

# **Detecció de patrons en l'Autisme mitjançant ML no supervisat**

**Diana Hueso Beneyto**

1



36

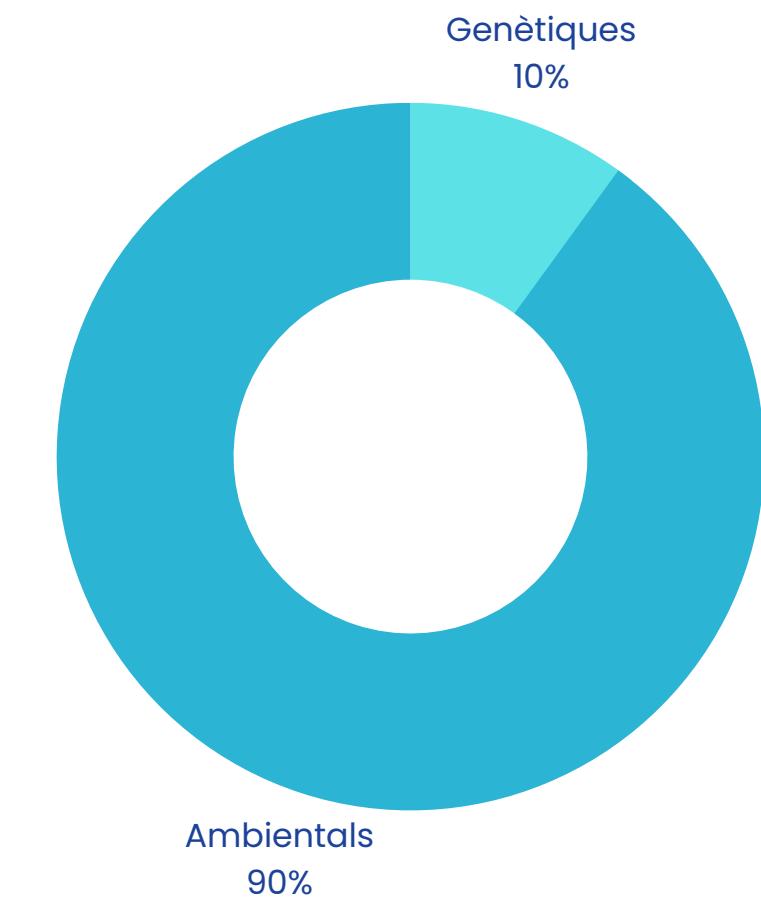


**El número de  
casos a Espanya i  
al món s'estima  
que és d'1-88.**

# Introducció

L'autisme és un trastorn del desenvolupament neurològic que afecta la comunicació, les habilitats socials i el comportament de les persones.

Hi ha hagut un gran augment de la prevalença de l'autisme en els últims anys a tot el món i va encara més en augment, hi ha qui ho defineix com "epidèmia silenciosa". Les causes poden ser o bé genètiques en un 10-15% dels casos i la resta són deguts a factors ambientals.



La meva idea de projecte era utilitzar algun model de ML no supervisat per ajudar a identificar patrons i característiques comunes entre les persones amb autisme.

# **DATASET**

**1**



# NSCH, cens nacional de salut infantil dels Estats Units

Està compost per 40 columnes i 119.241 files, i totes les dades són de l'any 2021. Finalment vaig utilitzar 40.000 registres i 20 columnes.

```
8 data screener;
9     set file.nsch_2021_screener;
10    label
11 C_FWS="Child Weight"
12 C_CSHCN="Special Health Care Needs Status of Child"
13 C_K2Q23="Child Treatment for Chronic Emotion Develop Behave"
14 C_K2Q22="Child Needs Treatment for Emotion Develop Behave"
15 C_K2Q21="Child Special Therapy for Health Condition for 12 Months"
16 C_K2Q20="Child Special Therapy for Health Condition"
17 C_K2Q19="Child Special Therapy"
18 C_K2Q18="Child Limited Ability from Health Condition for 12 Months"
19 C_K2Q17="Child Limited Ability from Health Condition"
20 C_K2Q16="Child Limited Ability"
21 C_K2Q15="Child Medical Care Currently for 12 Months"
22 C_K2Q14="Child Medical Care Used or Needed for Health Condition"
23 C_K2Q13="Child Needs or Uses More Medical Care than Others"
24 C_K2Q12="Child Medication Currently for 12 Months"
25 C_K2Q11="Child Medication Used or Needed for Health Condition"
26 C_K2Q10="Child Needs or Uses Medication Currently"
27 C_RACE_R_IF="Imputation Flag for C_RACE_R"
28 C_HISPANIC_R_IF="Imputation Flag for C_HISPANIC_R"
29 C_SEX_IF="Imputation Flag for C_SEX"
30 C_ENGLISH="Child Speak English"
31 C_SEX="Child Sex"
32 C_HISPANIC_R="Hispanic Origin of Child, Recode"
33 RACEAIAN="Race of Child, Recode, AIAN included. Reported for AK, AZ, NM, MT, ND, OK, SD."
34 RACEASIA="Race of Child, Recode, Asian included. Reported for CA, HI, MA, MD, MN, NJ, NV, NY, VA, WA."
35 RACER="Race of Child, Recode"
36 C_RACE_R="Race of Child, Detailed"
37 C_AGE_YEARS="Child Age - Years"
38 YEAR="Survey Year"
39 LINENUM="Child Line Number"
40 MPC_YN="Metropolitan Principal City Status"
41 METRO_YN="Metropolitan Statistical Area Status"
42 CBSAfp_YN="Core Based Statistical Area Status"
43 HHLANGUAGE="Primary Household Language"
44 TOTKIDS_R="Number of Children in Household"
45 FIPSST="State FIPS Code"
46 FWH="Household Weight"
47 STRATUM="Sampling Stratum"
48 HHIDS="Unique Household ID (Screener)"
49 TENURE="The Conditions under Which Land or Buildings Are Held or Occupied"
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119241 entries, 0 to 119240
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   FIPSST            119241 non-null   object 
 1   TOTKIDS_R          119241 non-null   float64
 2   TENURE             119241 non-null   float64
 3   HHLANGUAGE         118383 non-null   float64
 4   METRO_YN           107403 non-null   float64
 5   MPC_YN              98954 non-null   float64
 6   TENURE_IF           119241 non-null   float64
 7   FWH                119241 non-null   float64
 8   LINENUM             119241 non-null   float64
 9   C_AGE_YEARS         119241 non-null   float64
 10  C_RACE_R             119241 non-null   float64
 11  C_HISPANIC_R_IF      119241 non-null   float64
 12  RACER               119241 non-null   float64
 13  RACEASIA             20874 non-null   float64
 14  RACEAIAN             15144 non-null   float64
 15  C_HISPANIC_R          119241 non-null   float64
 16  C_SEX                119241 non-null   float64
 17  C_K2Q10              119029 non-null   float64
 18  C_K2Q12              15556 non-null   float64
 19  C_K2Q13              119041 non-null   float64
 20  C_FWS                119241 non-null   float64
dtypes: float64(20), object(1)
memory usage: 19.1+ MB
```



