Predicció del nivell d'estrès basat en factors ambientals a Barcelona

Lucía Revaliente Torres i Aránzazu Miguélez

Aprenentatge Computacional, 2024

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Índex

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

1. Introducció

En el marc de l'assignatura d'Aprenentatge Computacional, hem realitzat un projecte final amb l'objectiu de consolidar tots els coneixements adquirits al llarg del curs. Aquest projecte representa una oportunitat per posar en pràctica les competències apreses, des de la gestió de dades fins a l'aplicació de models d'aprenentatge automàtic, dins un context obert que fomenta la creativitat i la col·laboració en equip.

El nostre projecte, titulat "Predicció del nivell d'estrès basat en factors ambientals a Barcelona", té com a objectiu identificar i analitzar la relació entre diversos factors ambientals, com la qualitat de l'aire, i els nivells d'estrès a la ciutat. Per fer-ho, hem utilitzat un conjunt de dades extret de Kaggle, concretament el dataset "Air Pollution and Mental Health", que ens proporciona informació detallada sobre contaminació ambiental a Barcelona (BCN) i indicadors de salut mental.

Hem desenvolupat el projecte utilitzant Python com a llenguatge principal, gràcies a la seva versatilitat i a la seva àmplia gamma de biblioteques especialitzades en gestió de dades i aprenentatge automàtic. A més, hem integrat eines com GitHub per gestionar el codi, col·laborar de manera efectiva i mantenir un registre detallat de les versions del projecte.

El procés de desenvolupament ha estat constantment guiat per preguntes que ens ajudaven a validar les nostres decisions i a assegurar que cada pas tenia una base tècnica i conceptual sòlida. Aquest enfocament crític ens ha permès aprofundir en l'ús de recursos bibliogràfics i electrònics per recolzar les nostres decisions tècniques, alhora que ens ha ajudat a desenvolupar una visió més madura sobre les solucions implementades.

Finalment, aquest projecte també ens ha servit per explorar la capacitat dels models d'aprenentatge automàtic d'aportar informació útil en un tema tan rellevant com l'impacte ambiental sobre la salut mental. És important destacar que no explicarem el codi en detall, sinó que ens centrarem en descriure les idees principals, els passos seguits amb el seu perquè i els resultats obtinguts.

2. Objectius

Els objectius del nostre projecte són els següents:

1. Predir indicadors de salut mental utilitzant models de regressió, concretament en l'estrès.
2. Observar si les característiques més importants sobre salut mental formen clústers. És a dir, si hi ha una clara segmentació en funció de les variables influents en l'estrès.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

3. Dades

Com hem explicat en la introducció, en aquest projecte hem utilitzat el dataset titulat "Air Pollution and Mental Health". Aquest conjunt de dades proporciona informació detallada sobre la relació entre la contaminació de l'aire a BCN i la salut mental, permetent una anàlisi profunda de com els nivells de contaminació poden influir en diversos aspectes del benestar psicològic. Les dades, que han estat recopilades mitjançant enquestes, estudis psicològics i dades medioambientals, contenen variables com nivells de diferents contaminants (PM2.5, NO2, SO2, etc.), indicadors de salut mental (taxes de depressió, ansietat, etc.), i dades demogràfiques. Per una descripció detallada de cada característica, es pot consultar la documentació de Kaggle.

3.1. Importació

Les dades presentades estan disponibles en format CSV. L'script **/scripts/load_data.py** carrega les dades des de **data/CitieSHealth_BCN_DATA_ PanelStudy_20220414.csv** i les desa en un arxiu de tipus Pickle a la ubicació **data/dataset.pkl**.

Els motius pels quals hem decidit fer la transformació de formats són diversos. En primer lloc, Pickle és més ràpid que CSV quan es tracta de carregar dades en la memòria ja que és un format binari que permet una càrrega més eficient. En segon lloc, a diferència de CSV, que guarda les dades en format de text, Pickle conserva tots els tipus de dades originals, com ara números, dates o valors booleans, sense necessitat de convertir-los. Per últim, el fitxer Pickle ocupa menys espai en memòria que el CSV, fet que facilita la manipulació de conjunts de dades més grans.

Doncs, emprar Pickle davant CSV permet una major eficàcia i velocitat, manteniment dels tipus de dades i un menor ús dels recursos.

3.2. Anàlisi exploratori

En aquest apartat, exposarem un anàlisi general del dataset, el qual ens ha ajudat entendre millor les dades i acotar la variable objectiu del projecte. És per aquest motiu que exposarem els tipus de dades que manipularem, valors null, ouliers, distribucions, entre d'altres. Les comandes executades per obtenir la informació exposada es troba en l'arxiu **scripts/exploratory_analysis.py**.

3.2.1. Característiques i tipus de dades

En primer lloc, hem observat que el dataset té 3348 files (instàncies) i 95 columnes (característiques). Doncs, hem dividit les característiques del dataset en:

- **Dades temporals:** *year, month, day, hour, dayoftheweek, etc.* Informació útil per analitzar variacions temporals en la contaminació o en la salut mental.
- **Indicadors de salut mental:** *occurrence_mental, bienestar, energia, estrès, sueno, etc.* Dades reportades pels participants que poden servir com a variables dependents per predir l'impacte de la contaminació.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

- **Contaminació ambiental:** *no2bcn_24h, no2bcn_12h, pm25bcn, etc*. Concentracions de contaminants atmosfèrics. BCμg: Black Carbon, rellevant per a la salut respiratòria i mental. sec_noise55_day, sec_noise65_day: Exposició al soroll. hours_greenblue_day: Temps d'exposició a espais verds/blaus.
- **Dades demogràfiques i personals:** *age_yrs, gender, education, district, etc*. Context de l'individu per estudiar efectes diferenciats segons la població.
- **Factors relacionats amb la COVID-19:** *covid_work, covid_mood, covid_sleep, etc*. Canvis de comportament durant la pandèmia.
- **Condicions meteorològiques:** *tmean_24h, humi_24h, pressure_24h, precip_24h, etc.* Factors climàtics que poden influir en la salut mental.
- **Exposició espacial i temporal:** *min_gps, hour_gps, access_greenbluespaces_300mbuff*, etc. Dades de localització i accés a espais verds/blaus.

Com podem observar, hi ha diferents tipus de dades, els quals presentem a continuació:

1. **Dades numèriques:** dades numèriques són valors representats per nombres que poden ser manipulats matemàticament. P.e: *μgm3, BCμg, tmean_24h* o *humi_24h*.
2. **Dades categòriques:**
    a. **Ordinals:** categories que tenen un ordre o jerarquia natural, però les diferències entre elles no són mesurables numèricament. P.e: *education = ["primario o menos", "bachillerato", "universitario"]* o *covid_work = ["ha empeorado mucho", "ha empeorado un poco", "no ha cambiado", "ha mejorado un poco", "ha mejorado mucho"]*.
    b. **Nominals**: categories sense cap ordre inherent. Es tracta de grups únicament identificables per les seves etiquetes. P.e: *district = [“eixample”, “gràcia”, “sant martí”…]*.
    c. **Binàries**: representa dues categories o estats diferents. P.e: *mentalhealth_survey* o *Totaltime_estimated*.

Com veurem més endavant, aquesta diversitat requerirà de tècniques de preprocessament específiques, com la codificació de dades categòriques i l'escalat de dades numèriques, per garantir que totes les variables siguin adequades per als models utilitzats.

3.2.2. Missing values:

En segon lloc, hem identificat un **1,05% de valors null** en el conjunt de dades (3348 registres afectats de 318.060). A més, hem analitzat la distribució per columnes:

1. **Característiques amb <5% valors null:** ['ID_Zenodo', 'date_all', 'year', 'month', 'day', 'dayoftheweek', 'hour', 'mentalhealth_survey', 'occurrence_mental', 'bienestar', 'energia', 'sueno', 'horasfuera', 'actividadfisica', 'ordenador', 'dieta', 'alcohol', 'drogas', 'bebida', 'enfermo', 'otrofactor', 'stroop_test', 'no2bcn_24h', 'no2bcn_12h', 'no2bcn_12h_x30', 'no2bcn_24h_x30', 'min_gps', 'hour_gps', 'pm25bcn', 'tmean_24h', 'tmean_12h', 'humi_24h', 'humi_12h', 'pressure_24h', 'pressure_12h', 'precip_24h', 'precip_12h', 'precip_12h_binary', 'precip_24h_binary', 'maxwindspeed_24h', 'maxwindspeed_12h', 'gender', 'district', 'covid_work'].
2. **Característiques amb >5% i <10% valors null:** ['estrès', 'occurrence_stroop', 'mean_incongruent', 'correct', 'response_duration_ms', 'performance', 'mean_congruent', 'inhib_control', 'z_performance', 'z_mean_incongruent',

'z_inhib_control', 'no2gps_24h', 'no2gps_12h', 'no2gps_12h_x30', 'no2gps_24h_x30', 'BCµg', 'noise_total_LDEN_55', 'access_greenbluespaces_300mbuff', 'µgm3', 'incidence_cat', 'start_day', 'start_month', 'start_year', 'start_hour', 'end_day', 'end_month', 'end_year', 'end_hour', 'Totaltime', 'Totaltime_estimated', 'Houron', 'Houroff', 'age_yrs', 'yearbirth', 'education', 'covid_mood', 'covid_sleep', 'covid_espacios', 'covid_aire', 'covid_motor', 'covid_electric', 'covid_bikewalk', 'covid_public_trans'].

