**MIDA:**

# Activitat 1

EPSEVG
curs: 2023-24

Mariona Farré Tapias

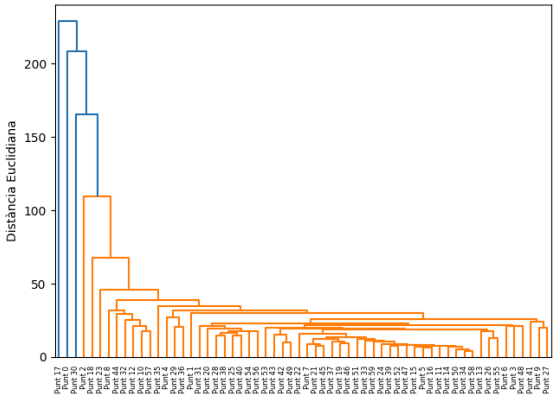
Índex:

1. **Les dades:**

Per aquest treball, s'analitzarà sobre un conjunt de dades dels moviments d'accions de 60 empreses al llarg de 2010 i 2015 dividits per cada dia de l'any.
Per reduir les dades originals, la nova estructura de les dades serà estructurada tal que cada mostra X conté els moviments de l'acció en 5 dies consecutius, i la mostra corresponent a Y conté el moviment de l'acció en el dia següent. Aquestes dades es ara ja es poden utilitzar per entrenar un model de machine learning per a la predicció de sèries temporals.
El codi també es fa una selecció aleatòria del 40% de les dades, ja que en la transformació explicada anteriorment, hi havia encara masses dades a tractar, fent impossible l'execució d'aquestes.

2. **Algorismes jeràrquics:**

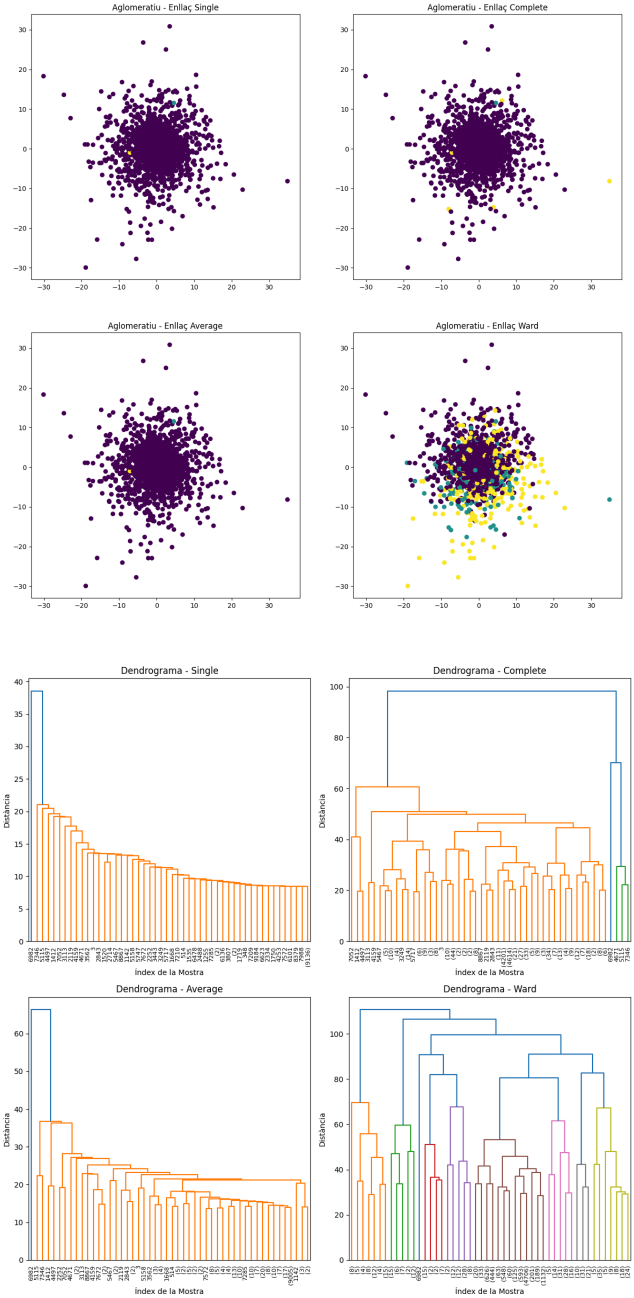
*Apliqueu els diferents algorismes jeràrquics que coneixeu sobre el conjunt de dades i avalueu els resultats.*
El més senzill per veure les dades inicials és amb la distància euclidiana, és una mesura de la distància entre dos punts en un espai d'n dimensions. Es calcula com la longitud de la línia recta que uneix aquests dos punts. En un espai bidimensional, la fórmula per a la distància euclidiana entre dos punts (x1, y1) i (x2, y2) és: $\sqrt{(x2 - x1)2 + (y2 - y1)^2}$




Aquest dendrograma representa totes les distàncies euclidianes calculades en aquest dataset.

