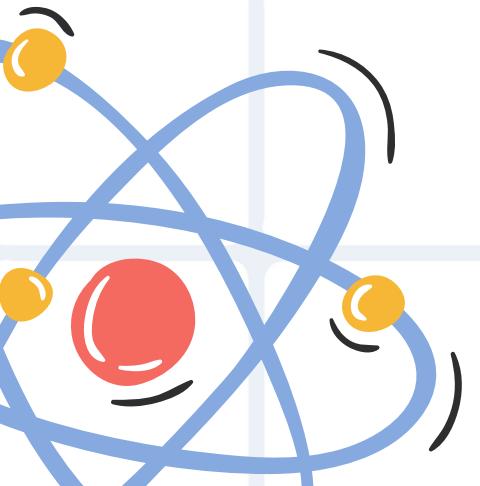
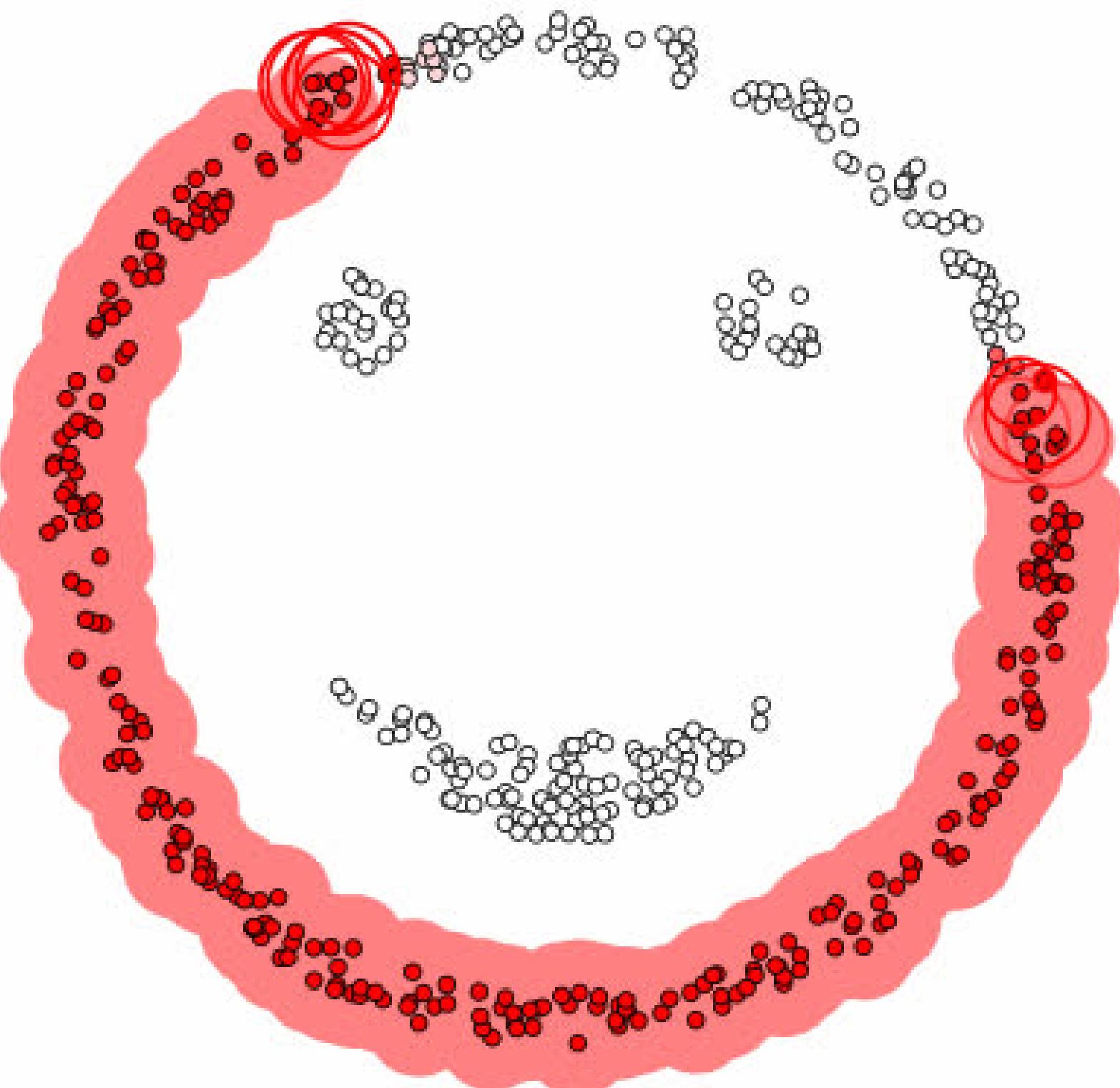


ROBUSTEZ

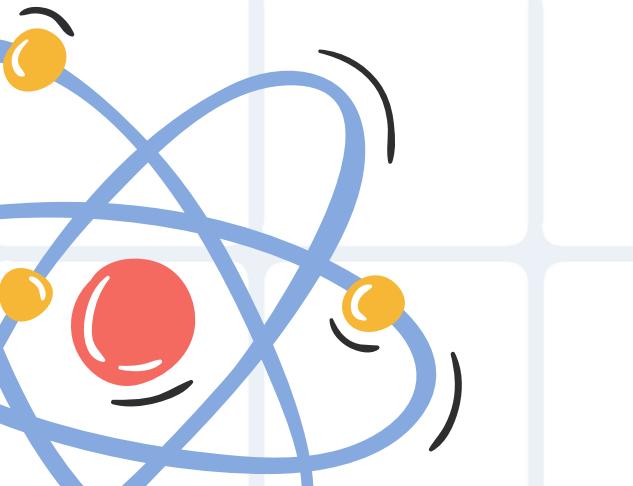
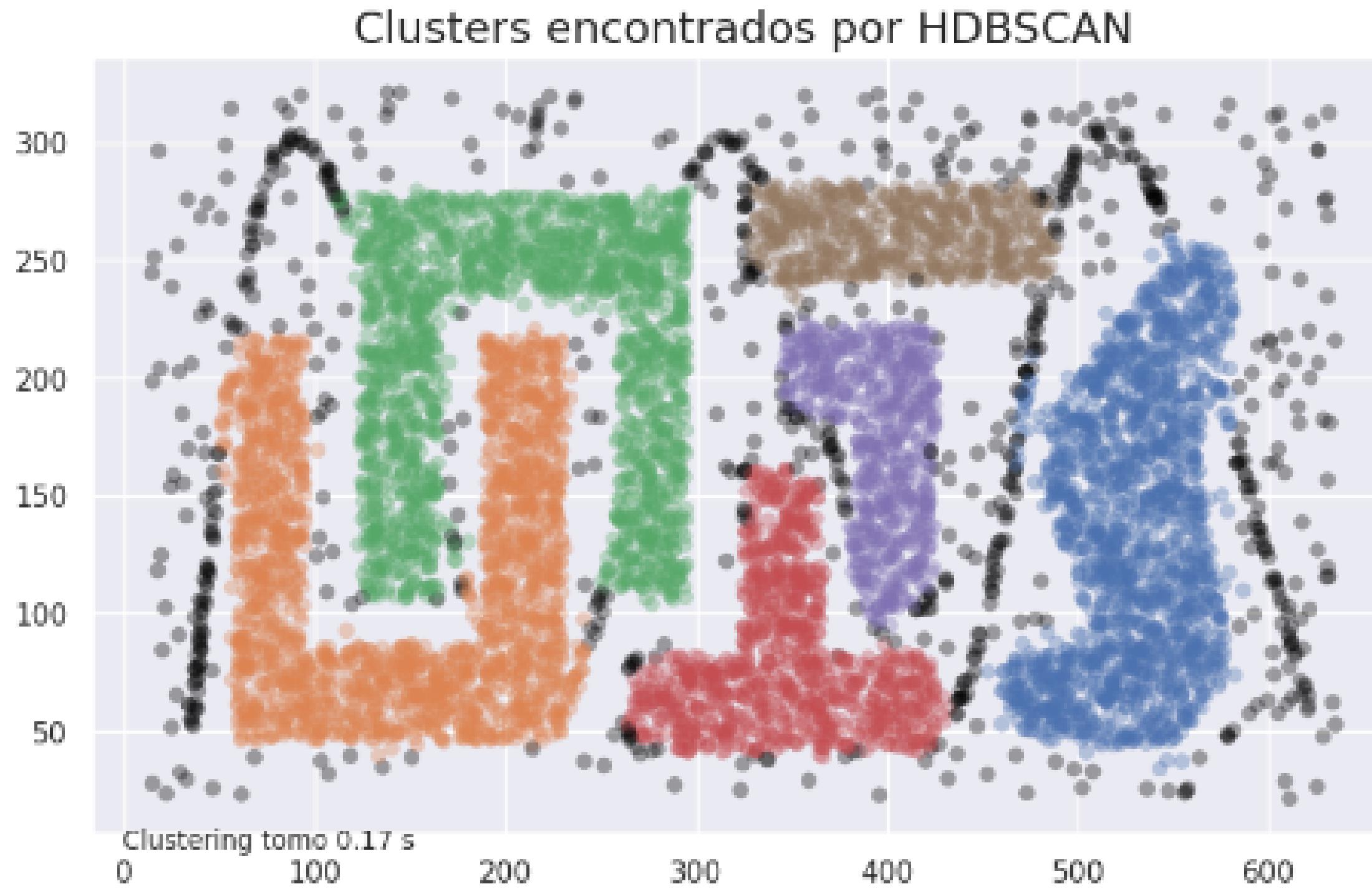
DENSIDAD