3. **Característiques amb >10% valors null:** ['sec_noise55_day', 'sec_noise65_day', 'sec_greenblue_day', 'hours_noise_55_day', 'hours_noise_65_day', 'hours_greenblue_day', 'smoke', 'psycho'].

Doncs, observem uns valors null dispersos al llarg de les columnes i files els quals no provoquen una gran pèrdua d'informació.

3.2.3. Outliers i distribucions de les dades

En tercer lloc, hem estudiat els **outliers** del dataset. En l'Annex 2, hi ha un anàlisi profund.

Concloem, mitjançant **boxplots**, que no hi ha cap valor incorrecte, només valors atípics dins del rang esperat. És a dir, els valors atípics observats no representen errors en les dades, sinó que són punts que, tot i estar lluny de la majoria de les observacions, encara es poden considerar dins d'un rang plausible o raonable per al conjunt de dades. Aquests outliers poden reflectir variabilitat natural o situacions excepcionals però no necessàriament haurien de ser eliminats, ja que podrien aportar informació valuosa per l'anàlisi.

Per tant, la seva presència no és un motiu per descartar les dades, sinó una oportunitat per investigar més a fons les causes d'aquests valors atípics i determinar si tenen alguna implicació per al model o per a les conclusions a obtenir.

Cal destacar que, posteriorment, vam fer servir **violin plots** ja que aquests ofereixen una visió més completa de la **distribució** dels valors, combinant informació sobre la densitat de les dades, els quartils i els outliers. A diferència dels boxplots, els violin plots permeten veure de manera més detallada la forma de la distribució i identificar possibles agrupaments o simetries en els valors. Això ens ha ajudat a comprendre millor la variabilitat dels valors i identificar possibles patrons que no són evidents amb altres tècniques de visualització.

Gràcies als violin plots, hem determinat que les distribucions de les dades són normals. La qual cosa ens facilita la presa de decisions en algortimes del preprocessament.

Les visualitzacions generades es poden observar en les carpetes del repositori: **visualizations/boxplots/boxplots.png** i **visualizations/violinplots/violinplots .png**. Destaquem que hem ajuntat totes les gràfiques per facilitar l'anàlisi i comparació entre ells.

3.2.4. Proporció de registres de salut mental

Com un dels nostres objectius és predir les variables de salut mental, hem estudiat la variable que indica si una persona té una malaltia mental o no és *mentalhealth_survey*. Gràcies al violin plot, entre d'altres, hem identificat un **desbalancejament significatiu**: només 13 individus diagnosticats amb problemes mentals enfront de 3335 individus sans.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Aquest desequilibri en la representació de les dades fa que aquesta variable sigui **inviable** com a objectiu del modelatge inicial, ja que podria introduir biaixos en les prediccions. En concret, el grup de malalts és tan petit que pot ser difícil extreure conclusions fiables sobre les seves característiques, la qual cosa dificultaria una predicció precisa per a aquest grup. A més, aquest esbiaix podria provocar que qualsevol model o anàlisi estigués esbiaixat cap al grup majoritari (els sans), ja que aquest tindria més pes en els càlculs estadístics i podria afectar les conclusions generals.

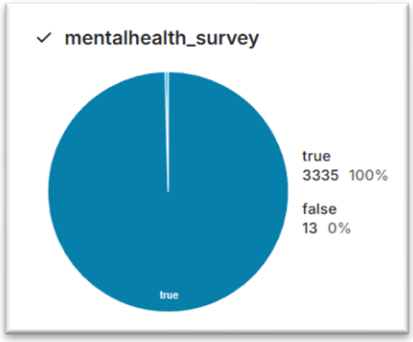


Figura 1. Proporció de valors en la variable binària *mentalhealth_survey,* que determina si una persona està malalta mentalment o no.

Per tant, **hem decidit descartar la variable *mentalhealth_survey* com a target** i hem optat per estudiar altres variables relacionades amb la salut mental, com el benestar, energia, estrès i son. Cal destacar que aquestes variables són subjectives, ja que es van recopilar mitjançant una enquesta, i prenen valors de 0 a 10. En la Figura 2, podem observar les seves distribucions.

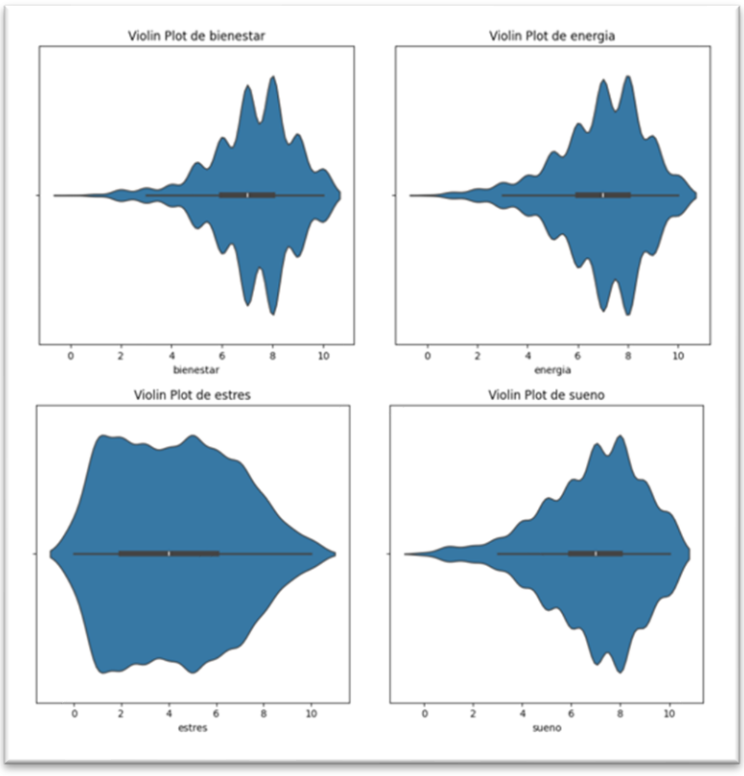


Figura 2. Distribucions de les variables de salut mental subjectives: bienestar, energia, estrès i sueno.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

3.2.5. Elecció de l'estrès com a variable objectiu

Després d'analitzar en detall la distribució de cada variable, considerem que **el nostre projecte es basarà en la variable estrès**, ja que és la que té una **variabilitat major**. Això significa que els valors d'aquesta variable mostren una àmplia dispersió, la qual cosa podria indicar una relació més estreta amb altres factors, com la qualitat de l'aire o altres condicions ambientals. A més, una major variabilitat pot permetre una millor capacitat per a la predicció i l'anàlisi de patrons, ja que proporciona més informació per modelar les dinàmiques entre les variables.

En conclusió, l'estrès ofereix un **potencial més gran** respecte la resta de variables per obtenir resultats significatius i utilitzar models d'aprenentatge automàtic de manera més efectiva.

3.2.6. Correlació i redundància entre variables

Un cop establerta la variable objectiu estrès, vam crear una **matriu de correlació (heatmap)** per analitzar les correlacions i redundàncies entre la resta de característiques. Doncs, en aquest apartat explicarem les variables amb correlació moderada o alta (positiva o negativa) que ens han cridat l'atenció, així com aquelles que poden tenir una rellevància important. En la Figura 3, es pot veure el resultat gràficament. També es pot visualitzar amb més qualitat en **visualizations/analisi_ correlacio/matriu_correlacio.png**.

Abans de continuar, ens agradaria repassar el concepte de correlació. Aquesta, mesura la força i la direcció de la relació lineal entre dues variables, representada pel coeficient de **correlació** r, que pot variar entre -1 i 1. Quan $r > 0$, indica que quan una variable augmenta, l'altra també tendeix a augmentar. En contraposició, quan $r < 0$, indica que quan una variable augmenta, l'altra tendeix a disminuir. Finalment, una correlació alta implica una relació forta entre dues variables, mentre que una baixa és una relació entre les variables és feble o inexistent.