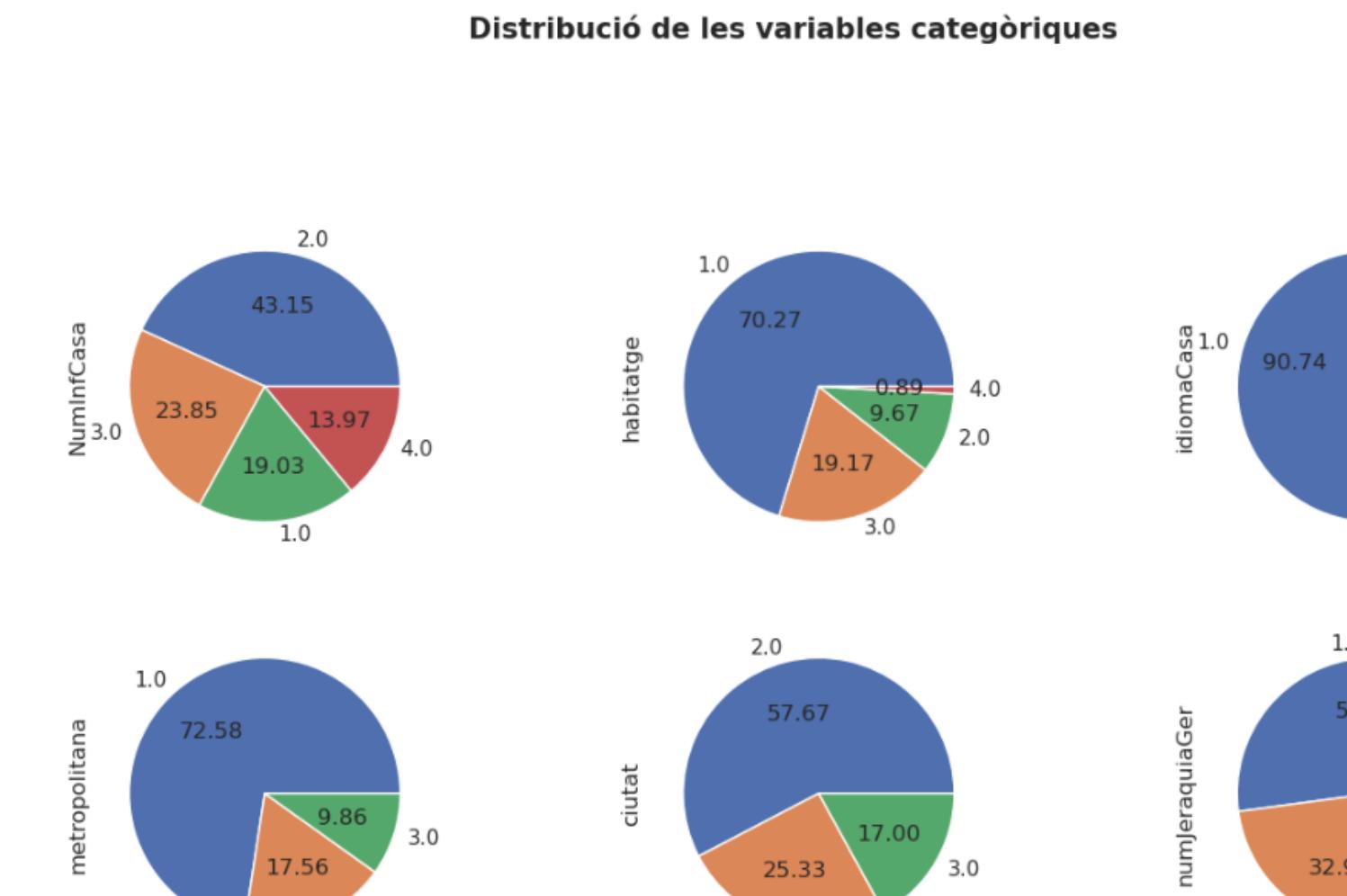
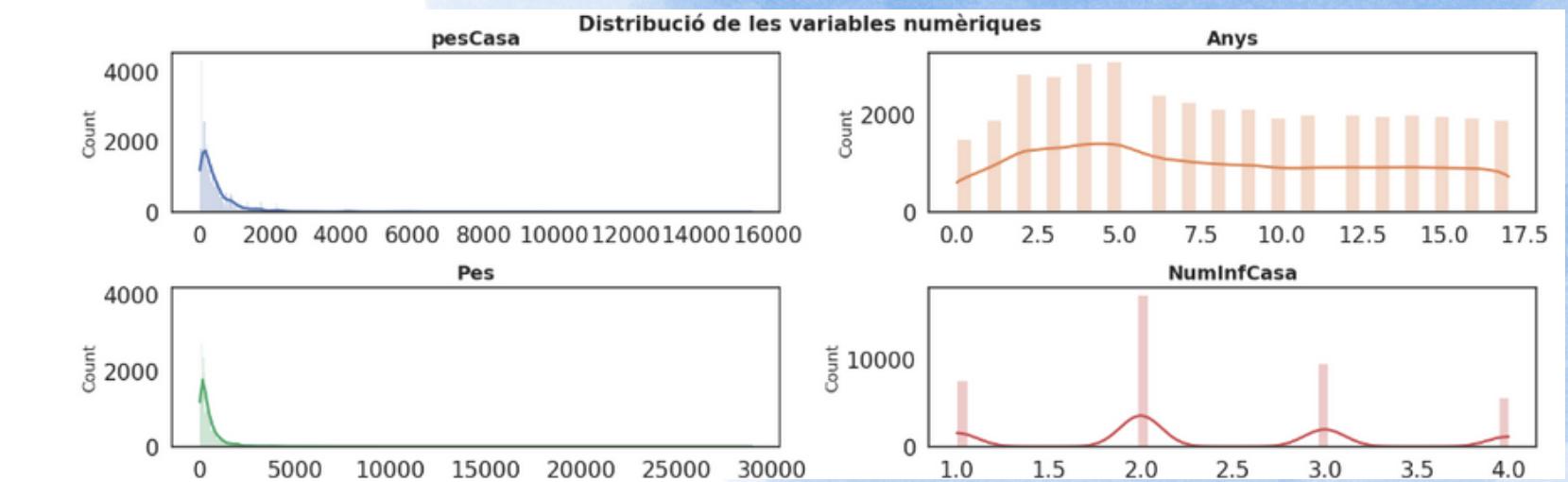
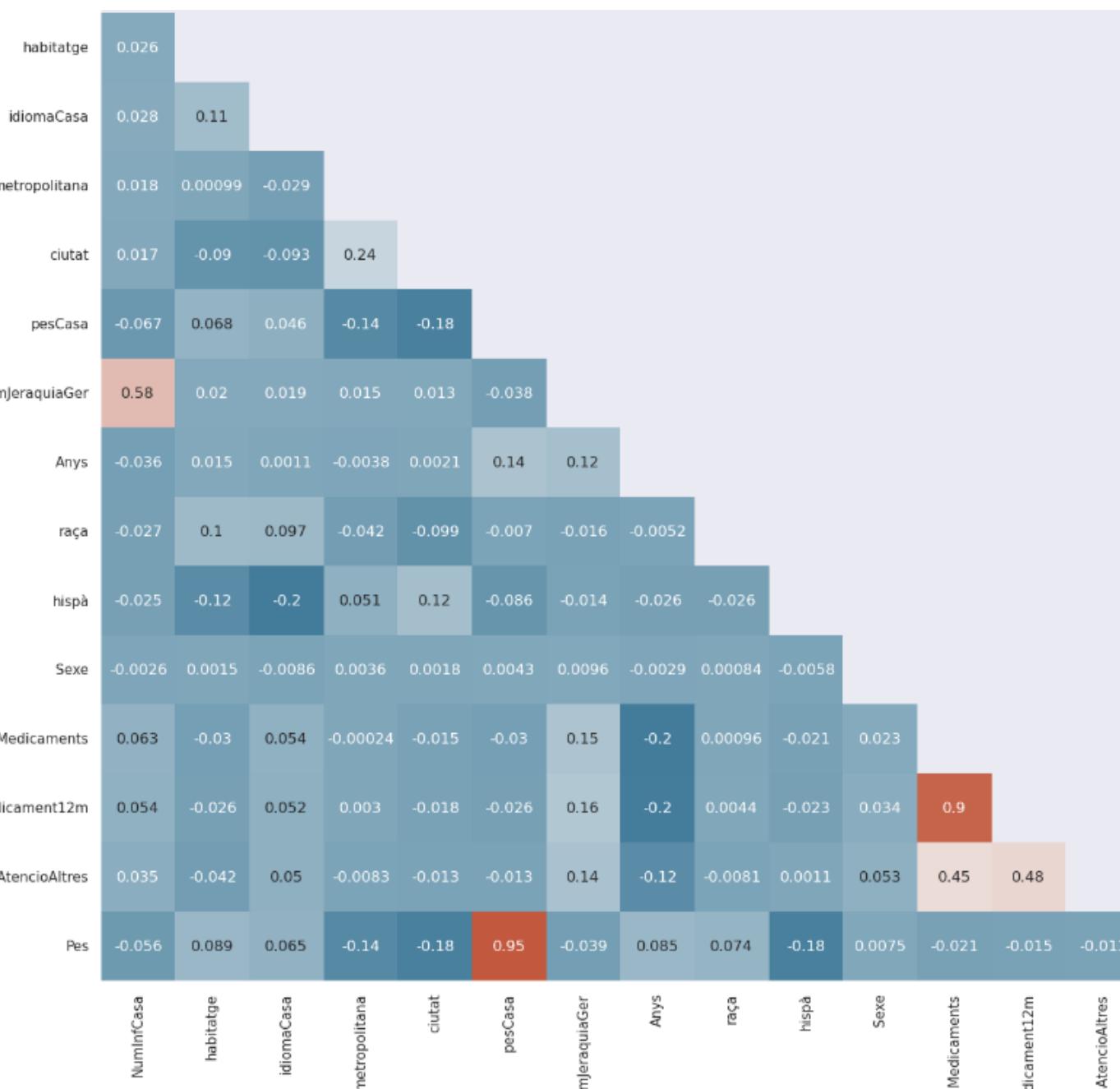