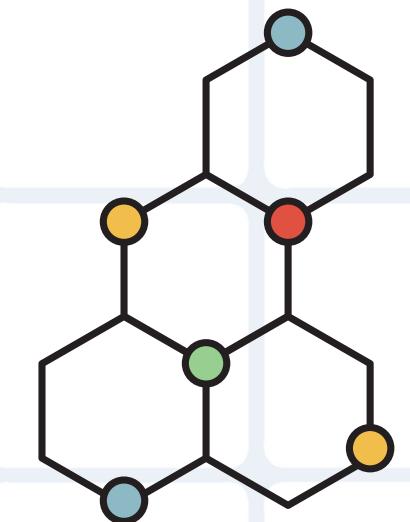


FUNCIONAMIENTO DEL HDBSCAN



EJEMPLO INICIAL





VENTAJAS Y DESVENTAJAS



UBICA LOS GRUPOS BASÁNDOSE EN LA DENSIDAD.



ENCUENTRA LOS INDIVIDUOS QUE SEGÚN SUS CARACTERÍSTICAS NO PERTENEcen A NINGÚN GRUPO.



MENOS PARÁMETROS Y MÁS INTUITIVOS DE AJUSTAR.

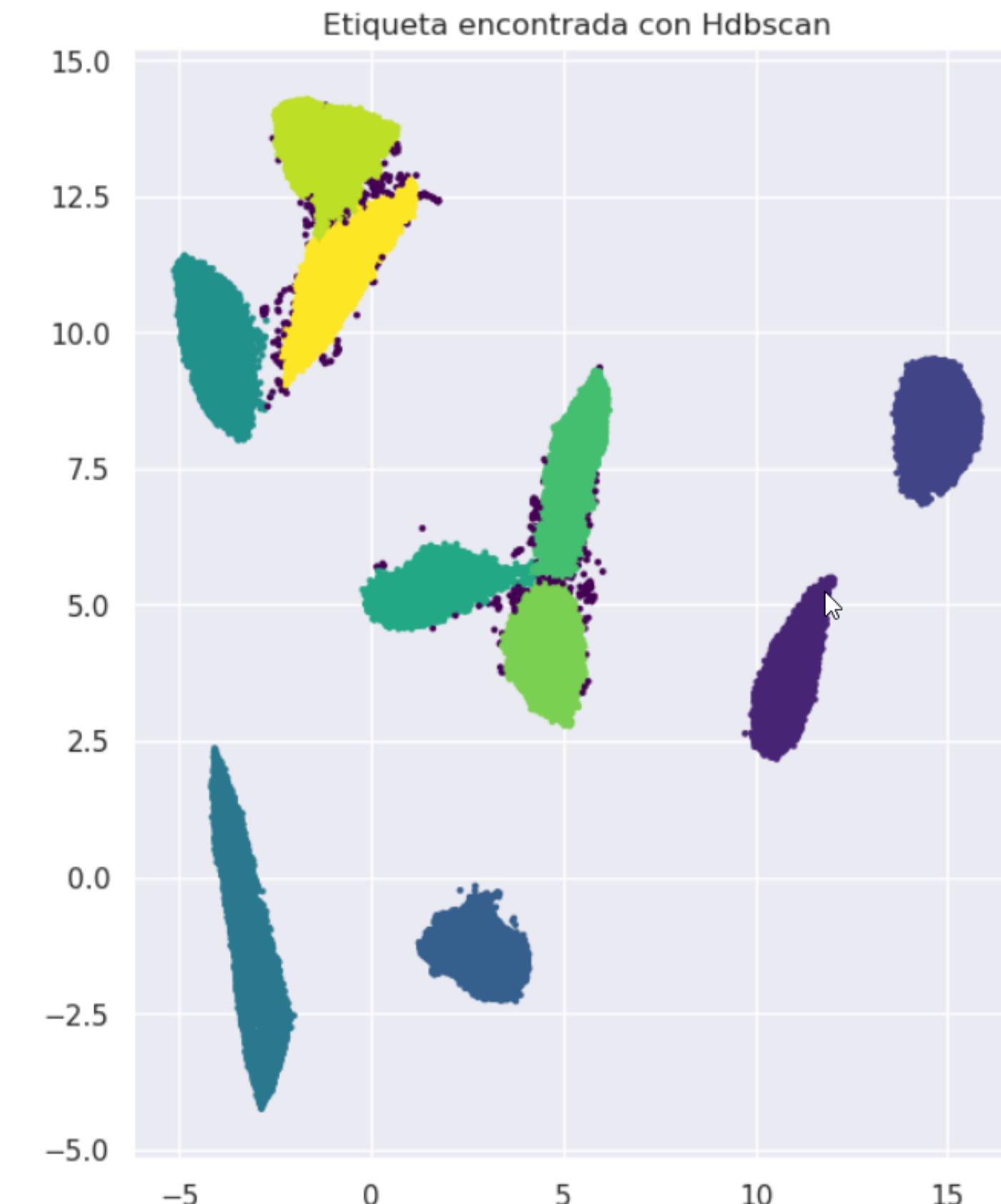
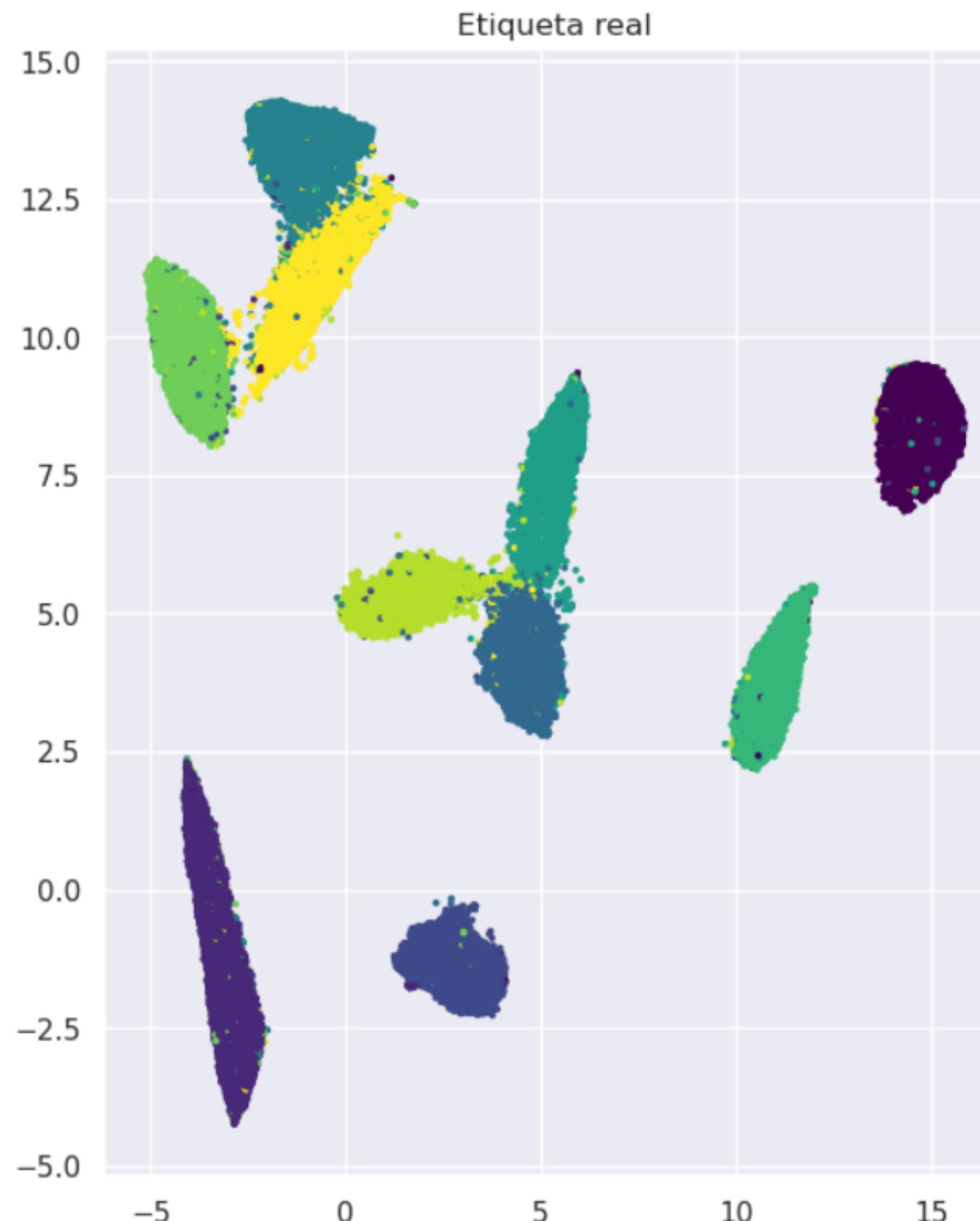
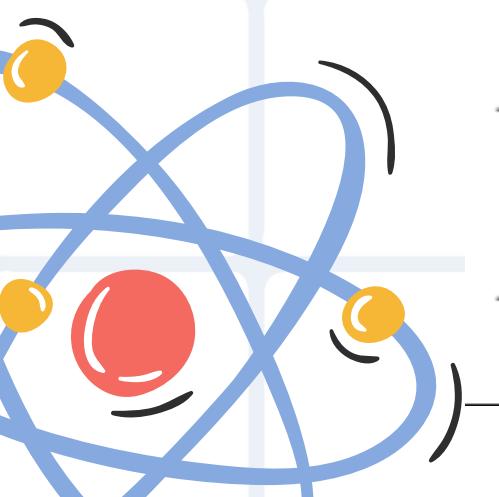