Tornant a l'anàlisi del dataset, algunes de les correlacions que hem analitzat són les seguents:

1. Observem una **correlació moderada (positiva)** entre *estrès* i ***no2bcn_24h / no2gps_24h***. Això podria suggerir que l'exposició a nivells més alts de NO_2 durant 24 hores està associada a majors nivells d'estrès.
2. Observem una **correlació moderada (negativa)** entre *estrès* i ***bienestar*** . Això podria suggerir que a mesura que augmenta el nivell d'estrès, el benestar disminueix. Això és coherent amb el que s'espera en termes de salut mental: a més estrès, menys benestar.
3. Observem una **correlació moderada (negativa)** entre *estrès* i ***dayoftheweek***. Això podria suggerir que a mesura que s'acosta el cap de setmana, l'estrès disminueix.
4. Observem una **correlació lleu (negativa)** entre ***sueno*** i *estrès*. Doncs, menys hores de son estan associades a majors nivells d'estrès. Aquesta relació és consistent amb la literatura sobre els efectes de la privació del son en la salut mental.
5. Observem una **correlació lleu (negativa)** amb ***tmean_24h*** i *estrès*. Per tant, dies amb temperatures extremes podrien influir negativament en l'estrès, tot i que cal un anàlisi més profund.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

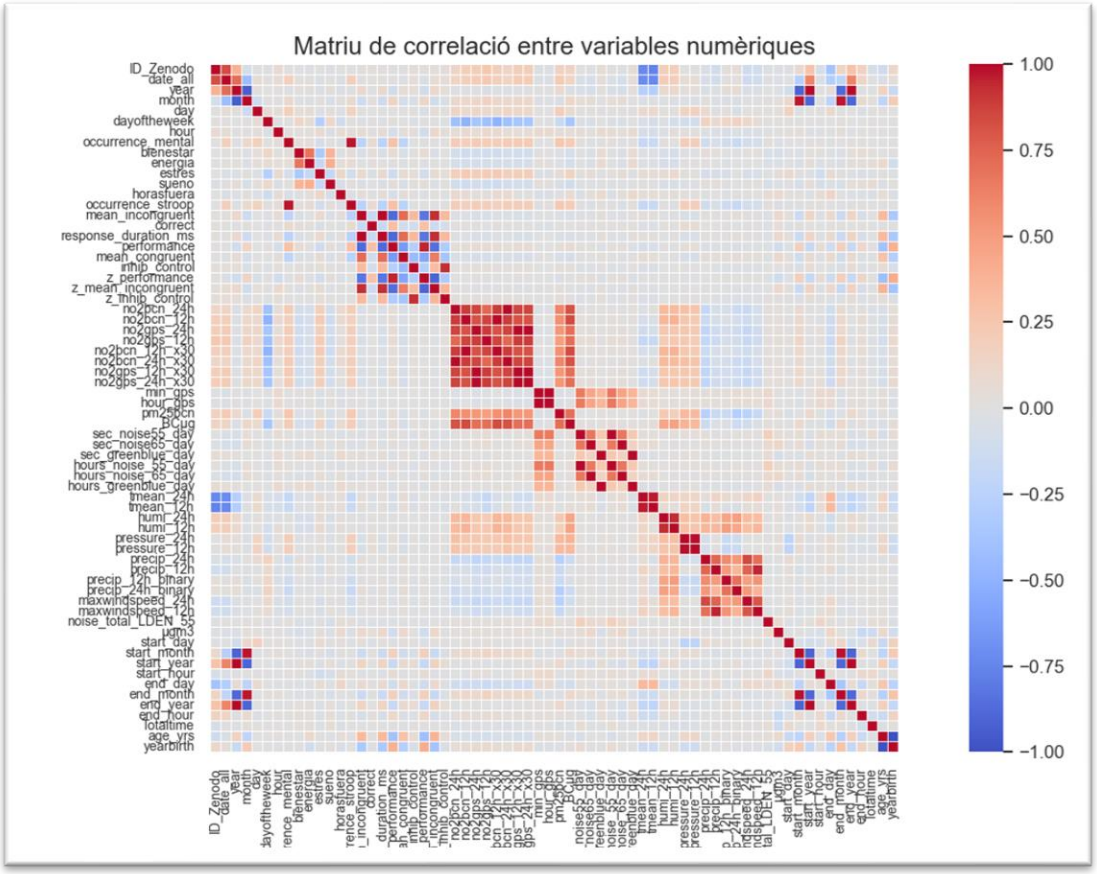


Figura 3. Matriu de correlació entre variables numèriques.

Cal destacar que, com encara no hem fet el preprocessament, només hem calculat la matriu amb les dades numèriques. El motiu és perquè aquesta es basa en fórmules que inclouen operacions matemàtiques com la mitjana i la desviació estàndard. Per tant, per incloure les dades categòriques, les hem de codificar.

Com no és l'objectiu del nostre projecte i simplement volem fer-nos una idea de les relacions inicials entre variables numèriques, no hem aprofundit en el càlcul de la matriu amb tot el conjunt de dades. Tot i això, en cas que es volgués continuar el treball, es podria recalcular al final del projecte per veure si les hipòtesi prèvies coincideixen amb els resultats de la regressió.

Finalment, hem determinat la **redundància** entre algunes variables. Per exemple, hem deteminat **innecessàries les dades temporals de l'enquesta** ja que no aporten informació, només serveixen per controlar les instàncies del dataset: *'date_all', 'year', 'month', 'hour', 'day', 'start_year','start_month', 'start_day', 'start_hour', 'end_year','end_month', 'end_day', 'end_hour', 'Houron', 'Houroff'*. D'igual forma, són redundants les variables que **només prenen un valor**: *'strooptest'*. No considerarem tampoc característiques que són **similars/factors d'unes altres**: *'yearbirth'* (igual que la variable age), *'no2gps_12hx30', 'no2gps_24hx30', 'no2gps_12h', 'no2bcn_12h_x30', 'no2bcn_24h_x30', 'no2bcn_12h', 'no2gps_12h_x30', 'no2gps_24h_x30'* (són factors de 30).

3.2.7. Hipòtesi inicial sobre les relacions entre variables

Finalment, per acabar l'anàlisi exploratori sobre el dataset, ens vam plantejar hipòtesis sobre com està influenciat l'estrès. Les més significatives són:

1. L'exposició a nivells més alts de NO_2 durant 24 hores està associada a majors nivells d'estrès.
2. A mesura que s'acosta el cap de setmana, l'estrès disminueix.
3. Contra menys benestar experimenti una persona, més estrèsada es sentirà.

En conclusió, aquelles variables amb una correlació forta (ja sigui positiva o negativa) influiran significativament en l'estrès d'una persona. No obstant això, és important destacar que aquest anàlisi s'ha fet només amb les variables numèriques, ja que la matriu de correlació no inclou dades categòriques . Per tant, la nostra hipòtesi queda parcialment validada, ja que podrien existir altres relacions importants amb les variables no numèriques que no hem considerat en aquesta etapa. Aquest fet posa de manifest la necessitat de complementar aquest tipus d'anàlisi amb un estudi més ampli que inclogui la codificació i la incorporació de totes les variables disponibles al conjunt de dades.

3.3. Neteja de les dades

Després d'analitzar les dades i identificar les seves característiques principals, el següent pas fonamental en el projecte és realitzar la **neteja de les dades**. Aquest procés és crucial per garantir la qualitat i fiabilitat del conjunt de dades abans d'aplicar qualsevol model d'aprenentatge automàtic. Per tant, en aquesta secció, detallem les accions que hem dut a terme per gestionar valors nuls, així com eliminar o transformar característiques que podrien introduir soroll o biaixos en l'anàlisi. Aquest procés ens assegura que les dades estiguin ben preparades per al preprocessament i modelatge posterior, mantenint-ne la consistència i la robustesa. L'script on hem realitzat la neteja és **scripts/data_cleaning.py**.

En primer lloc, hem **eliminat les característiques redundants** identificades prèviament. D'aquesta manera, hem millorat l'eficiència del model i hem reduït la complexitat, cosa que ens permet obtenir resultats més precisos i menys propensos a l'overfitting.

En segon lloc, hem **tractat els valors NULL**. Com que el percentatge de valors faltants és relativament petit però dispers al llarg de les files, hem decidit substituir-los i no eliminar-los. Bàsicament, perquè si no ens quedaríem sense instàncies.

En un inici, ens vam plantejar l'opció d'emprar la mitjana i la moda, però no ho vam fer servir perquè aquestes tècniques poden introduir biaixos importants. Aquestes mesures poden no reflectir correctament les relacions entre les diferents variables del conjunt de dades i, per tant, no són sempre les millors opcions per a la imputació de valors faltants.

Doncs, **KNNImputer** va ser preferit sobre aquest i altres mètodes perquè és una tècnica basada en l'aprenentatge automàtic que considera les relacions entre les variables per imputar els valors faltants de manera més informada i precisa. A més, KNNImputer permet tenir en compte la variabilitat i la complexitat del conjunt de dades, ja que utilitza les instàncies més properes per a cada valor faltant, la qual cosa fa que l'imputació sigui més adaptada al conjunt de dades en qüestió. Així, vam aconseguir una imputació més coherent amb les tendències globals del conjunt de dades.