**2**

# **NETEJA I PREPROCESSAT**

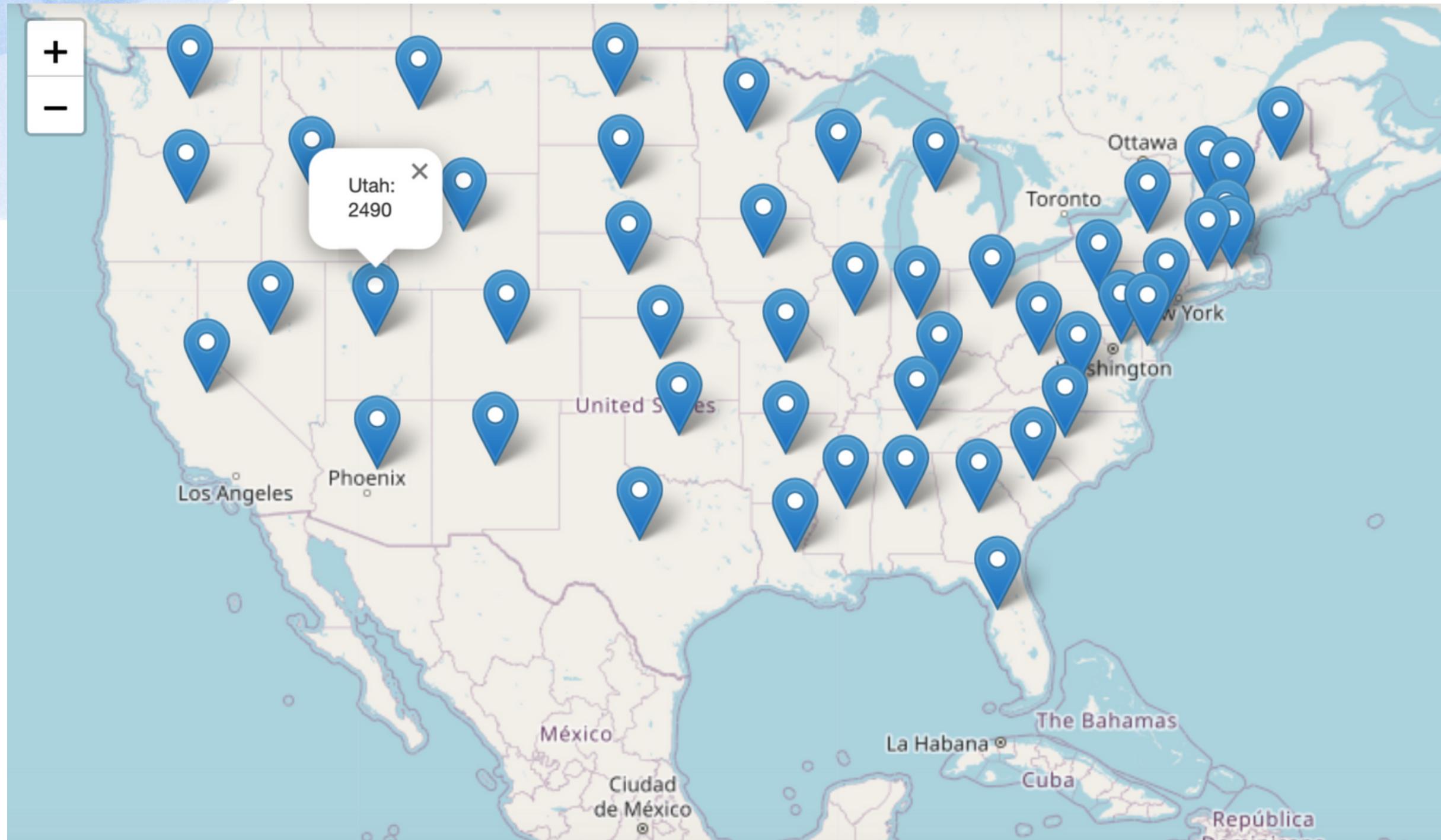


- Esborrar columnes que no volia utilitzar
- Canviar els nulls i NaNs per altres valors
- Canviar els noms de les variables
- Fer matriu de correlació
- Distribució de les variables numèriques i categòriques
- Hipòtesis test
- Transformació de variables amb outliers
- Mapa interactiu