TIENE UN MAL DESEMPEÑO CUANDO LA CANTIDAD DE OBSERVACIONES ES PEQUEÑA O LOS DATOS ESTAN MUY DISPERSOS.

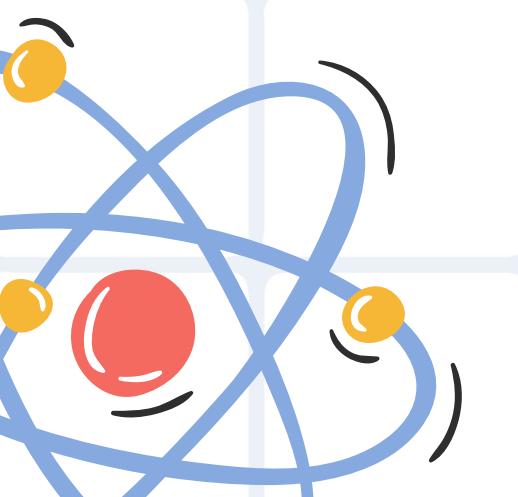
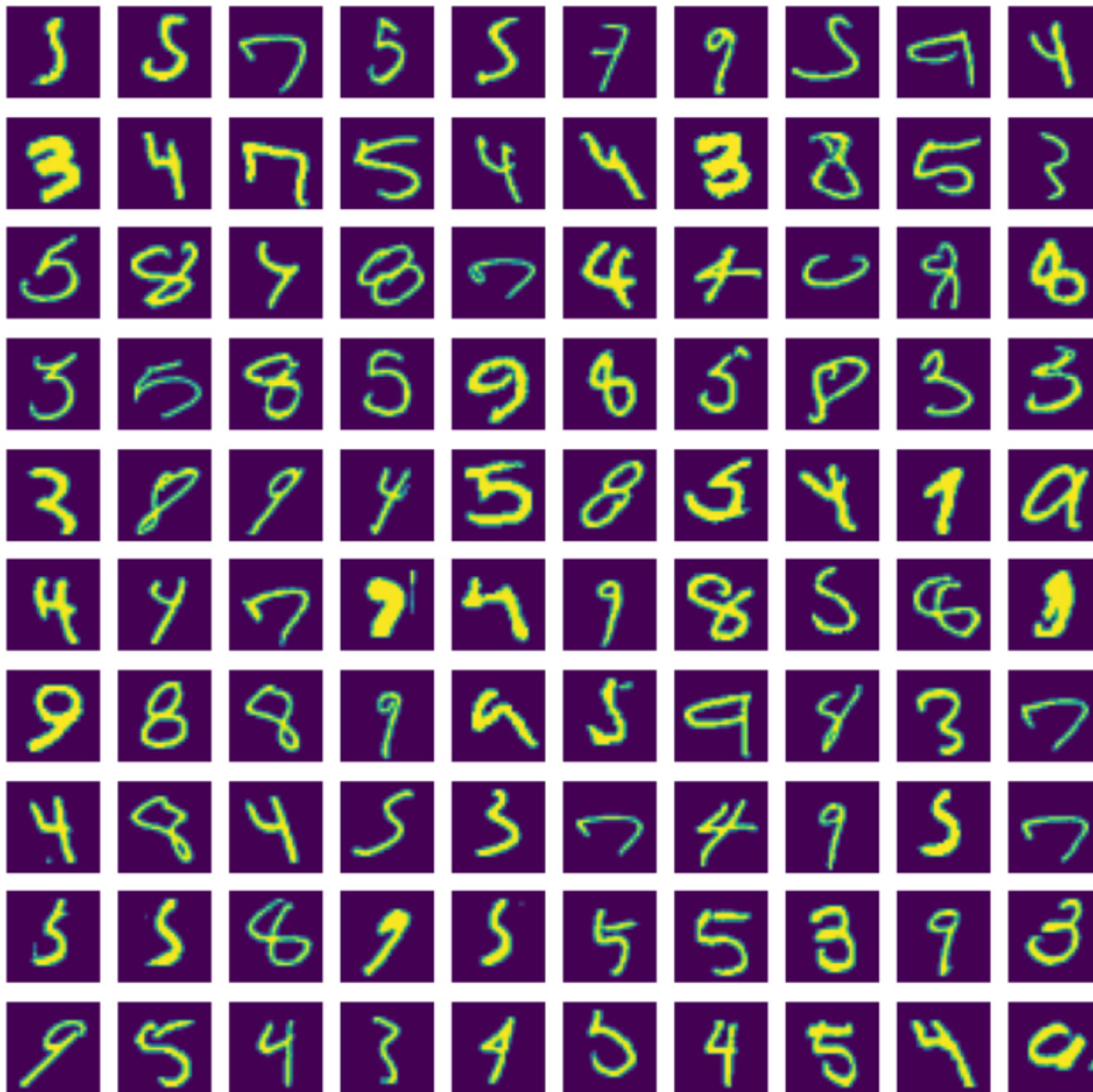


TIENE PROBLEMAS AL TRATAR CON DATOS DE MUCHAS DIMENSIONES.