En tercer lloc, hem intentat **eliminar les files duplicades**, de manera que el conjunt de dades es veiés més net i optimitzat. Tot i això, el resultat era el mateix dataset. Per tant, observem que no hi ha cap instància repetida. Aquesta situació suggereix que no era necessari realitzar cap acció de depuració en aquest aspecte, ja que no existien dades redundants que poguessin afectar l'anàlisi o modelatge posterior.

En quart lloc, hem fet una **conversió de tipus de dades incorrectes** per assegurar-nos que els formats eres adequats i no generaríen errors o biaxios en el processament posterior. Tot i que en la documentació de Kaggle s'estableixen els tipus esperats, en el dataset hi ha certes característiques que no tenen els valors o tipus esperats. Doncs, hem classificat les dades mal definides segons el seu tipus :

1. *transform_to_int = ['occurrence_mental', 'occurrence_stroop', 'correct', 'response_duration_ms', age_yrs', 'hour_gps', 'sec_noise55_day', 'sec_noise65_day', 'sec_greenblue_day', 'hours_noise_55_day', 'hours_noise_65_day', 'hours_greenblue_day', 'precip_12h_binary', 'precip_24h_binary', 'dayoftheweek', 'bienestar', 'energia', 'estrès', 'sueno']*
2. *transform_to_str = ['mentalhealth_survey', 'ordenador', 'dieta', 'alcohol', 'drogas', 'enfermo', 'otrofactor', 'district', 'education', 'access_greenbluespaces_300mbuff', 'smoke', 'psycho', 'gender', 'Totaltime_estimated']*

En cinquè lloc, hem **normalitzat i eliminat soroll de les dades categòriques**. Gràcies al mètode .value_counts(), hem observat que els valors de certes característiques contenen decimals, quan han de ser enters. Com hem explicat anteriorment, aquestes dades s'han recopilat mitjançant enquestes, i les possibles respostes són números en un rang del 0 al 10. Doncs, els valors float (soroll) poden generar inexactituds i complicar l'anàlisi. Les variables categòriques afectades són *bienestar, energia, estrès* i *sueno*. El resultat de la comanda esmentada es pot visualitzar en la Figura 4.

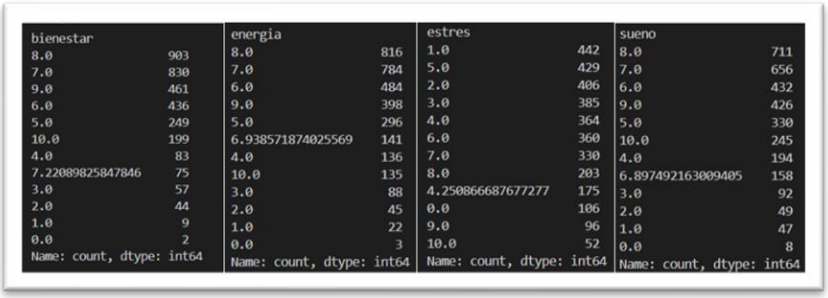


Figura 4. Soroll en les variables categòriques ordinals.

Per tal de corregir aquesta incongruència, simplement hem arrodonit tots els valors a l'enter més pròxim. Aquest procés de conversió ens ajuda a **optimitzar el conjunt de dades**, reduint les dades inexactes o inconsistents que podrien afectar els models d'aprenentatge automàtic.

Finalment, **no hem transformat els outliers** perquè no hi ha cap valor incorrecte, només valors poc comuns. **Tampoc hem estandaritzat les dades** textuals ja que no hem detectat cap error tipogràfic com "Barcelona" / "barcelona", espais en blanc innecessaris, etc.

En conclusió, aquesta fase de neteja de les dades **ha estat crucial per garantir que el conjunt de dades fos coherent, consistent i adequat per a l'anàlisi posterior**. Això ha

permès reduir el soroll, simplificar la complexitat del conjunt de dades i preparar-lo per ser esclat, transformat i utilitzat en els models posteriors amb més fiabilitat i eficàcia.

3.4. Preprocessament

Un cop finalitzada l'etapa de neteja de les dades, ens hem centrat en el preprocessament, un pas essencial per preparar el conjunt de dades per al modelatge. En aquesta fase, ens hem focalitzat en dues accions principals: la **codificació de variables categòriques** i l'**escalat de variables numèriques**, amb l'objectiu de garantir la compatibilitat de les dades amb els algoritmes d'aprenentatge automàtic: regressió i clustering.

El preprocessament s'ha implementat al fitxer **scripts/preprocess.py**, el qual no conté un número d'execució perquè serà importat en altres scripts del projecte. Aquesta estructura modular ens permet reutilitzar el codi de manera eficient en diverses etapes de l'anàlisi i el modelatge. A continuació, detallem cadascun dels passos seguits en aquesta fase.

Abans de començar, hem exclòs la variable objectiu (target) del dataset perquè les tècniques que aplicarem a continuació poden distorsionar la seva distribució i, per tant, afectar la capacitat del model per aprendre patrons significatius.

A continuació, hem **codificat les variables categòriques**, ja que els models de ML no poden treballar amb aquestes directament. Per tant, hem implementat tres tipus de codificació segons el tipus de característica:

1. **Codificació ordinal:** per a variables categòriques amb un ordre natural (com el nivell educatiu o les opinions sobre l'impacte de la COVID-19). El mètode emprat és *OrdinalEncoder*, el qual assigna valors numèrics consecutius a categories ordinals en funció del seu ordre natural. Doncs, hem especificat els ordres naturals per a cada variable. En l'script es pot observar la jerarquia establerta.
2. **Codificació binària:** per a variables amb dues categories, com ara "sí/no" o "home/dona". El mètode emprat és *el mapeig els valors a 1 i -1*.
3. **Codificació nominal:** per a variables sense ordre natural (com "ciutat" o "tipus de transport"). El mètode emprat és *OneHotEncoder*, el qual crea noves columnes binàries [0,1] per a cada categoria. D'igual forma que les variables binàries, els valors resultants els hem transformat a 1 i -1.

El motiu pel qual hem fet servir cada codificador bé motivat per les característiques dels nostres models i dades, així com per l'objectiu d'obtenir resultats coherents amb les propietats estadístiques del conjunt de dades. Recordem que el nostre dataset segueix una distribució normal.

Pel que respecta l'**OrdinalEncoder** l'hem utilitzat perquè preserva l'ordre de les categories, és simple i eficient. A més, respecta la distribució Gaussiana de les característiques. Pel que fa a la codificació binària, l'hem transformat a **1** i **-1** perquè les dades normalitzades es troben centrades al voltant de zero. Això manté la coherència amb la resta de les variables, ja que després de l'escalat, la mitjana de les dades és zero i la desviació estàndard és d'una unitat.

A més, codificar les dades així manté la simetria, assegurant que les seves propietats estadístiques són consistents amb la resta del conjunt de dades. Cal destacar que, amb aquesta codificació, no és necessari escalar les variables binàries, ja que, com es veu en la

Figura 5, entre 1 i -1 es presenta el 68.3% de les dades, concordant amb el criteri de la distribució normal. Per últim, algoritmes que fan càlcul de distàncies euclidianes es beneficien quan els valors estan centrats i tenen rangs similars, cosa que s'aconsegueix amb aquest mètode.

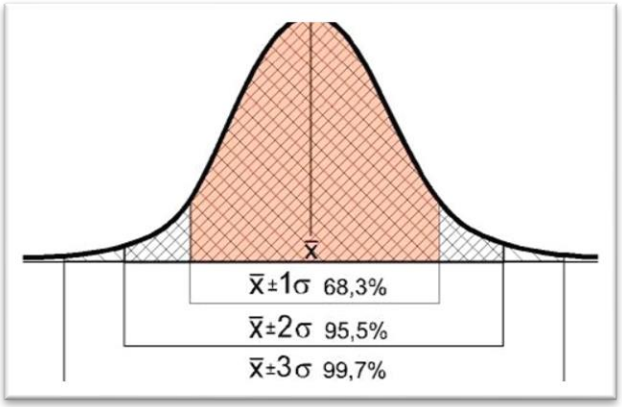


Figura 5. Proporció de les dades en la distribució gaussiana.

Seguidament, hem **escalat les variables numèriques**. Per ajustar-les, hem emprat el mètode **StandardScaler**, el qual aplica l'estandardització (Z-score), que transforma les dades a una distribució amb una mitjana de 0 i una desviació estàndard de 1. Aquesta tècnica és ideal per a dades normalment distribuïdes, ja que manté la forma de la distribució i fa que les dades siguin més comparables entre sí.

Volem ressaltar la importància de l'escalat en el nostre projecte ja que molts models d'aprenentatge automàtic són sensibles a l'escala de les variables. Si aquestes no estan a la mateixa escala, aquelles amb valors més grans poden dominar les decisions del model. Doncs, l'escalat assegura una representació equitativa de totes les variables en el model i optimitza el rendiment grantint resultats fiables.