# Mapa interactiu amb n° de casos a cada estat

Fet amb Folium i Geopandas



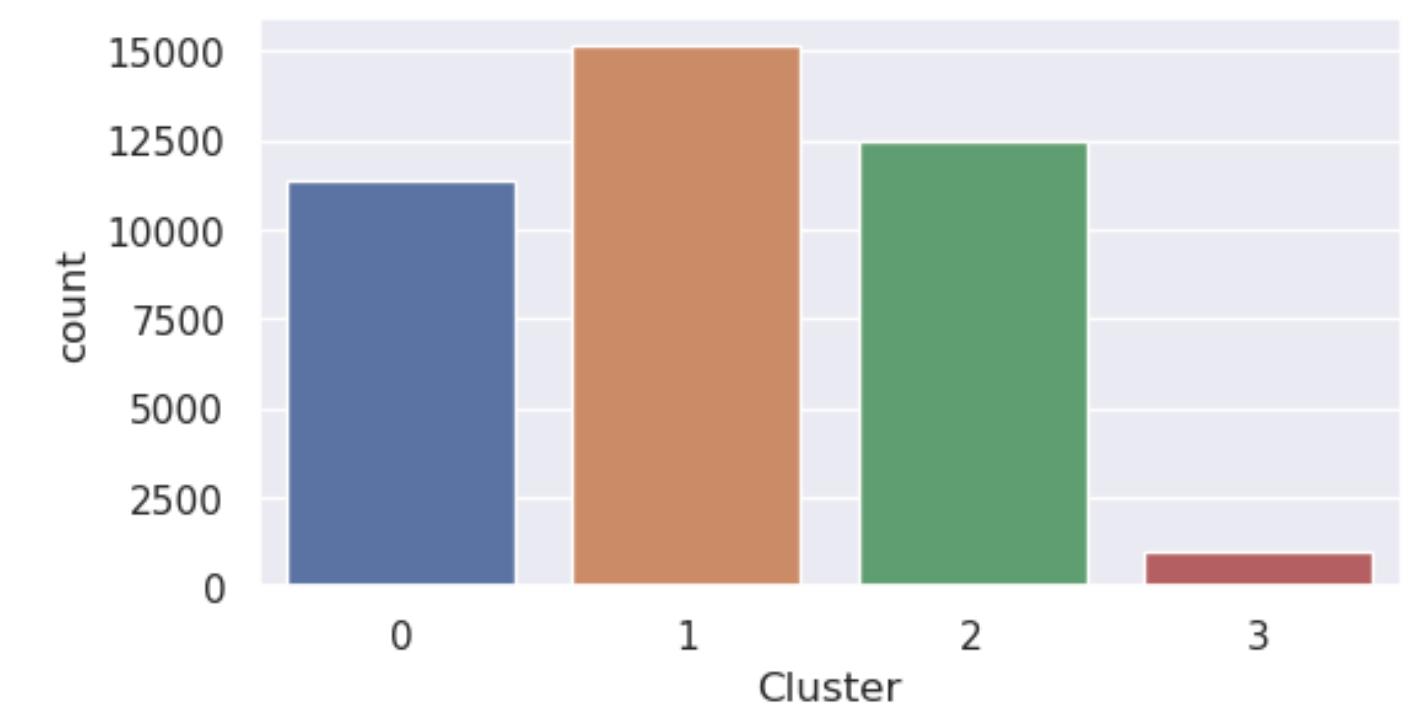