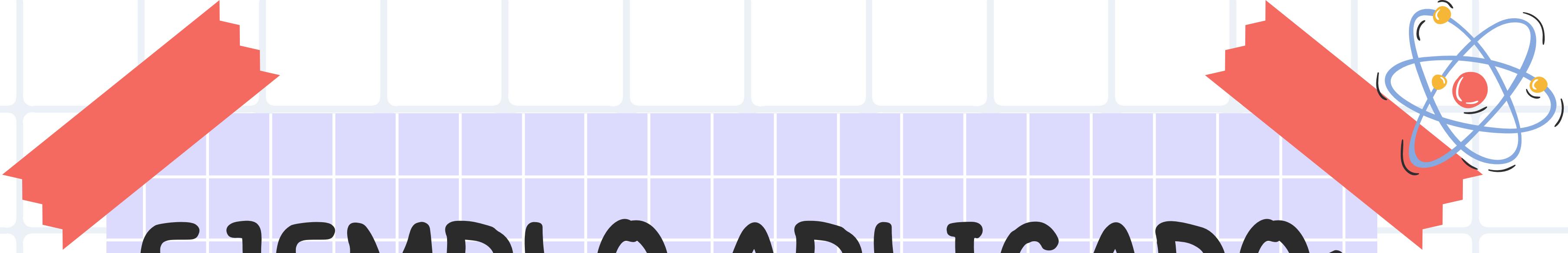
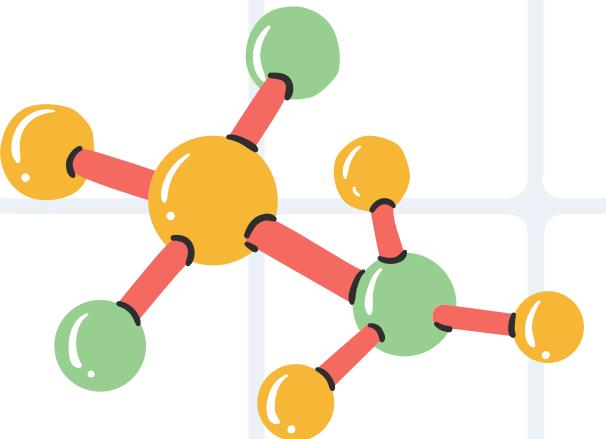
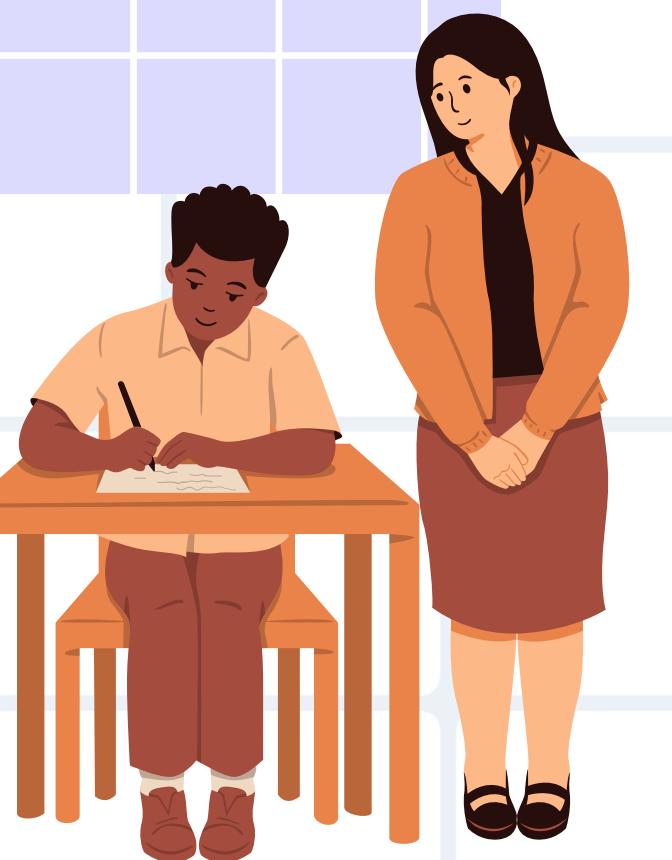
MNIST

