Detecció de patrons en l'autisme mitjançant ML no supervisat

Diana Hueso Beneyto
lutjens30@gmail.com

ITAcademy
Barcelona, Espanya

***Abstract*—A partir del resultat del model no supervisat de DBSCAN aplicat al conjunt de dades, s'han identificat diversos clústers, essent el més gran el grup zero. En analitzar aquest grup en detall, s'ha elaborat un perfil general de l'infant amb autisme, incloent característiques com edat aproximada de cinc anys, predominança de la llengua anglesa a casa, residència principal fora de les ciutats, habitatge comprat amb préstec hipotecari, primogènit de la família, presència de dos germans, raça caucàsica i gènere masculí. Pel que fa als aspectes mèdics, s'ha observat un patró de resposta predominant de "no" o desconegut quant a l'ús de medicació durant més de dotze mesos. Tot i que no s'ha pogut determinar amb precisió el pes de l'infant en quilograms, aquesta informació podria ser rellevant per a futures investigacions. Els objectius han estat identificar patrons o característiques comunes en el conjunt de dades utilitzant models no supervisats, com ara k-means, Agglomerative Hierarchical Clustering i DBSCAN."**

***Index Terms*—Autisme, DBSCAN, machine learning, Model no supervisat, patrons**

I. Introducció

L'autisme és un trastorn del desenvolupament neurològic que afecta la comunicació, les habilitats socials i el comportament d'individus de diferents edats. L'augment de la prevalença de l'autisme en els últims anys ha generat un interès creixent per comprendre millor aquest trastorn i desenvolupar mètodes eficaços per a la seva detecció i diagnòstic precoç.

En els últims temps, els models no supervisats basats en l'aprenentatge automàtic (Machine Learning) han emergit com a eines prometedores per identificar patrons i anomalies en grans conjunts de dades sense la necessitat de l'etiquetatge manual. Aquests models permeten descobrir relacions ocultes i generar coneixement a partir de dades no estructurades o amb poca informació prèvia.

En aquest context, el present treball té com a objectiu explorar l'ús d'un model no supervisat de Machine Learning per a la detecció de patrons en el context de l'autisme. Es pretén identificar patrons o característiques comunes en un conjunt de dades relacionades amb l'autisme de 119.241 registres.

Aquesta recerca té la finalitat d'obrir noves oportunitats per a l'exploració de factors clau en el trastorn. El desenvolupament d'un model no supervisat de Machine Learning per a la detecció de patrons en l'autisme podria millorar la capacitat de detecció precoç, permetent una intervenció més temprana i millorant els resultats i la qualitat de vida dels individus afectats.

II. Revisió de literatura

La investigació en el camp de l'autisme ha experimentat avenços significatius en les darreres dècades. Numerosos estudis s'han centrat a comprendre els aspectes clínics, genètics i neurològics del trastorn. No obstant això, hi ha encara molt de desconeixement sobretot quan la causa no es atribuïble a la genètica, el que es denomina factors ambientals.

En primer lloc, s'ha observat que el diagnòstic precoç de l'autisme és fonamental per a una intervenció i suport efectius. Tot i que s'han desenvolupat diverses eines d'avaluació i qüestionaris per detectar signes d'autisme, encara persisteix la necessitat de mètodes més precisos i eficients que permetin una detecció primerenca i fiable.

A més, la detecció d'anomalies específiques en l'autisme continua sent un repte. Tot i que s'han identificat alguns patrons i característiques comunes en les persones amb autisme, encara hi ha una necessitat d'una comprensió més profunda de les diferents manifestacions i subtipus d'aquest trastorn. La capacitat de detectar i classificar anomalies específiques podria contribuir a una avaluació més individualitzada i a un tractament més personalitzat.

En resum, malgrat els avenços en la investigació de l'autisme, encara hi ha grans mancances en el coneixement general més enllà de la mínima part atribuïble a factors genètics que són entre un 10 i 15 per cent. En altres paraules no se sap determinar la causa. L'aplicació de models no supervisats d'Aprenentatge Automàtic podria proporcionar noves perspectives en la detecció i la identificació d'anomalies en l'autisme. A continuació, es descriurà en detall la metodologia utilitzada en aquest projecte i es presentaran els resultats obtinguts.

III. Metodologia

En aquesta investigació, es va treballar amb un conjunt de dades que consistia en aproximadament 40.000 mostres extretes del Cens Nacional de Salut Infantil dels Estats Units sobre l'autisme. Es va realitzar una anàlisi exploratòria inicial i una neteja de dades per seleccionar només les columnes rellevants per aquest projecte. Aquestes columnes estaven compostes per 16 variables, de les quals 4 eren numèriques i