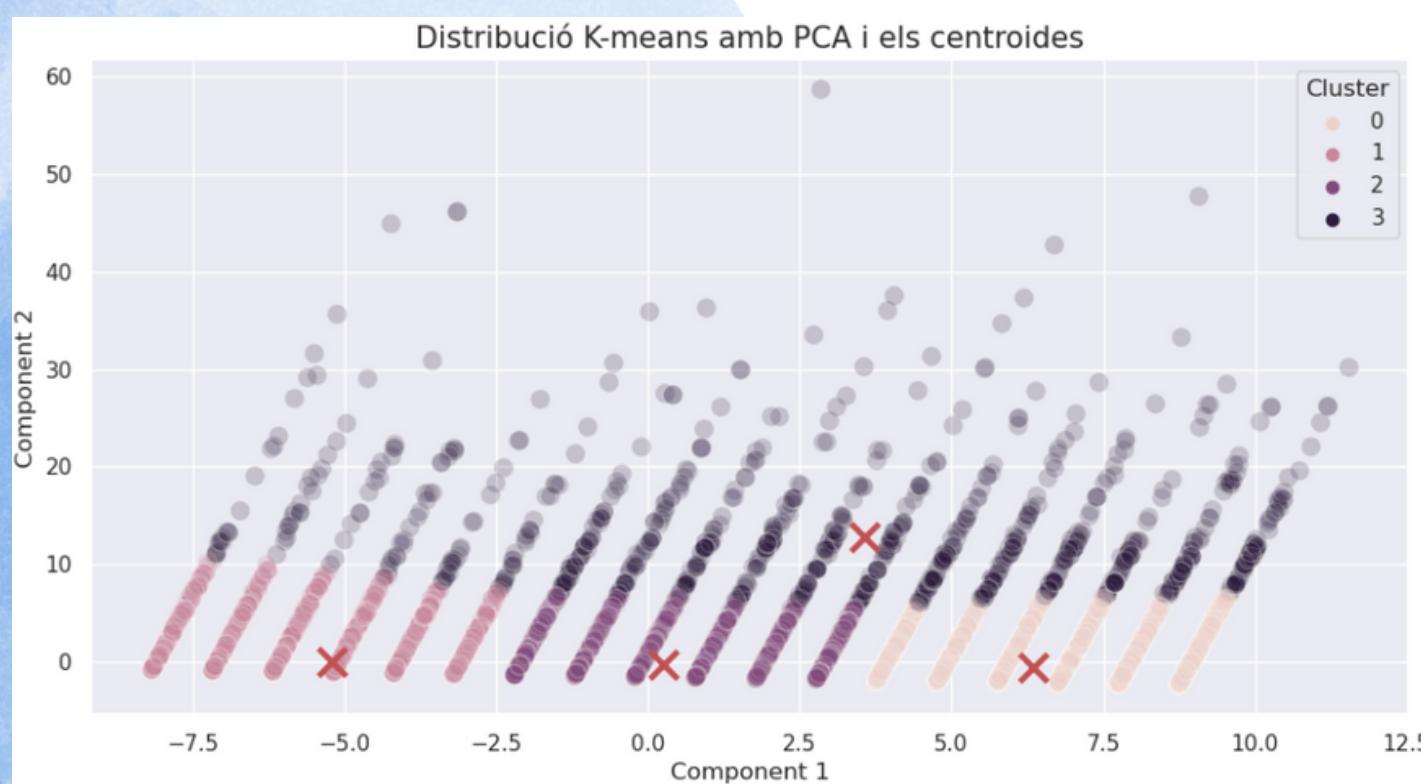
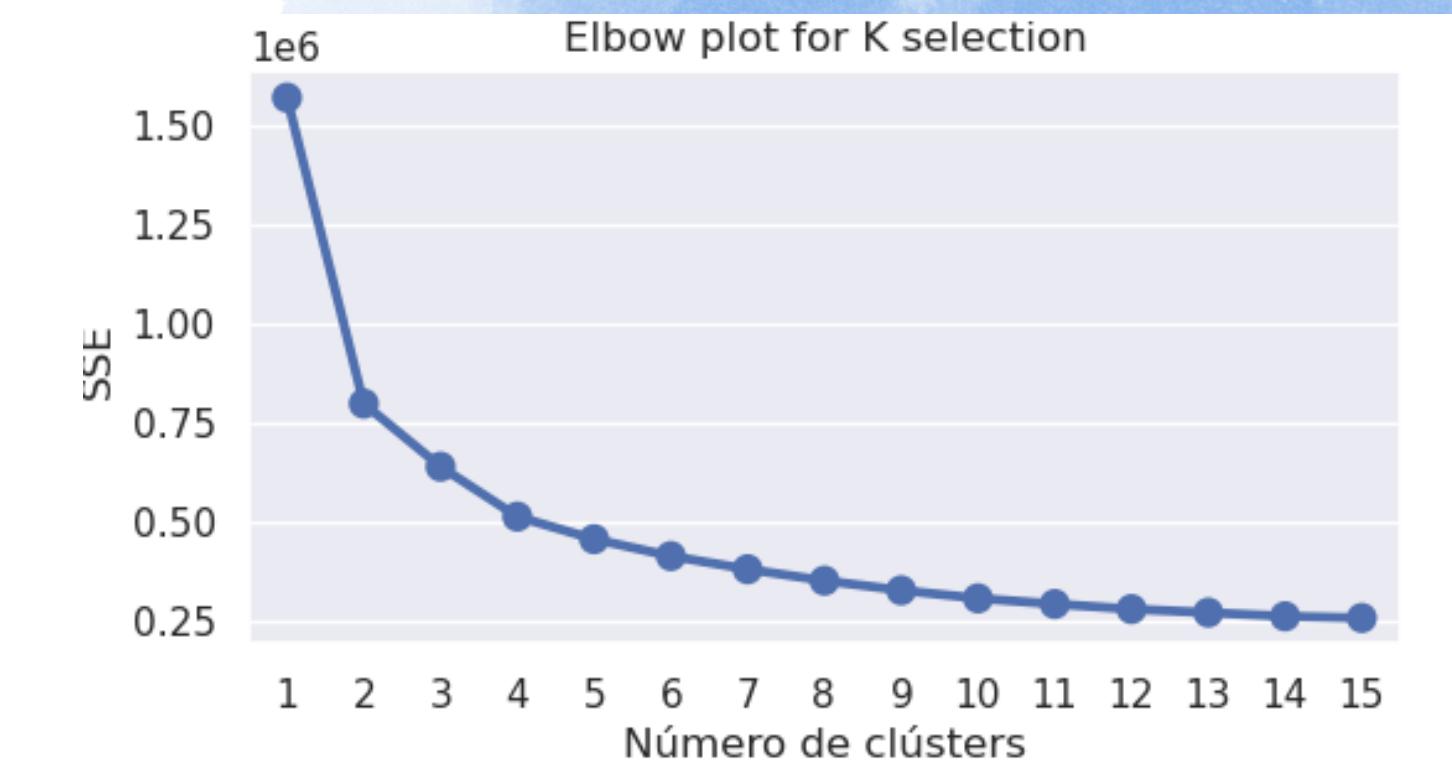
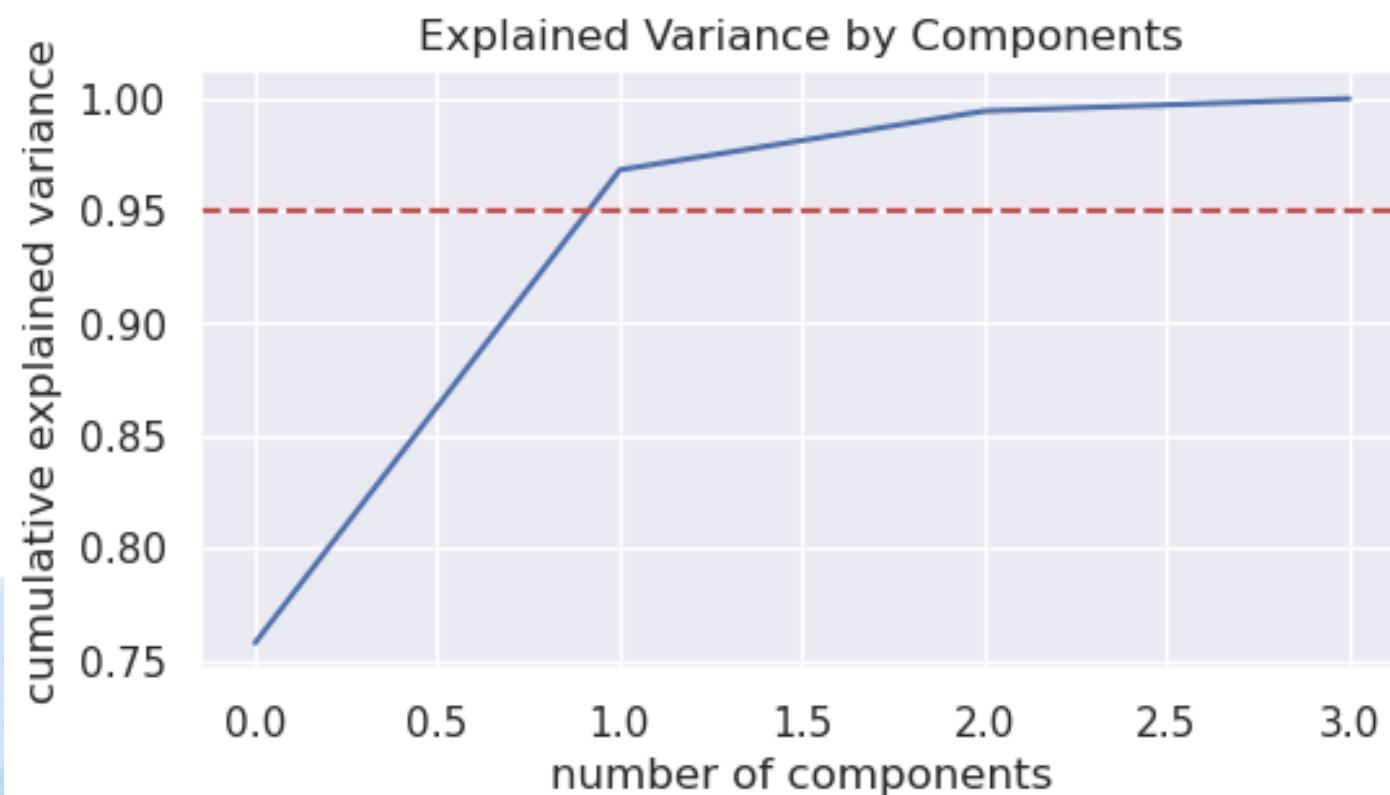