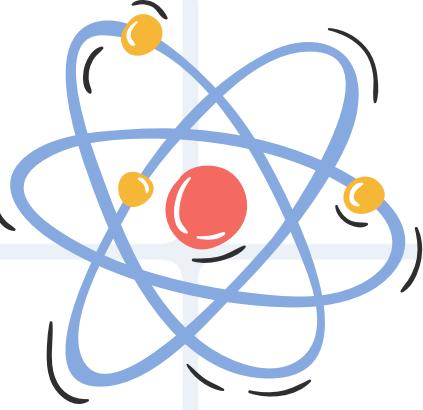


MNIST: DATOS NO CLASIFICADOS



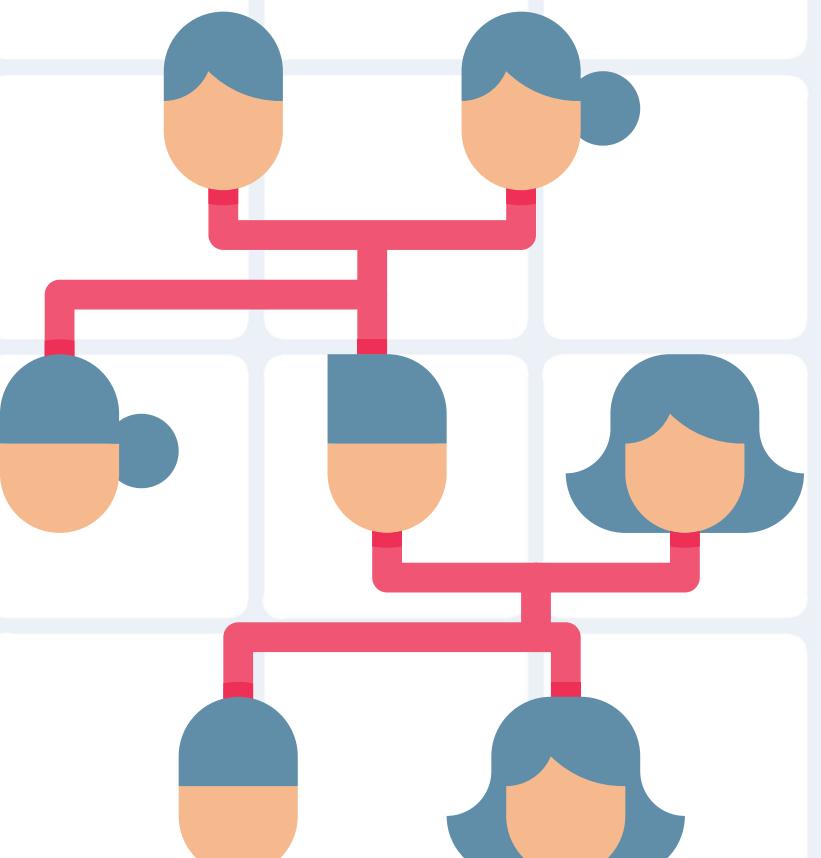
EJEMPLO APLICADO: PRUEBA SABER11



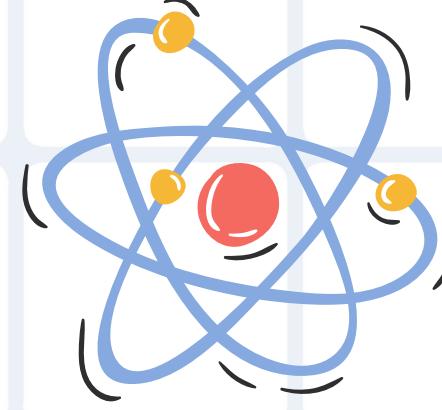


OBJETIVO

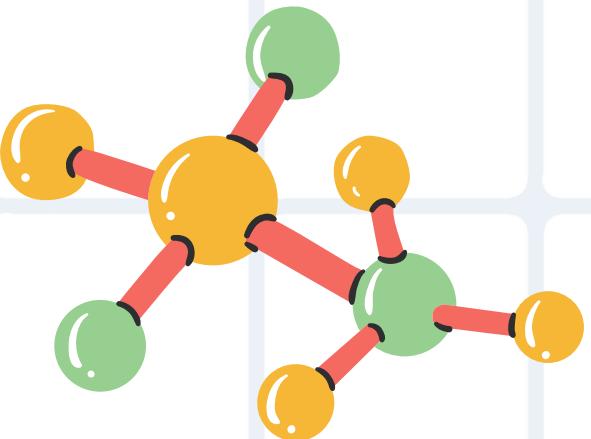
Segmentar y caracterizar a los estudiantes que presentaron la prueba saber 11 en el año 2021 calendario A en Bogotá por medio del algoritmo HDBSCAN



PRUEBA SABER 11



El Examen de Estado de la Educación Media, Saber 11°, es un instrumento de evaluación estandarizada que mide oficialmente la calidad de la educación formal impartida a quienes terminan el nivel de educación media.

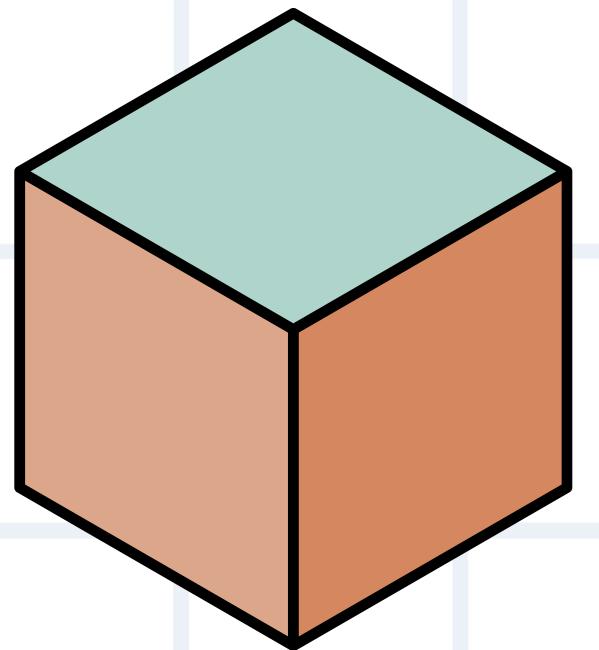


HDBSCAN

RESOLUCIÓN A UN EJEMPLO APLICADO DATOS ICFES 2011



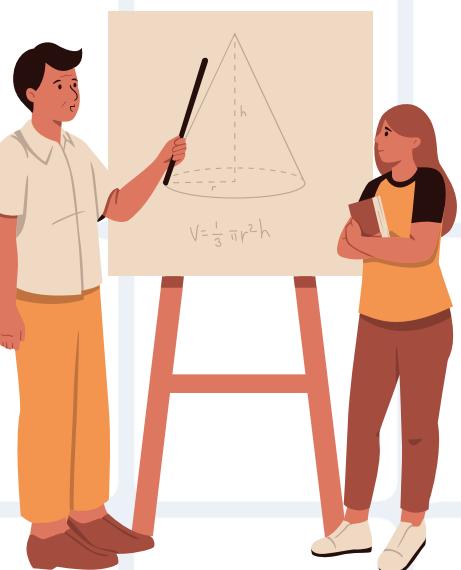
DATOS
ICFES



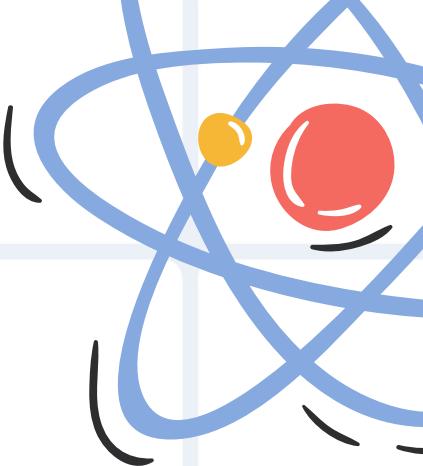
UMAP



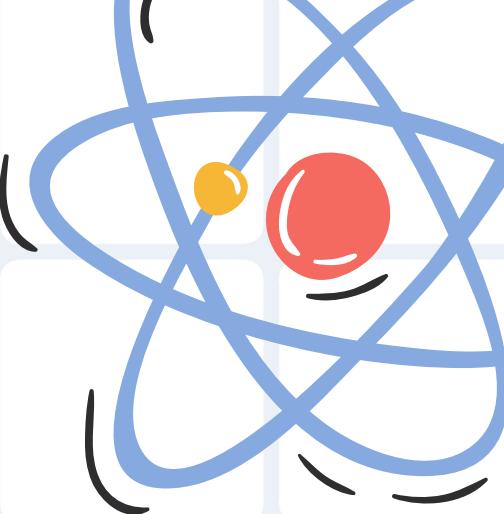
HDBSCAN



DESCRIBIR



ESTRUCTURA DE LOS DATOS



1

29 VARIABLES
CATEGÓRICAS Y 6
NUMÉRICAS

2

CARACTERÍSTICAS DEL
ESTUDIANTE: GÉNERO
ESTRATO, DEDICACIÓN
INTERNET Y LECTURA.

3

CARACTERÍSTICAS HOGAR:
COMPUTADOR, INTERNET,
CONSOLA VIDEOJUEGO,
CUARTOS HOGAR,
CONSUMO DE ALIMENTOS

4

CARACTERÍSTICAS
COLEGIO: CARÁCTER,
JORNADA, BILINGÜE,
CALENDARIO, UBICACIÓN.

5

REGISTROS PARA EL
ANÁLISIS: 69931

PRE-PROCESAMIENTO

RESOLUCIÓN AL PROBLEMA DE ESTA INVESTIGACIÓN

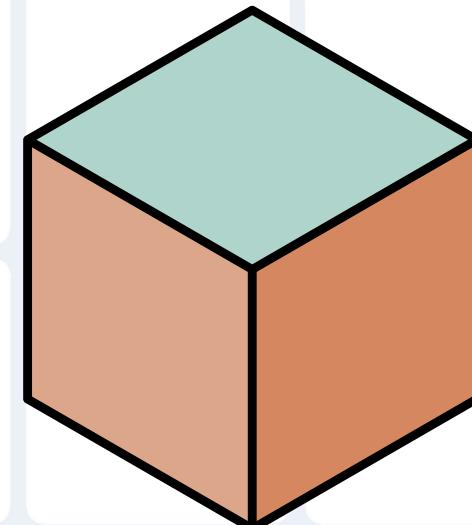
SE CONVIRTIERON LAS VARIABLES CATEGÓRICAS A VARIABLES ONE-HOT, RESULTANDO 130 VARIABLES



