Universidad Tecnológica Centroamericana

Facultad de Ingeniería

Sistemas Inteligentes

Mini Proyecto 1

Docente: Dr. Kenny Davila

Presentado por:

11541261 **Carlos Rivera**

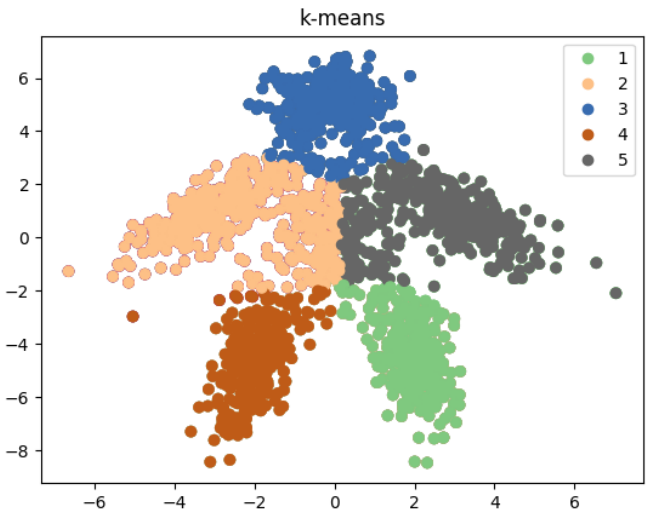
11641381 **Oswaldo Varela**

2 de diciembre del 2021

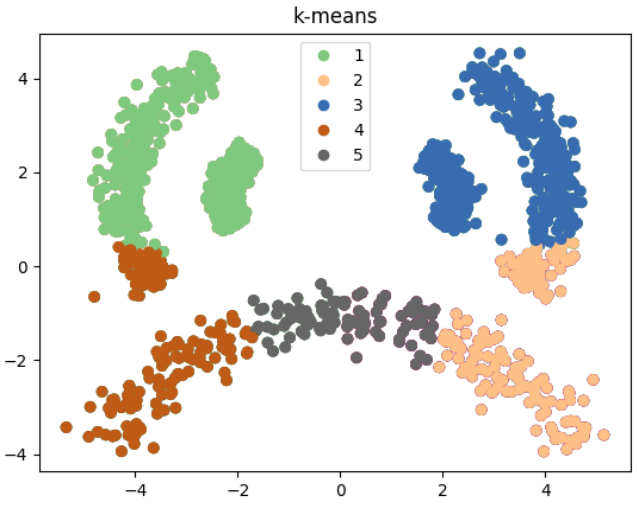
Parte 1: Clustering

Mejor función para el k-mean

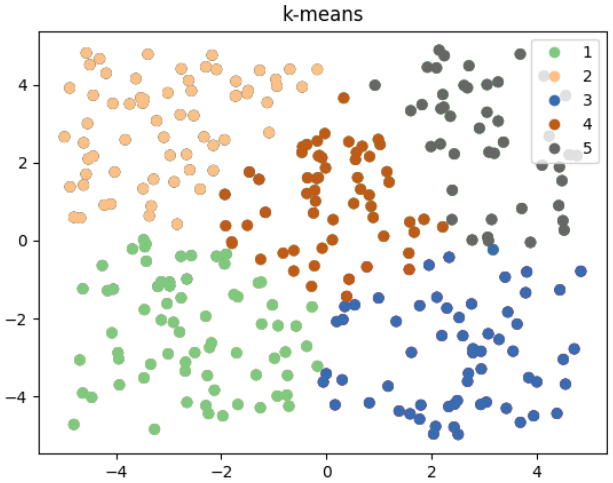
Dataset 1: el de 5 clusters, ya que el algoritmo divide en 5 sectores distintos los datos de una manera adecuada, agrupando los datos por clases



Dataset 2: el de 5 clusters, el algoritmo se divide en 5 sectores distintos de una manera adecuada, agrupando los datos por clase

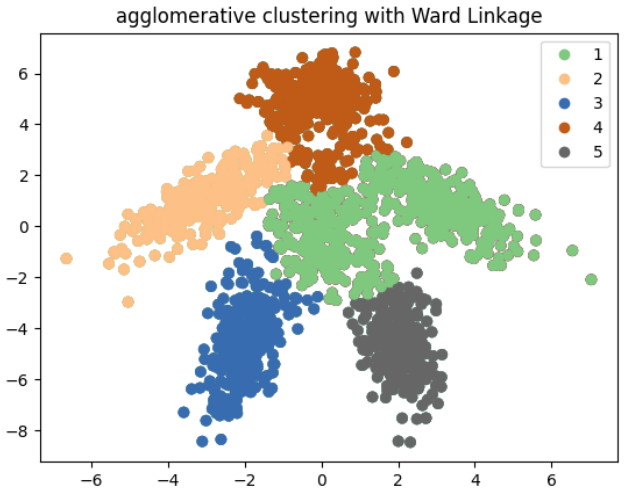


Dataset 3: El de 5 clusters, el algoritmo divide los datos en 5 sectores de una manera adecuada, es decir que los datos están bien agrupados en clases y no están regados