**3**

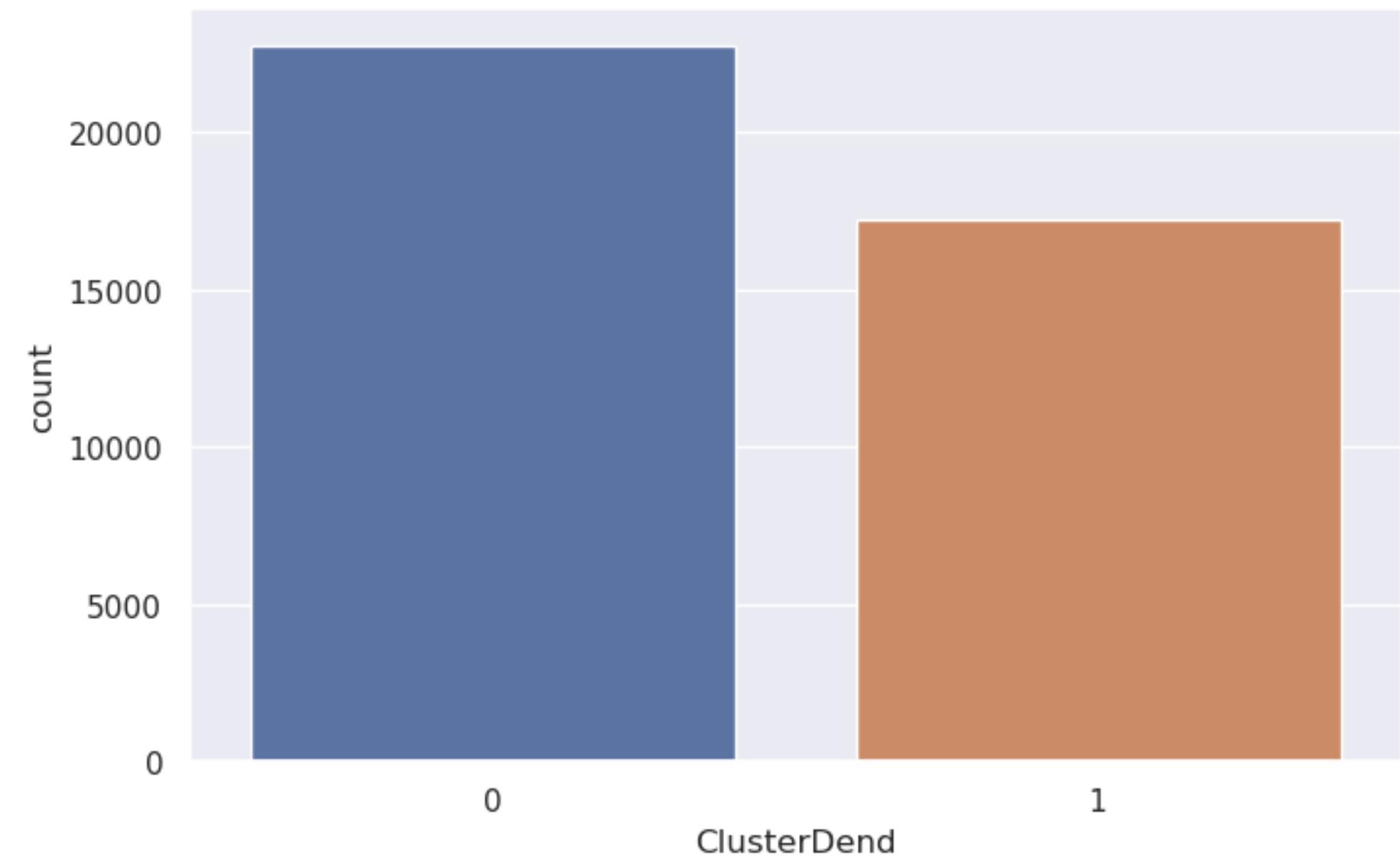
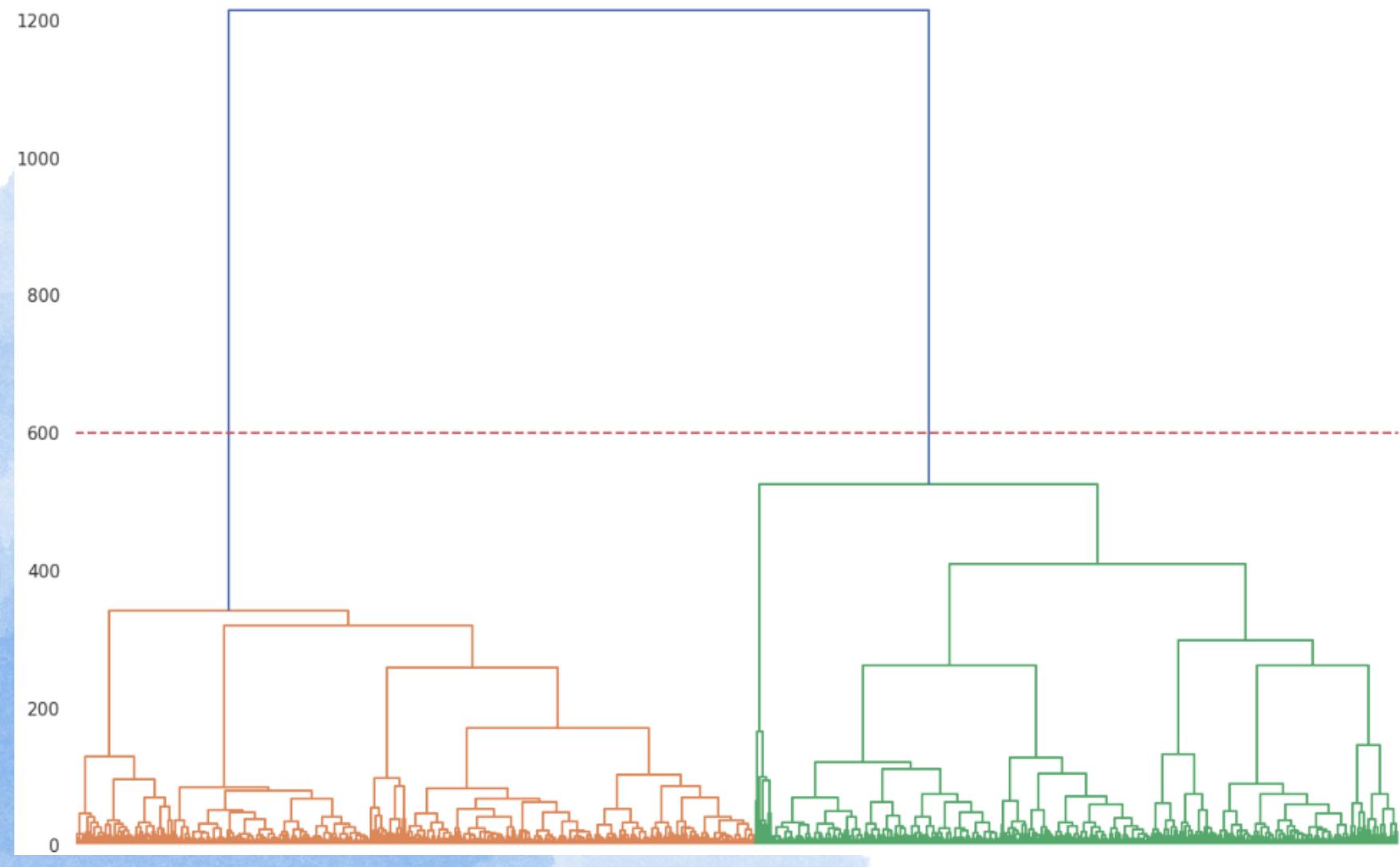
**MODELS**