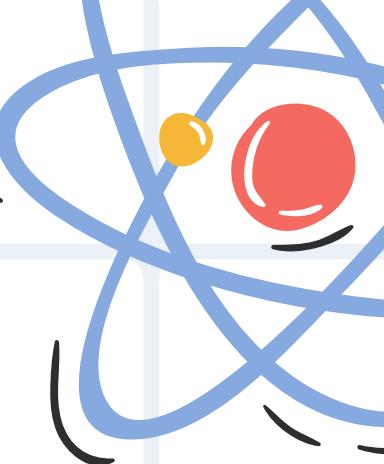
ACP A LAS VARIABLES ONE-HOT. 60 COMPONENTES 90% DE LA VARIANZA



REORGANIZACIÓN DE LAS 60 COMPONENTES CON UMAP



DIMENSIONES FINALES: 60





PRE-PROCESAMIENTO

RESOLUCIÓN AL PROBLEMA DE ESTA INVESTIGACIÓN

SE REALIZÓ LA TÉCNICA UMAP PARA
REDUCIR LAS 100 DIMENSIONES,
CONSERVANDO LA ESTRUCTURA
TOPOLÓGICA LOCAL

- HIPERPARÁMETROS:
- `N_NEIGHBORS = 10`
- `N_DIMS = 60`
- `LEARNING_RATE = 0.5`



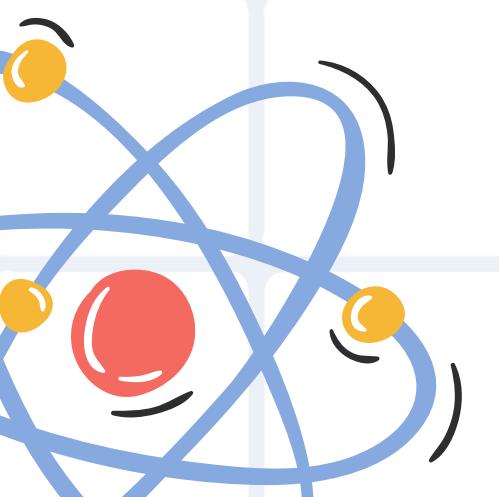
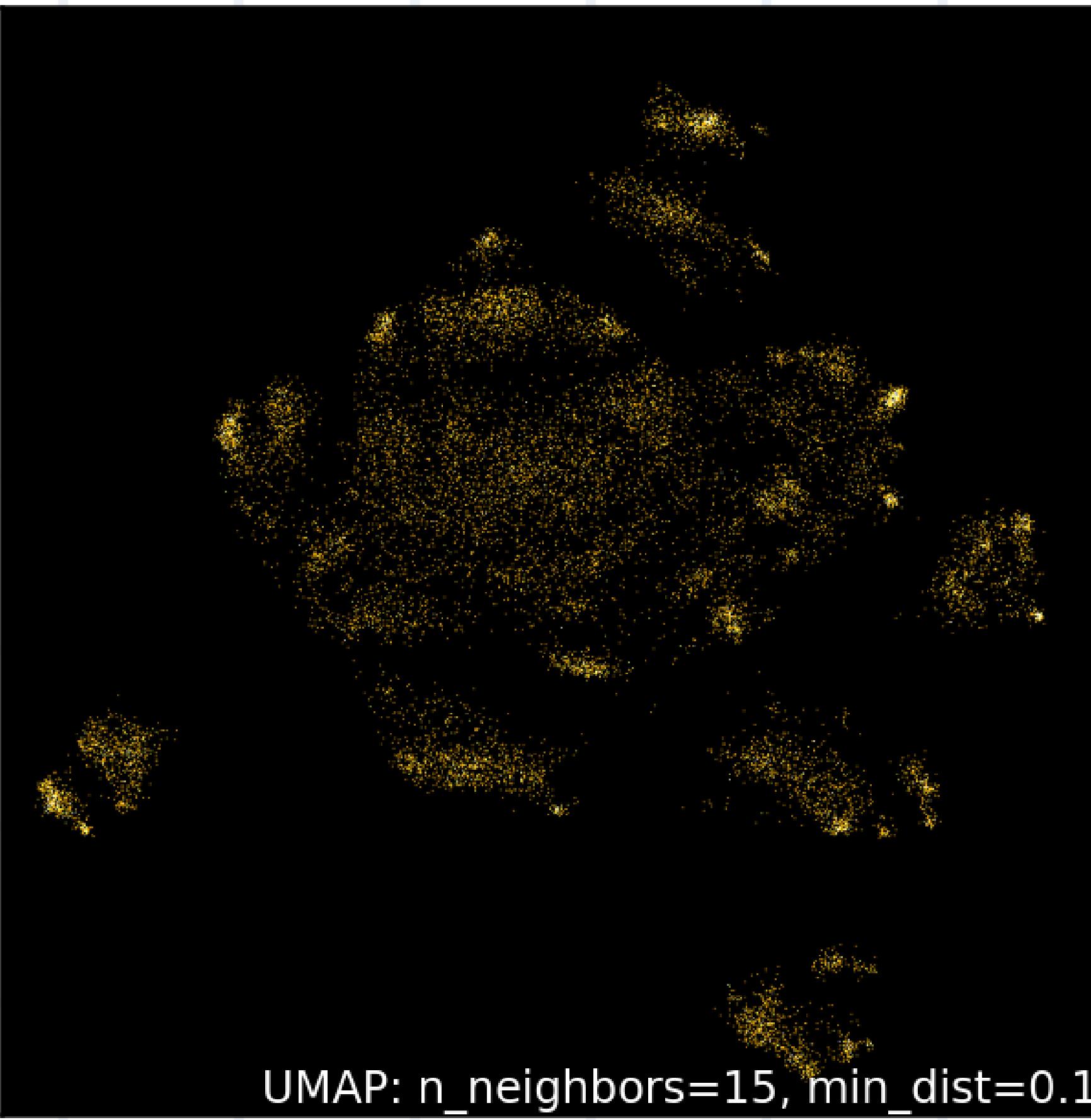
HDBSCAN