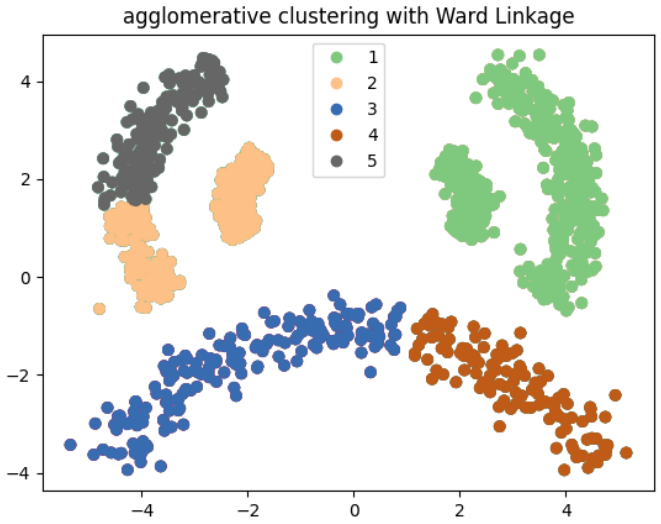


Mejor función para el Clustering Aglomerativo

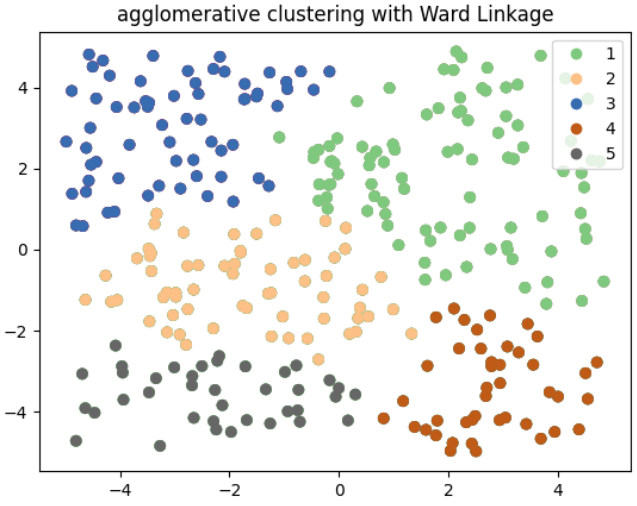
Dataset 1: La mejor función sería la de 5 clusters, ya que estas llega identificar 5 clases y las muestra bien agrupadas evitando la dispersión de datos



Dataset 2: La mejor función sería la de 5 clusters, ya que estas llega identificar 5 clases y las muestra bien agrupadas evitando la dispersión de datos

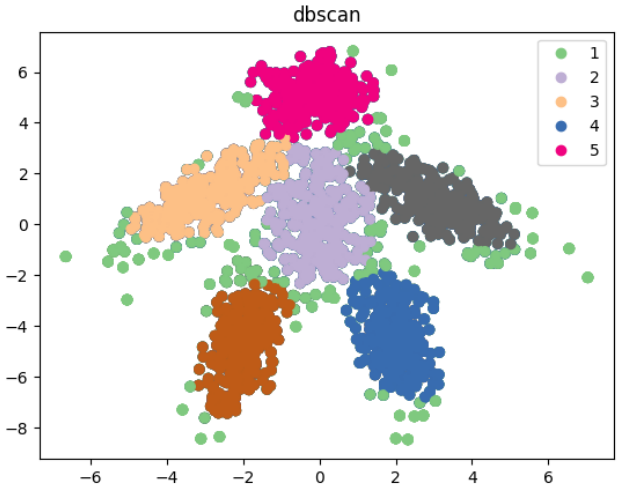


Dataset 3:La mejor función sería la de 5 clusters, ya que estas llega identificar 5 clases y las muestra bien agrupadas evitando la dispersión de datos