Finalment, hem afegit la variable target (*estrès*) al dataset i **realitzat una barreja aleatòria de les instàncies del conjunt de dades (shuffle)**. Les raons han sigut diverses: garantir la independència de les mostres, millorar la generalització del model, etc. D'aquesta manera, contribuim a l'entrenament de models més robustos i fiables. **El dataset processat conté 3348 files i 85 columnes.**

En conclusió, després d'haver aplicat els processos de codificació i escalat a les dades, hem aconseguit preparar el dataset adequadament per a les etapes posteriors del projecte. La codificació de les variables categòriques i l'escalat de les dades numèriques han estat passos fonamentals per assegurar que els models puguin operar de manera **eficient** i **precisa**.

Ara, amb la funció *preprocess()* de l'script de preprocessament, podem utilitzar les dades netejades i escalades per entrenar i avaluar models de regressió i clustering, obtinguent així resultats fiables i interpretables per a l'anàlisi de l'impacte de factors ambientals en l'estrès.

4. Arquitectures emprades

En aquest apartat es detallen les diferents arquitectures utilitzades per assolir els objectius del projecte, així com els models seleccionats per a l'anàlisi i predicció. L'elecció de l'arquitectura és un element clau en l'aprenentatge automàtic, ja que influeix en com els algoritmes tracten les dades, les tècniques emprades per a l'entrenament i la precisió de les prediccions generades.

En aquest projecte, hem utilitzat dues arquitectures principals per abordar els objectius plantejats: **regressió** i **clustering**. La regressió s'ha aplicat per predir el nivell d'estrès en funció de les característiques disponibles en el conjunt de dades processat. Aquest enfocament ens ha permès explorar com diferents factors influeixen en l'estrès i generar prediccions específiques per a noves instàncies, proporcionant una visió quantitativa i predictiva del problema.

Per altra banda, s'han implementat tècniques de clustering per analitzar la distribució de les dades a l'espai multidimensional generat per les variables més rellevants identificades en la fase de regressió. Això ha permès identificar grups d'individus amb patrons similars pel que fa als seus nivells d'estrès, facilitant la segmentació i la comprensió més profunda de les relacions entre característiques.

Aquest enfocament combinat ofereix una visió integral del problema. Mentre la regressió proporciona una predicció precisa i contínua del nivell d'estrès, el clustering complementa aquesta anàlisi permetent segmentar la població en perfils més detallats. Això facilita la comprensió dels factors clau i dona suport a la presa de decisions basada en dades, cobrint tant l'anàlisi individual com la general.

4.1. Regressió

La regressió és una de les tècniques més utilitzades en l'aprenentatge automàtic per modelar relacions entre una variable dependent (en aquest cas, el nivell d'estrès) i múltiples variables independents (factors ambientals, demogràfics i d'estil de vida). Aquesta metodologia permet predir valors numèrics continus i alhora identificar les variables més rellevants en el fenomen estudiat.

En el nostre projecte, la regressió ens ha servit per abordar el primer objectiu: predir amb precisió el nivell d'estrès a partir d'un conjunt de factors ambientals i salut mental. A més de ser una eina predictiva, també ens ha ajudat a comprendre millor les dinàmiques que influeixen en l'estrès, com la qualitat de l'aire, les activitats personals o el dia de la setmana. La combinació de models lineals i no lineals ens ha permès obtenir una perspectiva àmplia sobre les relacions existents en el conjunt de dades.

La motivació principal per utilitzar la regressió va sorgir de la necessitat d'abordar dos objectius específics. En primer lloc, construir models predictius que permetessin estimar el nivell d'estrès en funció de múltiples factors. En segon lloc, analitzar quins factors tenien un impacte significatiu sobre l'estrès, obtenint així informació útil per a la presa de decisions.

4.1.1. Models de Regressió

Un cop establits els objectius claus, es van seleccionar cinc models per capturar diferents perspectives i abordar la complexitat de les dades:

1. **XGBoost:** Aquest model basat en boosting destaca per la seva capacitat de manejar relacions no lineals i dades desequilibrades. És eficient en l'entrenament i ofereix un control detallat dels hiperparàmetres, cosa que el converteix en una opció ideal per conjunts de dades amb moltes variables i complexitat.
2. **Gradient Boosting:** Una variant del boosting amb implementació més senzilla, que complementa XGBoost. Aquest model permet validar si resultats similars poden ser obtinguts amb una configuració menys sofisticada.
3. **Random Forest:** Un model d'ensamblatge que combina múltiples arbres de decisió per millorar la robustesa i reduir el sobreajustament. És especialment útil per a conjunts de dades amb característiques redundants o correlacionades.
4. **SVR (Support Vector Regression):** Ofereix la capacitat de modelar relacions no lineals amb un enfocament basat en marges, sent adequat per a dades amb distribucions específiques o patrons no evidents.
5. **Regressió Polinòmica:** Aquesta tècnica estén la regressió lineal afegint termes polinòmics per capturar relacions no lineals senzilles, sent una bona opció per identificar patrons corbats o oscil·lants en les dades.

4.1.3. Sel·lecció de característiques

El procés de selecció de característiques es va realitzar després del preprocessament per identificar quines variables eren més rellevants en la predicció de l'estrès. Aquesta anàlisi va tenir dos objectius: reduir la complexitat del model i millorar-ne la interpretabilitat. Per aconseguir-ho, es van aplicar tècniques específiques segons la naturalesa de cada model.

Aquest procés va tenir com a objectiu identificar quines variables tenien més rellevància en la predicció del nivell d'estrès, permetent així optimitzar l'entrenament i millorar la interpretabilitat dels models. Per dur a terme aquesta anàlisi, es van utilitzar tècniques específiques segons la naturalesa de cada model. En els models basats en arbres, com Random Forest i XGBoost, es va fer ús de l'atribut feature_importances_, que mesura la contribució de cada variable a la reducció de l'impuresa dels nodes.

Per tal d'ajustar el model, es va realitzar una anàlisi de la importància de les característiques, que va permetre identificar quines variables tenen un major impacte en la predicció dels nivells d'estrès. A través de l'atribut feature_importances_ en els models basats en arbres i els coeficients en models lineals, es van destacar les següents variables:

- **dayoftheweek**: Indicant la influència del dia de la setmana en els nivells d'estrès.
- **bienestar**: Una mesura subjectiva de l'estat de benestar.
- **otrofactor i ordenador**: Relacionats amb activitats i factors personals.

- **Factors ambientals**: no2bcn_24h, no2gps_24h i smoke, que reflecteixen l'impacte de la contaminació i l'exposició al fum.

Aquestes variables van ser crucials en tots els models analitzats, reforçant la seva importància en l'estudi de l'estrès.

Per tal d'ajustar el model encara més amb les característiques que millor funcionen per les nostres dades, es va procedir a un procés iteratiu. Es van provar diferents configuracions de models per avaluar quina combinació maximitzava el rendiment i reduïa l'error de predicció.

Per tant, 000un cop realitzat aquest estudi, es va probar d'eliminar variables amb menor importància o correlació per reduir la complexitat del model i evitar problemes de multicolinealitat. Després de moltes proves amb combinacions diferents de les característiques més importants per a cada model, es va concluir que les que millor s'ajustaven eren **dayoftheweek, bienestar**, **otrofactor i ordenador**. Això va permetre centrar l'anàlisi en les característiques més rellevants.

Llavors, si ens fixem cap d'aquestes característiques implica factors ambientals. La següent taula justifica aquestes conclusions. Podem veure que l'error del model és menor, sense les característiques que impliquen factors ambientals. En aquest cas, per simplificar la taula hem posat les mètriques d'un sol model, el XGBoost.

	TRAIN MSE	**TEST MSE**	**MAE**
No factors ambientals	3,38	3,41	1,49
Factors ambientals	2,84	3,40	1,51

Taula 1: Comparativa d'errors de models amb diferents característiques.

4.1.4. Entrenament del Models

Un cop seleccionades les característiques rellevants, es va dividir el conjunt de dades en un 80% per entrenament i un 20% per test. Aquesta divisió va permetre validar els models amb dades no vistes, assegurant que el seu rendiment fos generalitzable.

Per millorar el rendiment dels models i assegurar que es maximitzés la seva capacitat predictiva, es van provar dues tècniques sistemàtiques d'optimització d'hiperparàmetres: **RandomizedSearchCV** i **GridSearchCV**. Aquestes tècniques no només permeten ajustar els paràmetres clau dels models, sinó que també ajuden a explorar diferents configuracions per obtenir resultats més robustos i fiables.

1. **RandomizedSearchCV:** Aquesta tècnica va ser la primera que es va aplicar gràcies a la seva flexibilitat i rapidesa en l'exploració d'un ampli espai de paràmetres. Es van generar combinacions aleatòries d'hiperparàmetres, i es van avaluar fins a 50 configuracions diferents (n_iter=50) per a cada model. A més, es va utilitzar la

validació creuada amb 5 particions (cv=5) per garantir que cada configuració fos avaluada en diverses subdivisions del conjunt d'entrenament. Això va permetre obtenir una mesura més precisa del rendiment general dels models.