les altres eren categòriques.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 40000 entries, 106920 to 109329
Data columns (total 16 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   FIPSST        40000 non-null  object
 1   TOTKIDS_R     40000 non-null  float64
 2   TENURE        40000 non-null  float64
 3   HHLANGUAGE    39712 non-null  float64
 4   METRO_YN      35983 non-null  float64
 5   MPC_YN        33129 non-null  float64
 6   FWH           40000 non-null  float64
 7   LINENUM       40000 non-null  float64
 8   C_AGE_YEARS   40000 non-null  float64
 9   C_RACE_R      40000 non-null  float64
 10  C_HISPANIC_R  40000 non-null  float64
 11  C_SEX         40000 non-null  float64
 12  C_K2Q10       39937 non-null  float64
 13  C_K2Q12       5223 non-null   float64
 14  C_K2Q13       39929 non-null  float64
 15  C_FWS         40000 non-null  float64
dtypes: float64(15), object(1)
memory usage: 5.2+ MB
```

Fig. 1. Variables seleccionades

Posteriorment, es va crear un mapa interactiu amb el nombre de casos per estat utilitzant les llibreries Geopandas i Folium.

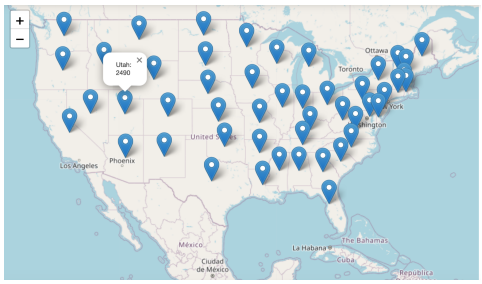

Fig. 2. Mapa d'estats amb els casos.

A continuació, es va dur a terme un pre-processament de les variables numèriques, ja que les variables categòriques ja havien estat transformades prèviament. A més, es va realitzar un Anàlisi de Components Principals (PCA) per reduir la dimensionalitat.

Seguidament, es van provar tres models no supervisats: k-means, AgglomerativeClustering i DBSCAN.

- El **K-means** és un algoritme de clustering que busca dividir el conjunt de dades en un nombre predeterminat de clústers, minimitzant la suma dels quadrats de les distàncies entre les mostres i els centroides dels clústers. Utilitza la distància euclidiana com a mètrica per avaluar la similitud entre les mostres.

- **L'AgglomerativeClustering** és un algoritme de clustering jeràrquic que agrupa les mostres successivament mitjançant l'enllaçament de punts de dades basat en diverses mètriques, com ara la distància euclidiana, la distància de Manhattan o la distància de Mahalanobis.

-**DBSCAN** és un algoritme de clustering basat en la densitat. Aquest mètode agrupa les mostres basant-se en la densitat de les seves regions i classifica les mostres com a nuclis, punts de frontera o soroll. Utilitza la distància de Mahalanobis com a mètrica per avaluar la similitud entre les mostres.

Per avaluar la qualitat dels clústers generats, es va utilitzar el coeficient de silueta (Silhouette). Després d'aquesta avaluació, es va seleccionar el millor model, que en aquest cas va ser el DBSCAN, ja que va demostrar un rendiment superior.

IV. Resultats

En primer lloc, es va utilitzar l'Anàlisi de Components Principals (PCA) per reduir la dimensionalitat del conjunt de dades. Després de realitzar aquest anàlisi, es va obtenir que dos components principals explicaven la major part de la variabilitat en les dades.

Seguidament, es va procedir a utilitzar el mètode del colze (elbow method) amb el model de k-means per determinar el nombre òptim de clústers i es va determinar que quatre clústers proporcionaven una bona agrupació de les dades.

A continuació, es va utilitzar un dendrograma per suggereix el nombre de clústers per al mètode d'AgglomerativeClustering . Segons el dendrograma, es va observar que dues agrupacions principals semblaven ser adequades per al conjunt de dades.

Finalment, es va aplicar el mètode DBSCAN, que no requereix especificar el nombre de clústers. Després de l'aplicació d'aquest mètode, es van obtenir tres clústers. Es va observar que el clúster -1 només agrupa soroll, anomalies o outliers, mentre que el clúster nº 1 tenia molt poques referències.

Per avaluar el rendiment dels models, es va utilitzar el coeficient de silueta (silhouette). Els resultats obtinguts van ser els següents:

- El resultat de l'AgglomerativeClustering: 0.40877
- El resultat del k-means: 0.31787
- El resultat del DBSCAN: 0.76201

A partir d'aquesta avaluació, es va seleccionar el model DBSCAN com el millor per aquest conjunt de dades, ja que va obtenir el coeficient de silueta més alt.

La interpretació del clúster nº 0, que el més gran, va permetre establir un perfil general de l'infant que inclou les següents característiques:

- 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.
- Són 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. Tot i aquest inconvenient, es considera que aquesta informació pot ser útil perquè, si algú pogués trobar la fórmula exacta, es tindria un gran valor per completar el perfil i potser detectar sobrepès, entre altres aspectes.

Cal destacar que el mètode DBSCAN, en comparació amb el k-means, és conegut per funcionar millor en casos on aquest últim falla. Això es reflecteix en el resultat més alt obtingut amb el DBSCAN en comparació amb el k-means.

V. Conclusions

Tot i que el resultat obtingut amb el model sigui discretament bo, cal destacar que, aquest projecte té algunes limitacions, com ara la disponibilitat de dades completes, el nombre de registres i l'ús de mètodes no supervisats, afectant els resultats, ja que depenen molt de les dades seleccionades.

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.

References

[1] Confederación Autismo España - *https://autismo.org.es/*
[2] United States Census Bureau - *https://www.census.gov/*
[3] Exponentis - *http://exponentis.es/ejemplo-de-uso-de-dbscan-en-python-para-deteccion-de-outliers*
[4] Scikit-learn - *https://scikit-learn.org/*
[5] DataScientits - *https://datascientest.com/es/machine-learning-clustering-dbscan*