# K-means



# Agglomerative Clustering

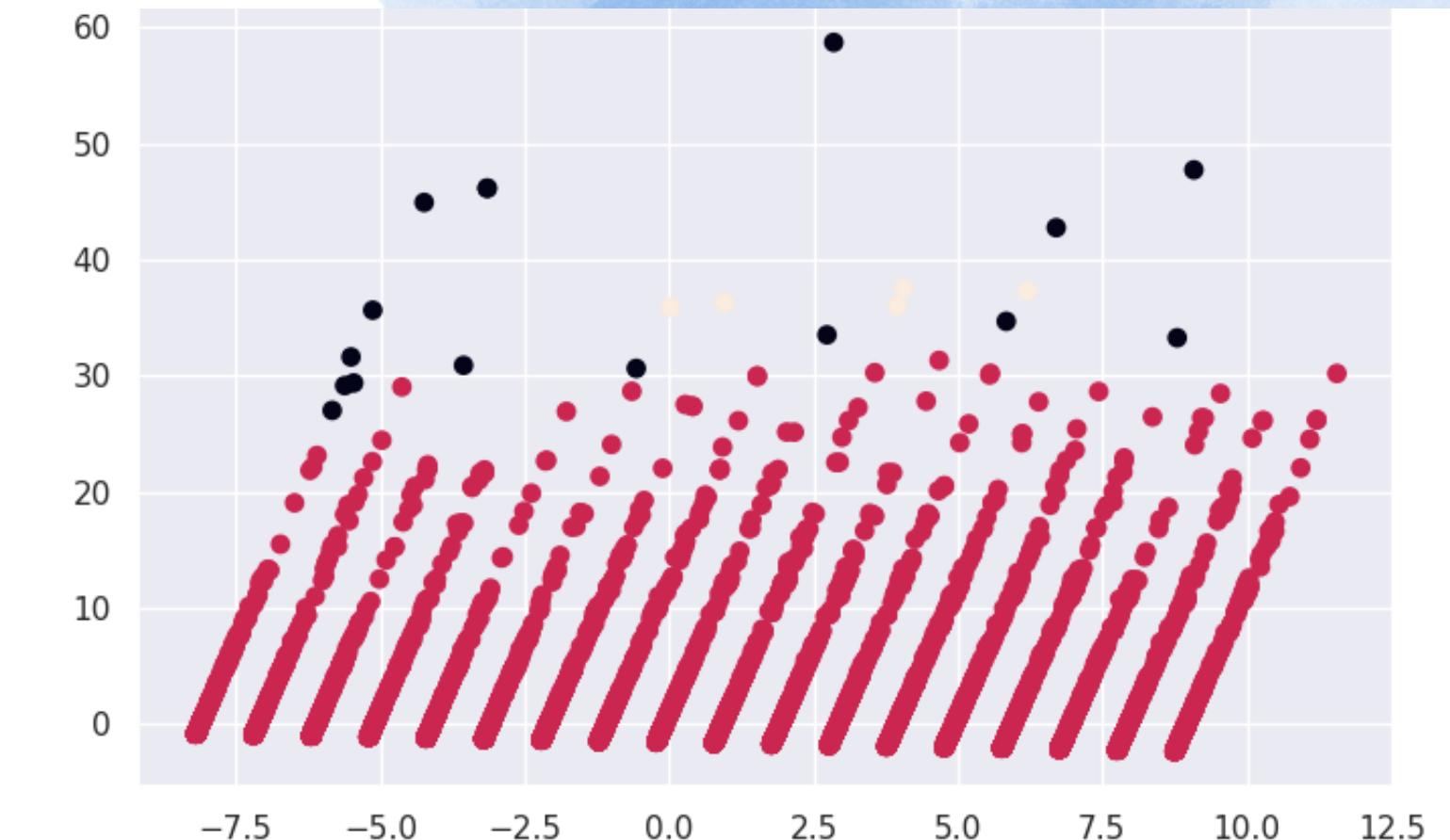


# DBSCAN

**DBSCAN és un algoritme de clustering basat en la densitat.**

**En lloc de considerar els clústers com a àrees amb densitats similars, defineix un clúster com un conjunt de punts que estan propers entre si i que tenen una densitat suficientment alta, separats per regions amb densitats més baixes.**

- Les anomalies se etiqueten con -1
- No requereix especificar el nombre de clústers
- En comparació amb el k-means, és conegut per funcionar millor en casos on aquest últim falla.
- Només s'han de definir dos paràmetres:
  - Epsilon ( $\varepsilon$ ): determina la distància màxima que s'utilitzarà per considerar dos punts com a propers.
  - MinPts: indica el nombre mínim de punts que han de ser considerats veïns d'un punt perquè sigui considerat un nucli.