Aquesta tècnica va resultar particularment útil per models com **XGBoost** i **Random Forest**, que tenen molts paràmetres ajustables. Per exemple, en el cas de XGBoost, es van optimitzar paràmetres com:

- n_estimators (nombre d'arbres en el model)
- max_depth (profunditat màxima dels arbres)
- learning_rate (taxa d'aprenentatge)
- subsample (proporció de mostres utilitzades per construir cada arbre)

Els millors resultats obtinguts amb RandomizedSearchCV es van utilitzar com a base per a la posterior optimització amb GridSearchCV.

2. **GridSearchCV:** Per refinar encara més els resultats obtinguts amb RandomizedSearchCV, es va aplicar GridSearchCV. Aquesta tècnica realitza una cerca exhaustiva i sistemàtica de totes les combinacions possibles dins d'un espai acotat de paràmetres. Tot i que és més lenta que RandomizedSearchCV, ofereix una exploració completa de l'espai de solucions, assegurant que no es passen per alt configuracions òptimes.

   No obstant això, es va observar que els temps de càlcul augmentaven significativament, especialment en models amb molts hiperparàmetres i conjunts de dades grans. A més, els resultats finals obtinguts amb GridSearchCV no van mostrar millores significatives respecte als de RandomizedSearchCV, indicant que l'espai d'hiperparàmetres ja havia estat ben explorat inicialment. Per tant, RandomizedSearchCV es va establir com la tècnica principal d'optimització en el projecte.

Aquest doble enfocament sistemàtic va assegurar que cada model fos ajustat de manera òptima, maximitzant la seva capacitat predictiva alhora que es reduïen riscos de sobreajustament (overfitting). Les tècniques basades en validació creuada també van permetre avaluar el rendiment dels models de manera robusta, garantint que les mètriques reportades (com l'error quadràtic mitjà -**MSE** - i l'error absolut mitjà - **MAE**) fossin fiables i representatives

4.1.5. Problemes trobats

Durant el procés d'entrenament i avaluació dels models, es van identificar diversos problemes que van requerir ajustos en la implementació per millorar el rendiment i la interpretabilitat dels resultats.

El primer problema observat va ser l'**overfitting** en alguns models, especialment en aquells amb una gran capacitat de complexitat, com **Random Forest**, **Gradient Boosting** i **Polynomial Regression**. Aquests models mostraven una discrepància significativa entre

les mètriques d'entrenament i les de test, amb un error MSE molt baix en entrenament però considerablement més alt en test (com es pot observar en la figura). Aquesta diferència indicava que els models s'ajustaven massa a les dades d'entrenament, capturant soroll en lloc de patrons generalitzables.

Per mitigar aquest problema, es van implementar diverses estratègies:

- **Regularització:** En models com Gradient Boosting i Polynomial Regression, es van aplicar tècniques de regularització per penalitzar la complexitat dels models, reduint així l'ajust excessiu als punts específics de l'entrenament.
- **Pruning en Random Forest:** Es van limitar la profunditat dels arbres i el nombre de fulles per controlar la complexitat del model.
- **Cross-validation:** L'ús de validació creuada va ajudar a detectar ràpidament casos d'overfitting, assegurant que els resultats fossin consistents en diferents subconjunts de dades.

Aquestes mesures van aconseguir reduir parcialment el problema d'overfitting, millorant les mètriques de test sense sacrificar massa la precisió en entrenament.

Podem veure les mètriques dels models amb overfitting a la primer gràfica, on es nota claramenr que els models memoritzen. Això ho veiem ja que l'error de train és molt més alt que el de test, cosa que indica que el model s'adapta a les dades d'entrenament i no busca els patrons adeqüats.

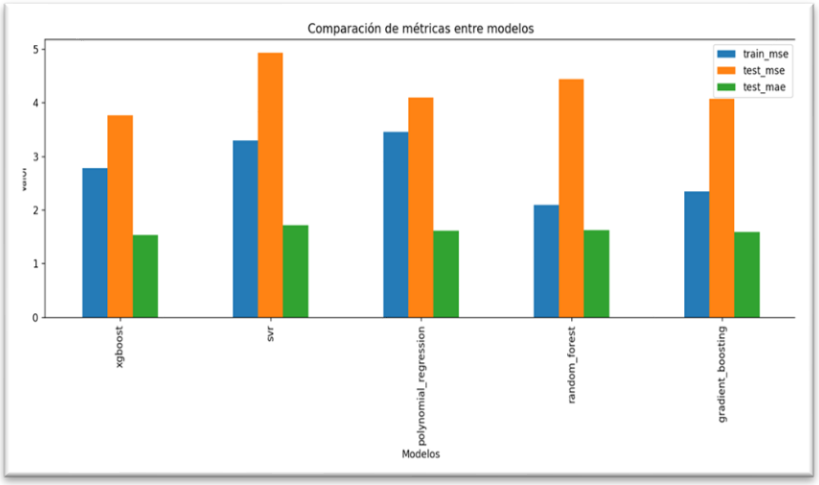
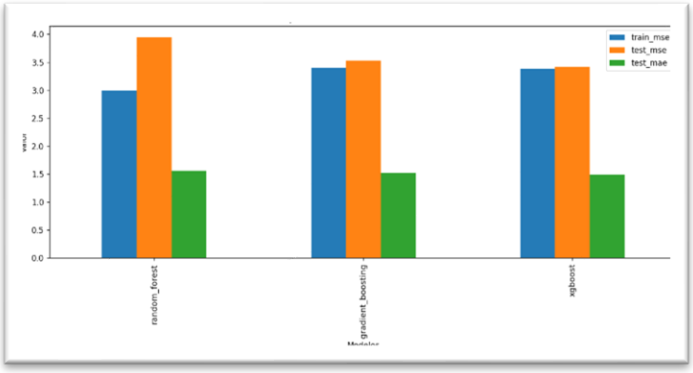
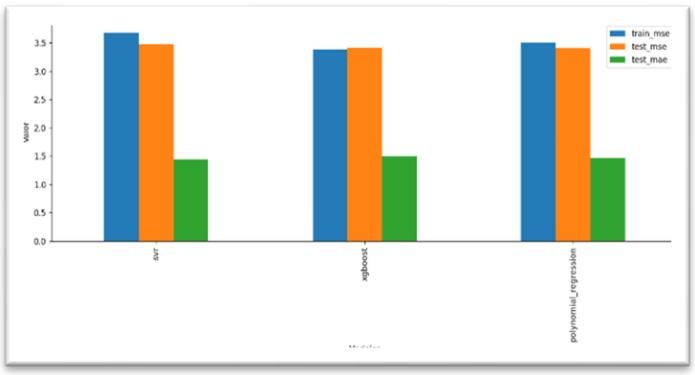


Figura 6: Comparació de mètriques dels models amb Overfitting.

En les següents gràfiques, es pot veure que, després d'aplicar aquestes millores, els errors de **train** i **test** són molt més propers, la qual cosa indica que els models ja no estan memoritzant les dades d'entrenament. Els models mostren un rendiment molt més equilibrat, cosa que significa que han après patrons generalitzables.

En particular, models com **Gradient Boosting** i **XGBoost** mostren una clara reducció de la discrepància entre els errors de train i test. Això confirma que aquests models s'adapten millor a les dades sense sacrificar la seva capacitat de generalització. A més, els errors MAE i MSE es mantenen constants entre els dos conjunts, reflectint un comportament robust.





Figures 7 i 8: Comparació de mètriques dels models sense Overfitting.

El segon problema identificat va ser un **desequilibri significatiu** en les classes extremes dels nivells d'estrès. Aquest desequilibri va provocar que els models tendissin a predir amb més precisió els nivells d'estrès més comuns, mentre que les prediccions per a nivells extrems eren menys fiables. Aquest biaix va afectar negativament la capacitat del model per generalitzar.

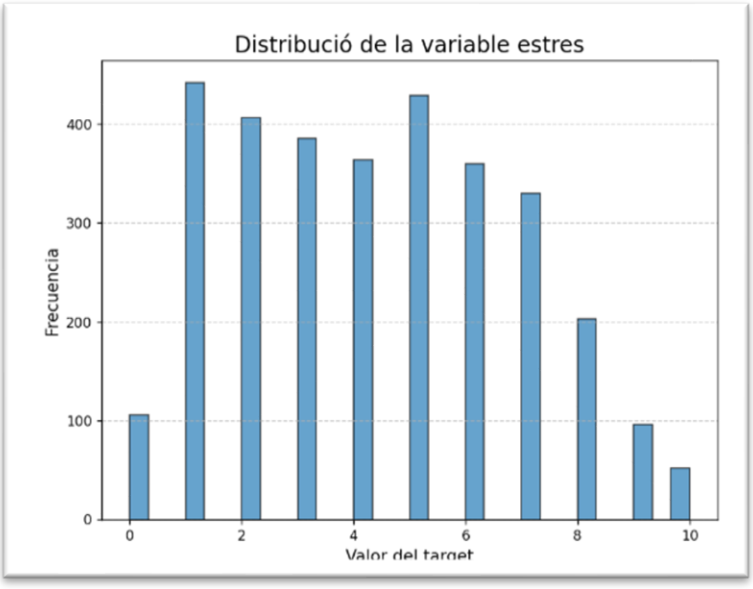


Figura 9: Distribució de la variable estrès.