Per fer-ho més amè de comparar en el codi posarem els 4 tipus d'algorismes jeràrquics i el seu Dendograma en la mateixa graella:

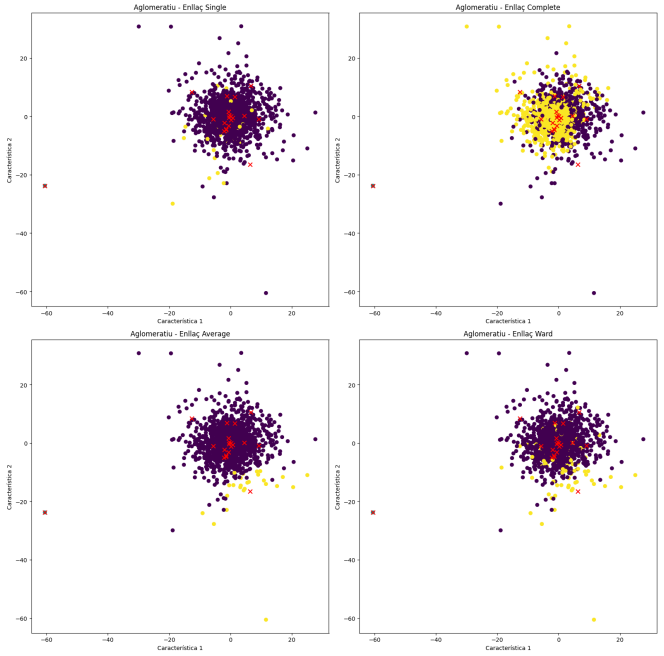
- Single: Aquest algorisme utilitza la distància mínima entre qualsevol parell de punts, un de cada clúster.
- Complete: Aquest algorisme utilitza la distància màxima entre qualsevol parell de punts, un de cada clúster.
- Average: Aquest algorisme utilitza la distància mitjana entre tots els parells de punts, un de cada clúster.
- Ward: Aquest algorisme intenta minimitzar la suma dels quadrats dins dels clústers.

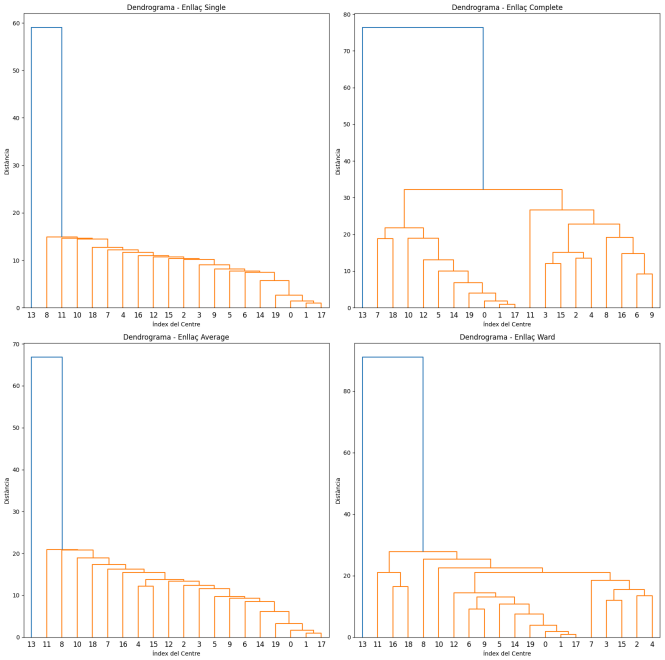



En la primera graella veiem que utilitzen quatre tipus diferents d'algorismes, com podem veure els resultats l'únic algorisme que realment fa alguna mena de classificació és el ward donant gran quantitat de punts als 3 tipus de clústers, mentre que els altres tenen un clúster principal amb alguns punts solts dins del graf.
En la segona graella veiem que utilitzen quatre tipus diferents d'algorismes per trobar el dendograma d'aquests, el dendrograma single té molts grups petits i uns quants de més grans, aquest mètode tendeix a produir clústers llargs i allargats. El dendrograma Complete, mostra grups més equilibrats en comparació amb el mètode Single.  Aquest mètode tendeix a produir clústers més compactes i equilibrats.el dendrograma average proporciona un punt intermedi amb grups de mida moderada i equilibrats. i el dendrograma Ward tendeix a crear grups de mida igual i esfèrics, clústers que són més compactes i equilibrats que els altres mètodes.
Els 3 primers donen a suposar una divisió d'una sola agrupació o màxim dos (complete), però amb l'algorisme ward és el que ens classifica més específicament les dades tenint uns 8 agrupacions delimitades.