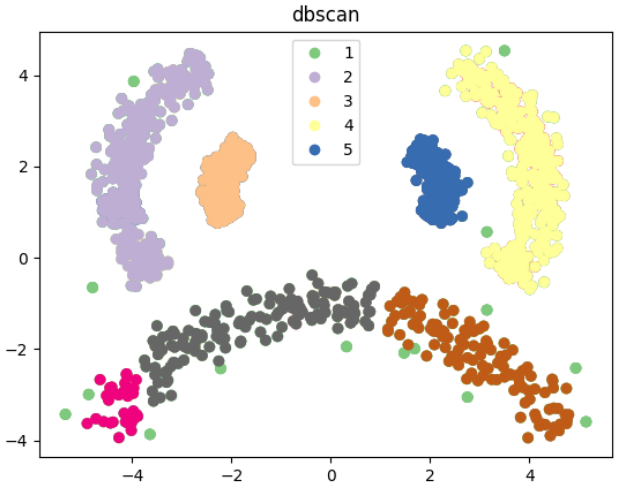


Mejor función para el DBScan

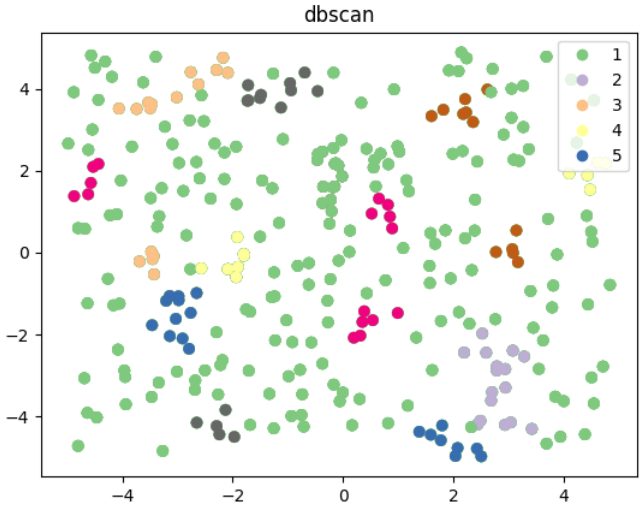
Dataset 1: su mejor función es eps 0.5 min samples 15, este logra identificar 5 clasificaciones distintas, y se muestran agrupadas a diferencia de otros que solo mostraban un poco de las clasificaciones



Dataset 2: su mejor función es eps 0.35 min samples 2, este logra identificar más de 5 clasificaciones distintas, y se muestran agrupadas a diferencia de otros que mostraban las clasificaciones un poco dispersas.



Dataset 3: su mejor función es eps 0.5 min samples 5, este logra identificar más de 5 clasificaciones distintas, y se muestran algo agrupadas y algunos de sus datos se encuentran dispersos.



Análisis de dataset

Dataset 1: Los datos en este dataset se encuentran más agrupados, por lo cual el mejor algoritmo para estos sería el k-mean. Haciendo que los otros dos algoritmos tuvieran más dificultad en dividir sus datos. Posiblemente se utilizaron 4 a 5 clases para generar los datos del dataset.

Dataset 2: Los datos en este dataset se encuentran un poco más dispersos, por lo cual el mejor algoritmo para su clasificación sería el agglomerative.

Dataset 3: Los datos en este dataset se encuentra muy dispersos, haciendo que el algoritmo de dbscan tuviera dificultades en poder agrupar sus datos, pero los algoritmos de k-mean y agglomerative pudieron agrupar bien sus datos

Parte 2: K-NN

k	Accuracy total	Recall	Precisión	F1-score	Time (seg)
1	0.4733	0.4891	0.5191	0.4817	0.0199
3	0.5022	0.5145	0.5883	0.5057	0.0229
5	0.5177	0.521	0.5742	0.5126	0.0209
7	0.5488	0.5486	0.6048	0.5538	0.0239
9	0.5333	0.5348	0.598	0.5382	0.0219
11	0.5688	0.565	0.6064	0.5648	0.0209
13	0.5866	0.5795	0.6146	0.577	0.0239
15	0.5688	0.5616	0.5879	0.5591	0.0229

En un overall de todos, el valor de k que tiene un mejor rendimiento es el k=13, ya que en su accuracy, recall, precision y f1-score tiene los valores más altos, donde este tiene un mejor entrenamiento ya probando el testing, y parte de tiempo está casi similar a los demás no con mucha diferencia.

Comando para la ejecución

python k-means.py “./datasets/datos_1.csv” 1

python agglomerative_clustering_ward.py “./datasets/datos_1.csv” 1

python dbscan.py “./datasets/datos_1.csv” 1

python knn.py "./datasets/genero_peliculas_training.csv"
“./datasets/genero_peliculas_testing.csv"