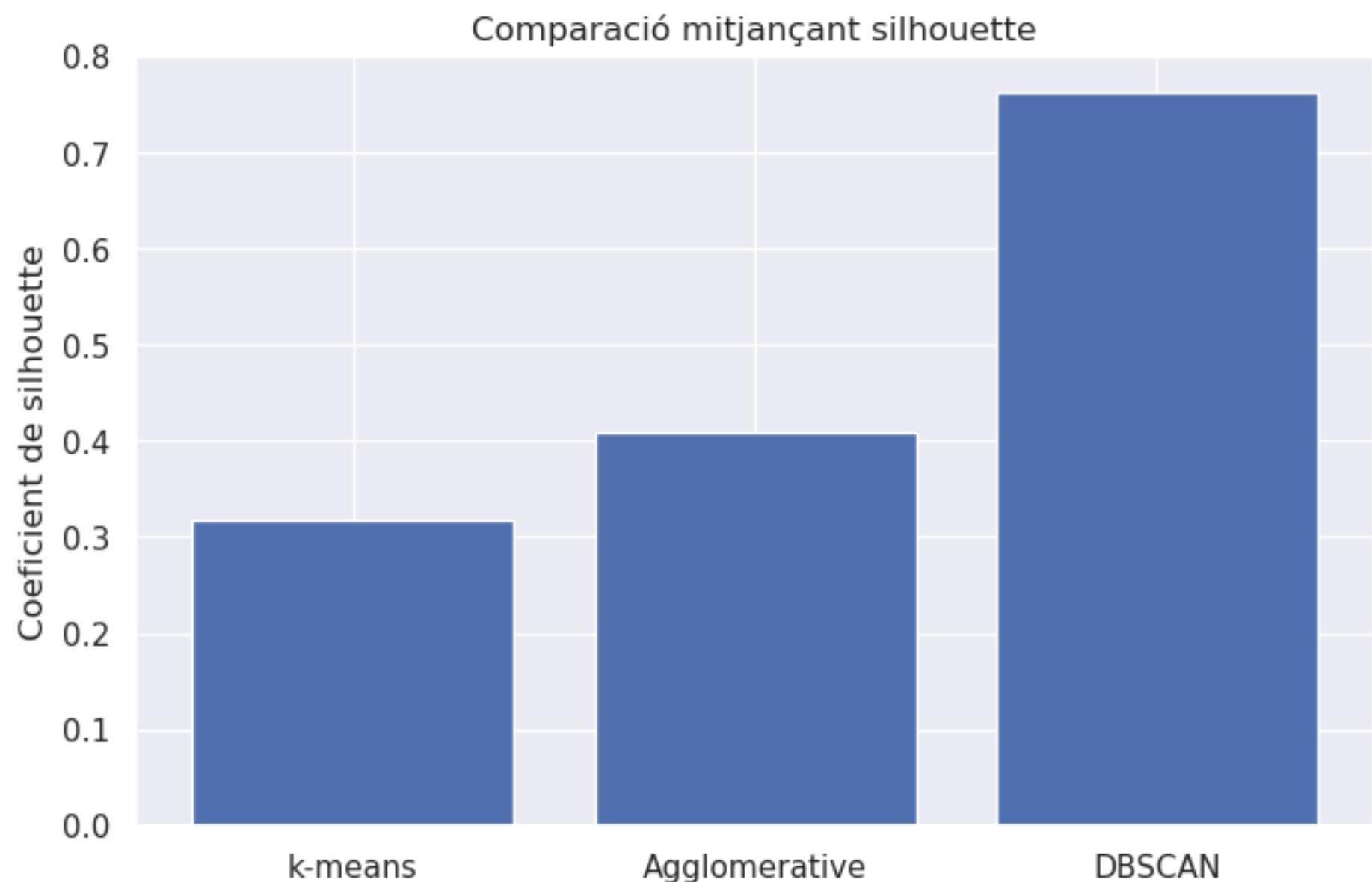
# Resultats

## Avaluació dels models amb Silhouette.

El resultat de Agglomerative és: 0.408772

El resultat de k-means és: 0.317871170659

El resultat de DBSCAN és: 0.7620171968695



**Sens dubte el millor és el DBSCAN amb un 76% contra un 31 i un 40.**

# Característiques de l'infant

- Té al voltant de cinc anys
- La llengua predominant a casa és l'anglès
- Viuen majoritàriament fora de les ciutats
- Les famílies viuen en una casa que ha sigut comprada i actualment tenen un préstec hipotecari
- L'infant en qüestió sol ser el primer dels germans.
- Hi ha dos infants a casa
- Són de raça caucàsica i de sexe masculí
- Respecte a les variables de medicació, en tots els casos s'ha indicat majoritàriament "no" o "desconegut" en el cas de més de dotze mesos, suggerint que aquest grup no utilitza habitualment medicació i, si ho fa, seria de manera ocasional.
- El rang de pes de l'infant es troba entre -0.562634 i -0.554728, ambdues opcions proporcionant el mateix resultat.
- El pes de la llar seria -0.568319. Malauradament, no es va poder obtenir el pes en kg degut a la falta de dades.





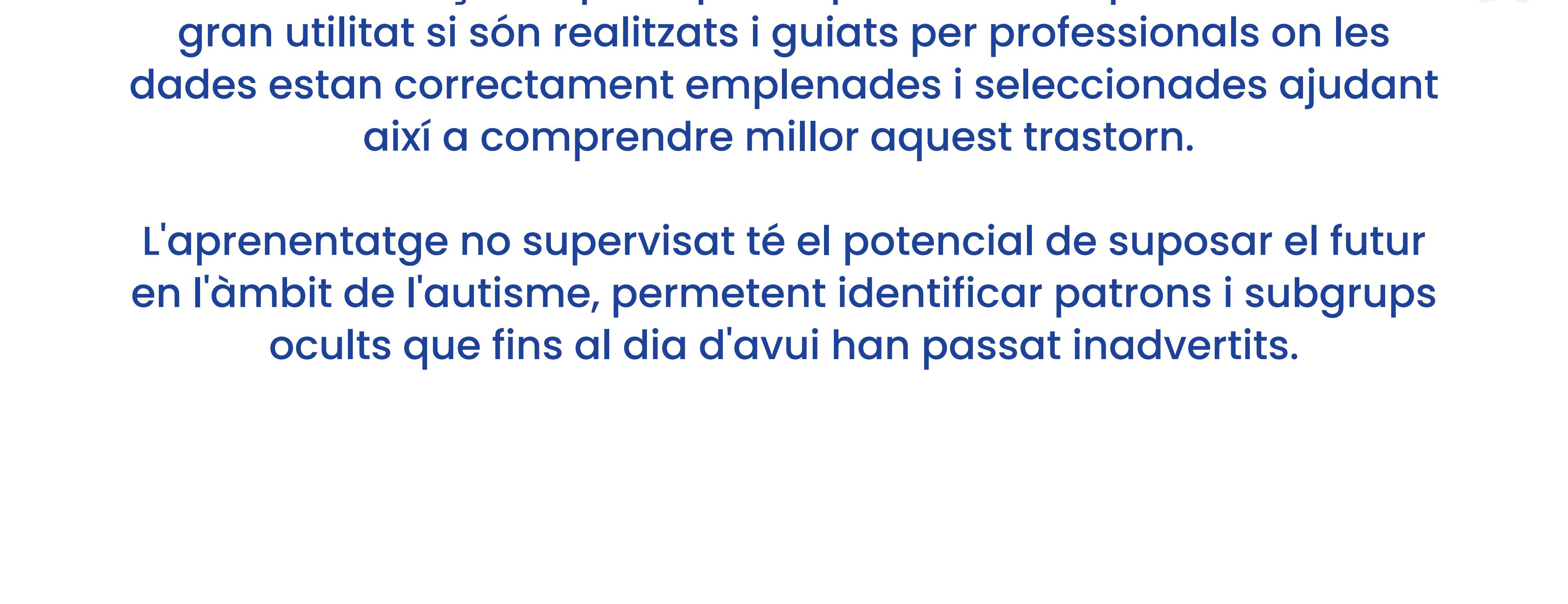
**4**

# **CONCLUSIONS**





Estic convençuda que aquest tipus d'estudis poden ser de gran utilitat si són realitzats i guiats per professionals on les dades estan correctament emplenades i seleccionades ajudant així a comprendre millor aquest trastorn.



L'aprenentatge no supervisat té el potencial de suposar el futur en l'àmbit de l'autisme, permetent identificar patrons i subgrups ocuts que fins al dia d'avui han passat inadvertits.

Moltes  
gràcies!