Per tant vam voler comprovar que aquest desbalanceig estaba causant problemes a l'hora de predir les classes minoritàries.

Per això vam visualitzar la següent gràfica. Aquestarepresenta un diagrama de dispersió en el qual es comparen els valors reals de l'estrès amb les prediccions generades pel model per a classes minoritàries (0, 9, i 10). La línia diagonal representa una predicció ideal, és a dir, on els valors predits serien exactament iguals als valors reals.

Es pot observar que per a les classes minoritàries, especialment la classe 10, les prediccions tendeixen a estar significativament desalineades de la línia ideal. Això reflecteix la dificultat del model per capturar patrons adequats per a aquestes classes

menys representades. En contrast, les classes amb més dades (com les intermèdies) es mostren més properes a la línia ideal, indicant una millor precisió en les prediccions.

Aquest tipus d'anàlisi és crucial per identificar problemes específics en les prediccions i ajustar les estratègies, com per exemple, millorar la qualitat de les dades d'entrenament o optimitzar els hiperparàmetres dels models, per tal d'obtenir un rendiment més equilibrat entre totes les classes.

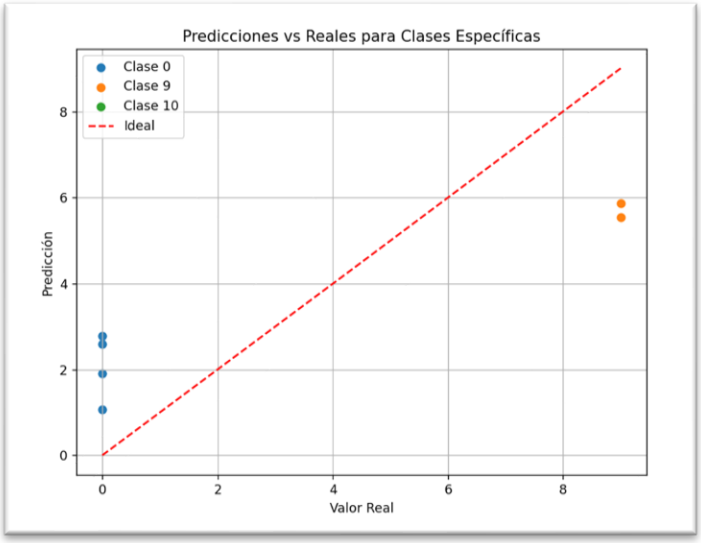


Figura 10: Prediccions vs Valors Reals de classes minoritàries.

Per abordar aquest problema, es van explorar dues solucions:

La tècnica **de Synthetic Minority Oversampling Technique (SMOTE**) es va utilitzar com una primera aproximació per solucionar aquest problema. SMOTE genera mostres sintètiques per a les classes minoritàries, augmentant així la seva representació en el conjunt d'entrenament. Aquest enfocament busca equilibrar la distribució de dades, permetent que els models tinguin més informació sobre les classes menys representades.

Tot i això, els resultats obtinguts amb SMOTE van ser decebedors. Els models no van mostrar una millora significativa en les mètriques després d'aplicar aquesta tècnica. L'error MAE (Mean Absolute Error) es va mantenir pràcticament igual abans i després de l'aplicació de SMOTE. Aquest resultat suggereix que, tot i tenir més mostres de les classes extremes, els models continuaven tenint dificultats per aprendre patrons significatius en aquestes classes. A més, l'ús de mostres sintètiques podria haver introduït soroll en el conjunt d'entrenament, afectant la capacitat general del model per capturar patrons reals.

Davant la limitada efectivitat de SMOTE, es va optar per una estratègia alternativa basada en **l'agrupació de classes**. Aquesta tècnica consisteix a agrupar les classes amb nivells d'estrès similars, especialment aquelles situades en els extrems, per reduir la complexitat del problema i equilibrar millor la distribució. Per exemple, les classes 9 i 10 es van agrupar en una sola classe, i el mateix es va fer amb les classes més baixes (com la 0 i la 1).

Els resultats d'aquesta aproximació van ser més prometedors. L'error MAE es va reduir lleugerament, passant del 15,1% al 14,9%, cosa que indica que l'agrupació de classes va ajudar els models a predir amb una mica més de precisió. Tot i que la millora en les

mètriques no va ser molt significativa, aquesta estratègia va demostrar ser més efectiva que SMOTE per al conjunt de dades utilitzat. A més, l'agrupació de classes va simplificar la tasca de predicció, permetent als models centrar-se en patrons més generals i menys sorollosos.

4.1.5. Avaluació dels Models

Per avaluar el rendiment dels models desenvolupats, s'han utilitzat dues mètriques clau: el **Mean Squared Error (MSE)** i el **Mean Absolute Error (MAE)**. Aquestes mètriques proporcionen informació complementària sobre la capacitat dels models per predir de manera precisa els nivells d'estrès.

El **MSE** calcula l'error mitjà al quadrat entre els valors predits pel model i els valors reals. Aquest indicador penalitza fortament els errors grans gràcies al seu component quadràtic, cosa que el fa especialment útil per identificar models que cometen errors significatius en algunes prediccions. No obstant això, el **MSE** és sensible a valors extrems, ja que amplifica errors grans de manera desproporcionada.

El **MAE** calcula l'error absolut mitjà entre els valors reals i les prediccions. A diferència del MSE, el MAE no penalitza tant els errors grans i proporciona una mesura més interpretativa en termes de l'error mitjà absolut en la mateixa unitat que les prediccions (en aquest cas, nivells d'estrès). Aquesta mètrica és menys sensible a valors extrems i és ideal per avaluar models amb errors més uniformes.

La combinació d'aquestes dues mètriques permet una avaluació completa dels models.

A continuació,2 podem veure les mètriques finals que hem obtinut dels models. A les dues figures veiem les mètriques del model XGboost ja que més en davant parlarem d'ell.