### 3. Algorismes jeràrquics sobre els centres:

*Apliqueu els algorismes jeràrquics sobre els centres de K-means amb k elevada (p.e. k=20) i avalueu-los:*



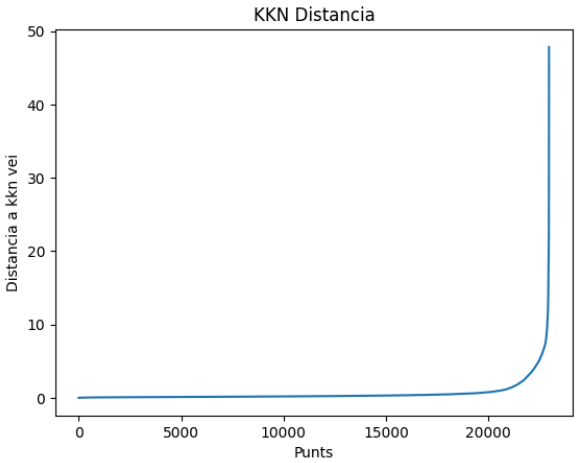




Les graelles anteriors també segueixen la divisió per diferents tipus del càlcul de les distàncies : single , complete, average i ward.
Comparat amb els anteriors algorismes jeràrquics (no s'executen sobre els centres), aquests aconsegueixen una mica més de classificació dels punts, sobretot el Complete seguit pel Ward, els altres dos continuen tenint un gran clúster amb alguns punts solts de l'altre clúster. En els dendogrames es veu el mateix que el Complete aconsegueix una divisió més equilibrada dels punts mentre que la resta continuen tenint divisions generalment iguals..

4. **DBScan:**

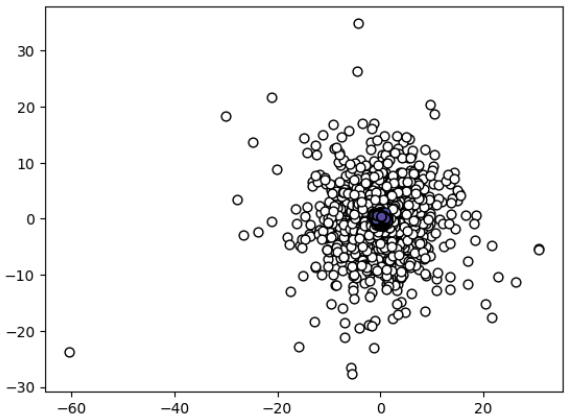
*Mireu quins serien bons parametres per DBScan, apliqueu al conjunt de dades i avalueu resultats*

Per trobar els millors paràmetres de DBScan, trobar el valor de eps trobant el kkn, *kneareset neighbors* trobar els k-n veïns més a prop:




On el millor eps és 0.23 i la distància kcloser es va actualitzant a cada iteració.

I el DBScan d'aquestes dades dona:




Aquest graf no segueix les formes que hauria de tenir un DBScan de mitja lluna, aixo es produeix perquè le dades són molt disperses i amb molts punts aïllats, fent que aquest algorisme no sigui el més adequat per aquest tipus de dades de la borsa.

5. **Resultats:**

*Compareu els resultats de tots els mètodes i digueu quin mètode dona millors resultats i si creieu que hi ha algun motiu per això*

Veient tots els algorismes anteriors, és prou clar quin ha sigut el millor algorisme en poder classificar les dades de les transaccions de borsa, al ser unes dades tan disperses i sense seguir un patró en específic crear clústers amb aquestes és una tasca complicada.

Però l'algorisme que ha pogut dividir les dades amb millors resultats és: Algorismes de Ward original , sense centrar-se en els centres dels clústers.

Això és degut  a que té en compte la variança de les dades, minimitza la dispersió dins dels clústers creats, té més flexibilitat i s'adapta millor a formes més complexes que els altres algorismes.
Els algorismes single(min)  és sensible  a la unió de les dades fent que es puguin fusionar clústers que estiguin massa a prop tenint clústers allargats. El algorismes complete (max) és separació de les dades, dividir clústers fins que es un punt es trobi massa lluny tenint petits clústers i dispersos.
Aquests no han pogut donar bons resultats perquè la majoria dels punts els posaven en un sol clúster sense fer-hi cap tipus de classificació.
I finalment el DBScan és sensible a les densitats, si les dades tenen diferents densitats o formes irregulars, com aquestes dades, no dóna bons resultats marcant casi tots els punts com outliers.