RESOLUCIÓN AL PROBLEMA DE ESTA INVESTIGACIÓN

SE IMPLEMENTÓ EL ALGORITMO
HDBSCAN CON LAS 60 DIMENSIONES
ENCONTRADAS CON UMAP

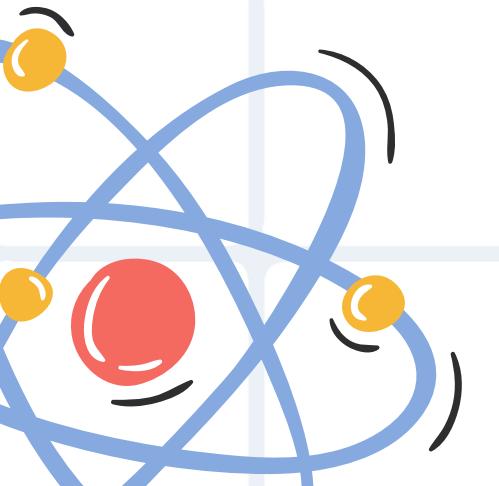
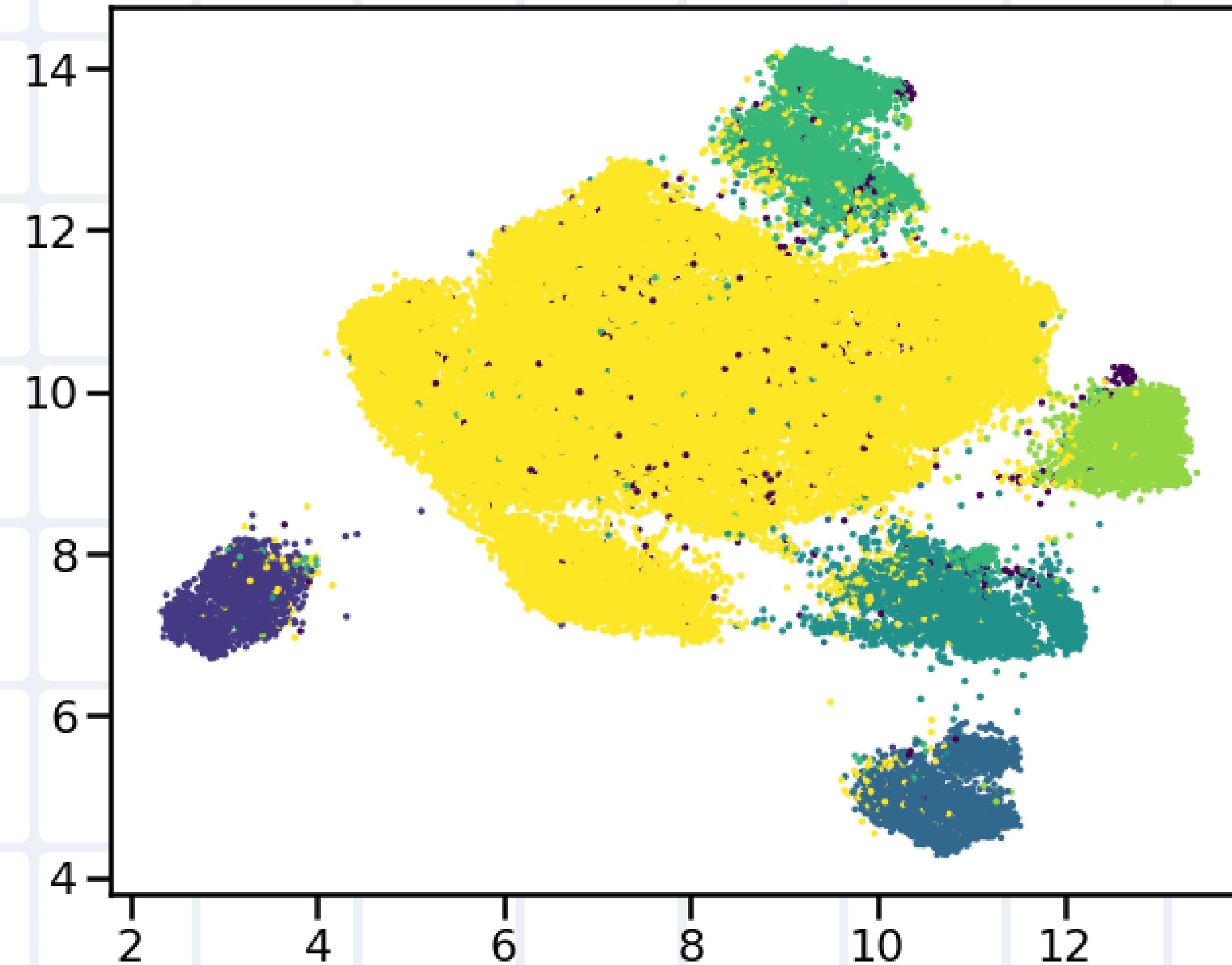
- HIPERPARÁMETROS:
- METRIC = "EUCLIDEAN"
- MIN_CLUSTER_SIZE = 300

UMAP



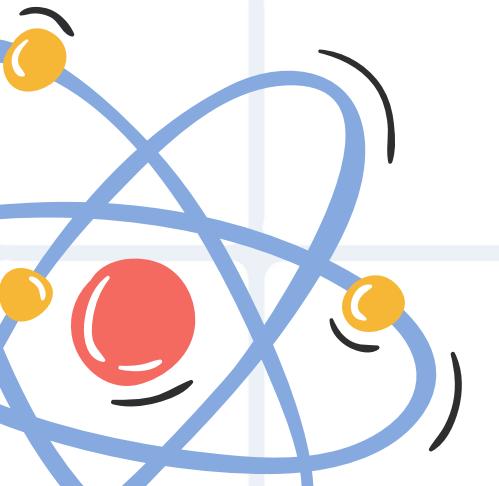
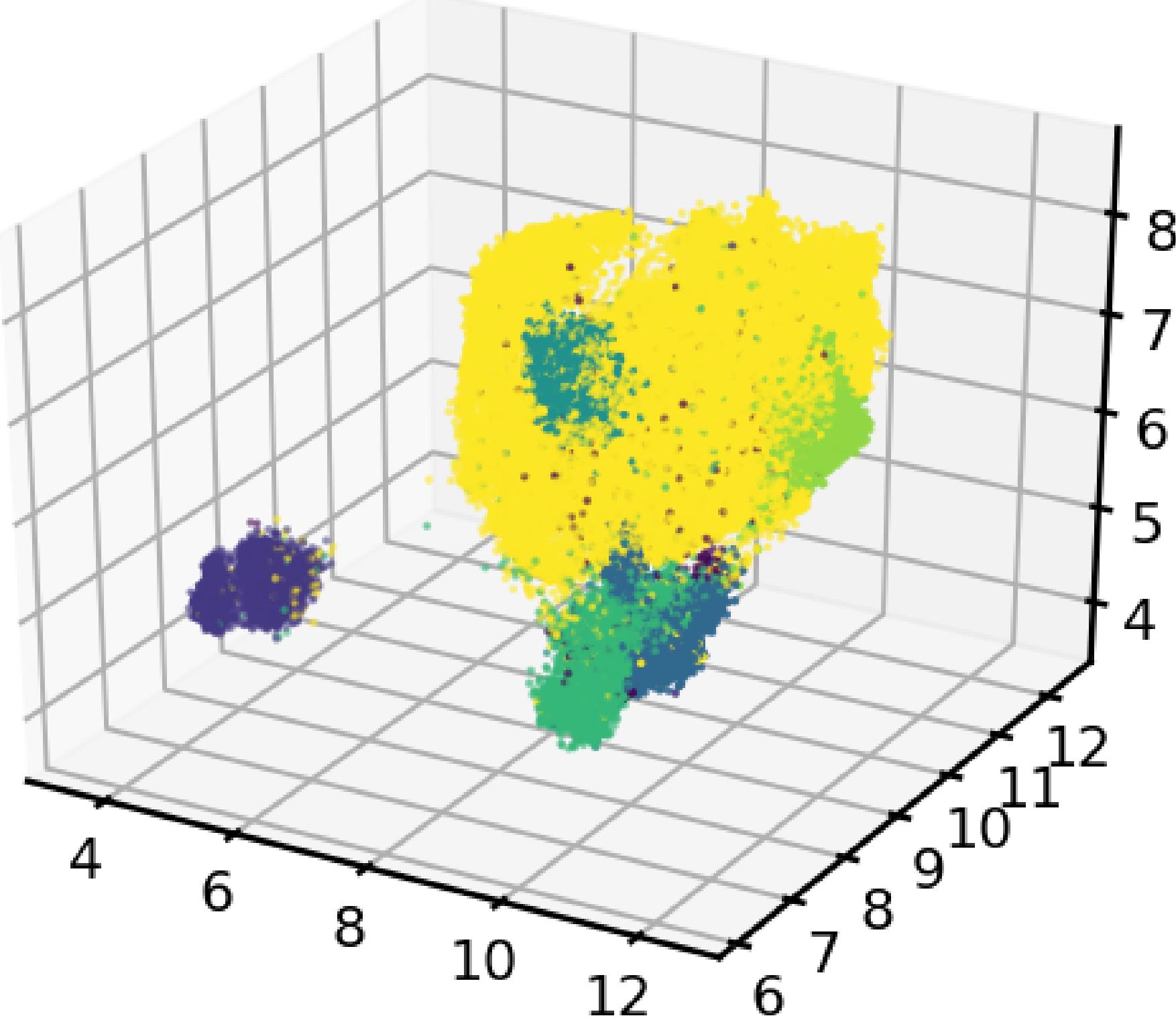
UMAP

Visualización de los datos en 2d con UMAP



UMAP

Visualización de los datos en 3d con UMAP



DESCRIPCIÓN DE LAS CARACTERÍSTICAS POR GRUPOS ENCONTRADAS CON HDBSCAN

LOS GRUPOS SON CONFORMADOS PARA
LOS ESTUDIANTES DE BOGOTÁ

AYUDA DE FACTOMINER PARA EXPLORACION DE
DATOS MULTIVARIADOS



GRUPO 1

3706



- EL 100% NO TIENE INTERNET
- 80.2% ESTRATO BAJOS (1 Y 2)
- EL 56% DE ELLOS PERTENECE A GENERACIÓN E GRATUIDAD
- NO TIENEN BIENES COMO TV, MOTO, TV, AUTOMOVIL
- 88% COLEGIOS OFICIALES
- EL 60.9% SON MUJERES
- BAJO CONSUMO DE CARNES, LECHE Y LEGUMBRES

PUNTAJE GLOBAL 234.81



3714

GRUPO 2

NINGÚN ESTUDIANTE PERTENECE A UN
COLEGIO MIXTO.

71.5% SON MUJERES

64.5% SU FAMILIA TIENE AUTOMOVIL

64.5% PERSONAS DE ESTRATO 3-4

68.4% SON COLEGIOS FEMENINOS

TENENCIA DE BIENES COMO AUTOMOVIL,
LAVADORA, HORNO, INTERNET, TV
COLEGIOS DE JORNADA COMPLETA



PUNTAJE GLOBAL 301

4735

GRUPO 3



- 99.9% TIENEN MÁS DE 100 LIBROS EN SUS CASAS
- TENENCIA DE BIENES
- COLEGIOS NO OFICIALES 70.4%
- 60.5% TIENEN AUTOMÓVIL
- PUNTAJE GLOBAL 297.6
- 71% PERSONAS DE ESTRATOS MEDIOS (3 Y 4)



5144

GRUPO 4

99.11% EL PADRE ES DUEÑO DE UN NEGOCIO PEQUEÑO
99% ESTUDIAN EN COLEGIO MIXTO



PUNTAJE GLOBAL 270.49



GRUPO 5

3490

99.4% SON DE ESTRATO 4

95.6% NO SON DE GENERACIÓN E-GRATUIDAD

87.9% SON DE COLEGIOS NO OFICIALES

79.1% TIENEN AUTOMOVIL

78.9% JORNADA COMPLETA

73.2% COME CARNE PESCADO O HUEVO TODOS LOS DIAS



PUNTAJE GLOBAL 303



48500

GRUPO 6

99.5% INTERNET, TV, COMPUTADOR

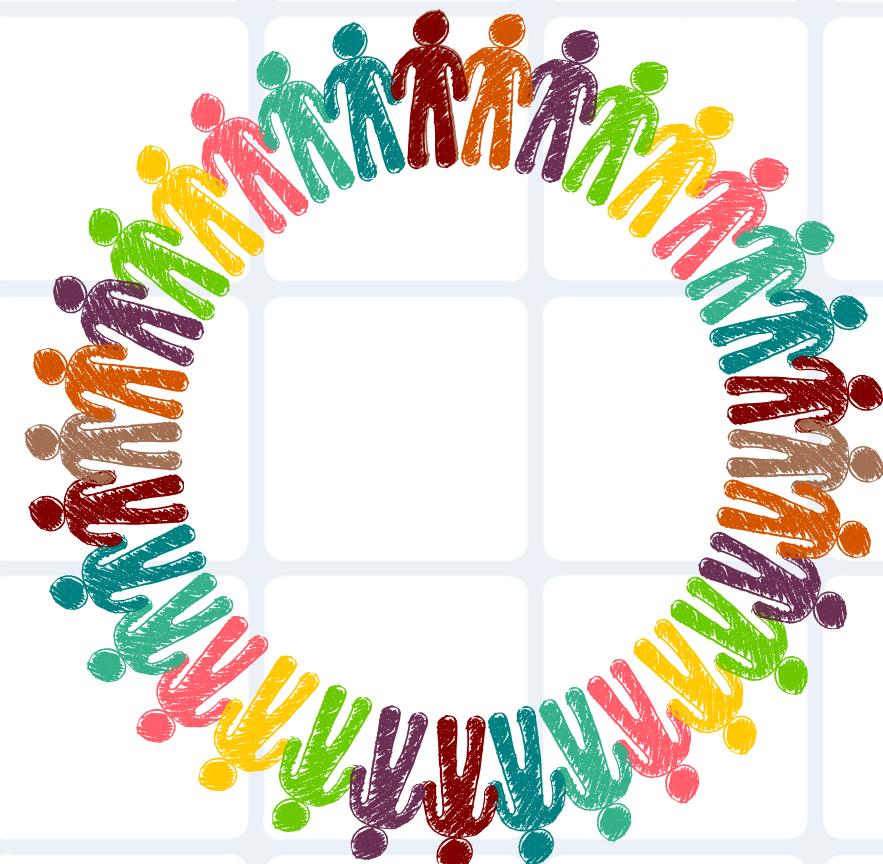
67.8% NO TIENE AUTOMOVIL

65.8% NO TIENE CONSOLA DE VIDEOJUEGOS

62% SON DE COLEGIO DE NATURALEZA OFICIAL

86% SON ESTRATO 2 Y 3

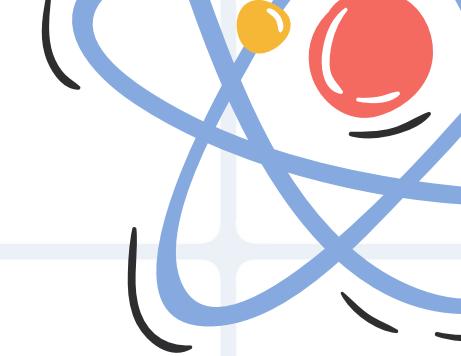
PUNTAJE GLOBAL 264.4



#

642

SIN AGRUPACIÓN



ESTUDIANTE MASCULINO, ESTRATO 1, EL PADRE NO ESTUDIÓ, LA MANDRE TERMINÓ PRIMARIA, EL PAPÁ ES AGRICULTORESTUDIA EN JORNADA TARDE, OBTUVO UN PUNTAJE DE 353

FEMENINO, ESTRATO 2, PADRE TRABAJA POR CUENTA PROPIA, MADRE NO TRABAJA, CONSUMO DE CARNES 1 O 2 VECES POR SEMANA, JORNADA MAÑANA. PUNTAJE 379

MASCULINO, ESTRATO 2, PADRES CON EDUCACIÓN DE TÉCNICOS, EL PADRE ES DUEÑO DE UN NEGOCIO PEQUEÑO, LA MADRE NO TRABAJA, PUNTAJE GLOBAL: 365. GENERACIÓN E- EXCELENCIA NACIONAL

MASCULINO, ESTRATO 6, PADRE NO ESTUDIO, MADRE BACHILLERATO INCOMPLETO, PADRE ES AGRICULTOR, PESQUERO O JORNALEROBAJO CONSUMO BAJO DE CARNES Y DERIVADOS DE LECHE, ALIMENTACIÓN PRINCIPAL DE FRUTAS Y LEGUMBRES, NO NAVEGA INTERNET, PUNTAJE GLOBAL: 234. GENERACIÓN E- GRATUIDAD

BIBLIOGRAFÍA



1. CAMPELLO, R.J.G.B., MOULAVI, D., SANDER, J. (2013). DENSITY-BASED CLUSTERING BASED ON HIERARCHICAL DENSITY ESTIMATES.
2. L. MCINNES AND J. HEALY, "ACCELERATED HIERARCHICAL DENSITY BASED CLUSTERING,"
3. HASTIE, TREVOR, ROBERT TIBSHIRANI, AND JEROME FRIEDMAN. THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION. SPRINGER SCIENCE & BUSINESS MEDIA, 2009.
4. [HTTPS://WWW2.ICFES.GOV.CO/DOCUMENTS/39286/8165657/GU%C3%ADA+DE+ORIENTACI%C3%B3N+SABER+11.%C2%BO+2021-1+PDF+ACCESIBLE.PDF](https://www2.icfes.gov.co/documents/39286/8165657/GU%C3%ADA+DE+ORIENTACI%C3%B3N+SABER+11.%C2%BO+2021-1+PDF+ACCESIBLE.PDF)