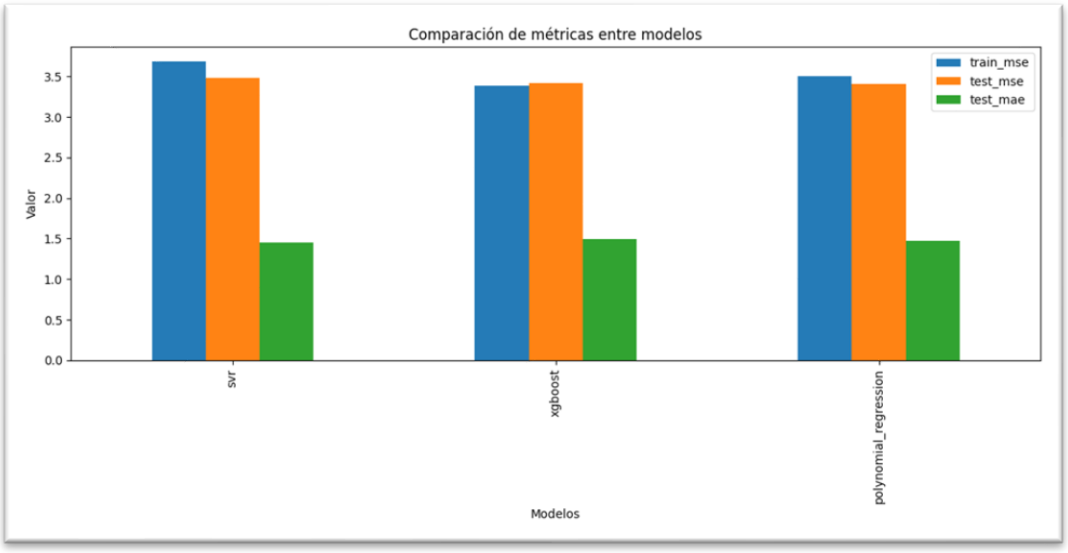


Figura 11: Comparativa de mètriques entre models

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

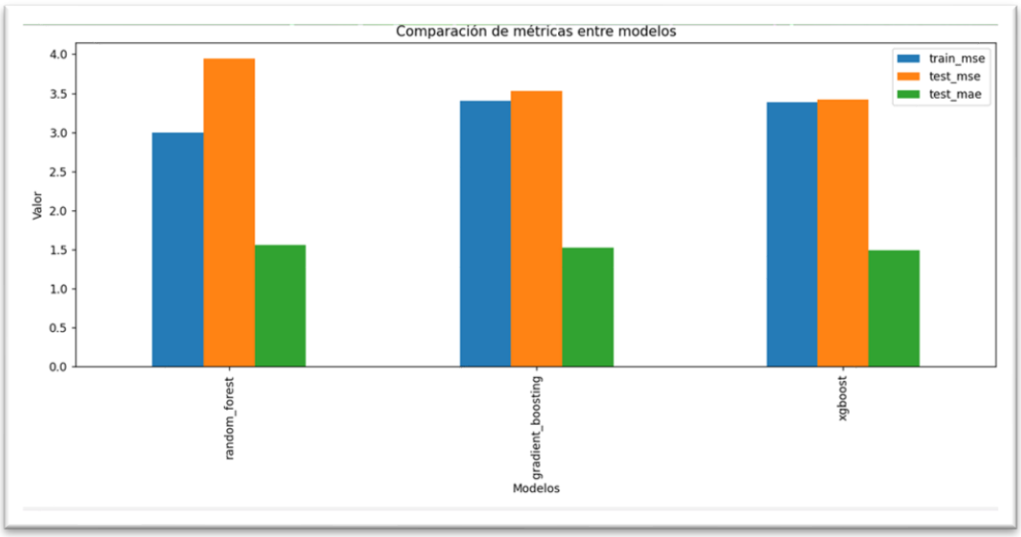


Figura 12: Comparativa de mètriques entre models

En la gràfica presentada, es comparen les mètriques d'error **MSE** i **MAE** per als conjunts de dades d'entrenament i test en diferents models: **XGBoost**, **SVR**, **Polynomial Regression**, **Random Forest** i **Gradient Boosting**.

A continuació podem veure els valors de les mètriques per entendre millor i poder interpretar-los.

	Random Forest	**Gradient Boosting**	**XGBoost**	**SVR**	**Polynomial**
MSE_train	3,00	3,40	3,38	3,68	3,50
MSE_test	3,94	3,53	3,41	3,47	3.41
Diferència train-test	0,94	0,13	**0,02**	0,21	0,09
MAE	1,55	1,52	1,49	1,45	1,47

Taula 2: Mètriques numèriques dels models

Per tant, la taula presentada resumeix les mètriques d'avaluació dels models analitzats en termes de **MSE (Mean Squared Error)**, **diferència entre MSE d'entrenament i test** i **MAE (Mean Absolute Error)**. Aquestes mètriques ens permeten comparar els models segons la seva capacitat de predicció, generalització i robustesa.

Random Forest té un rendiment equilibrat, però presenta una diferència considerable (0.94) entre les mètriques de train i test. Això suggereix una certa tendència a l'overfitting, ja que el model és més precís en dades d'entrenament que en dades noves.

Gradient Boosting mostra una excel·lent generalització amb una **diferència molt baixa (0.13)** entre train i test, indicant que el model no està sobreajustat. Aquest és un dels models amb millor comportament global, amb errors baixos i consistents.

XGBoost és el model amb la **menor diferència entre train i test (0.02)**, cosa que indica una gran capacitat de generalització. També té el MAE més baix (1.49), mostrant una precisió destacada en termes absoluts. Aquestes característiques fan que sigui el millor model per al conjunt de dades.

SVR mostra un comportament adequat amb un **MAE molt competitiu (1.45)**. Tot i això, la diferència entre train i test (0.21) és més alta que en els millors models, indicant que la seva capacitat de generalització podria millorar.

Polynomial Regression presenta una diferència moderada entre train i test (0.09), cosa que indica que el model generalitza bé, però el seu MAE no és tan competitiu com els millors models.

La taula evidencia que **XGBoost** és el model més adequat per a aquest conjunt de dades, gràcies a la seva capacitat de mantenir una precisió consistent entre entrenament i test i els errors més baixos. Gradient Boosting és una alternativa fiable i robusta, mentre que altres models, tot i ser competitius, presenten més limitacions en termes de generalització. Per tant, el nostre model final presenta un 14,9 % de MAE amb les característiques **dayoftheweek, bienestar**, **otrofactor i ordenador.**

4.1.6. Resultats i anàlisi

Després de completar l'entrenament i l'avaluació dels models seleccionats, hem obtingut resultats que ens permeten extreure conclusions rellevants sobre la capacitat predictiva dels models i la seva relació amb les variables més importants en l'estrès. Aquest apartat analitza els resultats, posant especial èmfasi en l'acceptació o rebuig de les hipòtesis plantejades inicialment.

Com s'ha analitzat anteriorment, XGBoost ha estat el model amb un millor rendiment global, destacant per:

- **Diferència mínima entre train i test (0.02):** Això demostra que el model té una excel·lent capacitat de generalització i no pateix problemes d'overfitting.
- **MAE més baix (1.49):** Indica que les prediccions són molt properes als valors reals en termes absoluts, la qual cosa fa que aquest model sigui el més precís.
- **MSE_test consistent (3.41):** Reflexix que l'error mitjà quadràtic també és dels més baixos, demostrant estabilitat en les prediccions.

En comparació:

- **Gradient Boosting** també ha obtingut bons resultats, amb una diferència petita entre train i test (0.13) i un MAE competitiu (1.52), però lleugerament inferior a XGBoost.
- **Random Forest** ha mostrat signes d'overfitting amb una diferència significativa entre train i test (0.94), cosa que limita la seva utilitat per a generalització.
- **Polynomial Regression i SVR** han estat adequats, però no tan robustos com XGBoost o Gradient Boosting.

A més, els resultats han confirmat que les característiques **dayoftheweek**, **bienestar**, **otrofactor** i **ordenador** són les més rellevants per predir l'estrès. Això s'ha determinat després d'una anàlisi iterativa i la comprovació de l'impacte d'incloure o excloure altres variables.

Tot i que inicialment els **factors ambientals** es consideraven rellevants, la seva inclusió no ha millorat els resultats del model, tal com es reflecteix a la **Taula 1** (on l'error del model augmenta lleugerament quan es consideren aquests factors).

Les Variables personals i d'estil de vida han estat determinants en la predicció de l'estrès, destacant la influència del benestar i el dia de la setmana.

Aquestes conclusions indiquen que l'estrès està més condicionat per factors subjectius i d'estil de vida que no pas per variables estrictament ambientals, almenys en el context d'aquest estudi.

4.1.7 Acceptació/Rebuig de les Hipòtesi Plantejades

Les hipòtesis inicials es basaven en la relació entre l'estrès i variables com la qualitat de l'aire, el benestar i el dia de la setmana. Aquí detallem la seva acceptació o rebuig:

1. **Hipòtesi 1:** *L'exposició a nivells més alts de NO_2 durant 24 hores està associada a majors nivells d'estrès.*
    - **Rebuig:** Tot i que s'ha detectat una correlació positiva moderada entre NO_2 i l'estrès, aquesta variable no ha tingut un impacte significatiu en els models de predicció. Això indica que el NO_2 podria influir en l'estrès, però el seu efecte no és prou fort en comparació amb altres variables.
2. **Hipòtesi 2:** *A mesura que s'acosta el cap de setmana, l'estrès disminueix.*
    - **Acceptació:** S'ha confirmat que **dayoftheweek** és una de les variables més rellevants per a la predicció de l'estrès. Aquesta troballa és coherent amb la correlació negativa observada entre el dia de la setmana i l'estrès, ja que les persones tendeixen a sentir-se menys estrèssades a mesura que s'apropa el cap de setmana.
3. **Hipòtesi 3:** *Contra menys benestar experimenti una persona, més estrèssada es sentirà.*
    - **Acceptació:** La variable **bienestar** ha estat consistentment identificada com una de les més importants en la predicció de l'estrès, corroborant aquesta hipòtesi. La correlació negativa moderada entre benestar i estrès també valida aquesta relació.

L'anàlisi ha permès validar parcialment les hipòtesis inicials, posant de manifest que l'estrès és una variable altament influenciada per factors personals i d'estil de vida. Tot i que els factors ambientals poden tenir un cert impacte, no són determinants en el conjunt de dades utilitzat.

Finalment, l'ús de models avançats com XGBoost ha permès obtenir prediccions robustes i fiables, mentre que l'anàlisi de característiques ha proporcionat una millor comprensió

dels factors més influents en l'estrès. Aquest projecte ha complert amb els objectius plantejats i ofereix una base sòlida per a estudis futurs que vulguin aprofundir en la relació entre factors ambientals i salut mental.

4.2. Clustering

En el context del nostre projecte, l'objectiu d'aquest apartat és observar si les característiques més importants que influencien l'estrès formen clústers, és a dir, analitzar si hi ha una separació de les dades en l'espai.

Recordem que el **clustering** és una tècnica d'aprenentatge no supervisat que ens permet segmentar les dades en diversos grups (o clústers), on els individus dins d'un mateix grup són més semblants entre si que amb els membres de grups diferents. Doncs, aquesta tècnica ens ajudarà a comprendre millor les relacions subjacents entre les característiques.

4.2.1. Sel·lecció de variables d'entrada

Les variables d'entrada del nostre model, han canviat en funció del propòsit de cerca. A continuació llistem les diferents implementacions que hem realitzat, amb les variables d'entrada respectives:

1. **Clustering per verificar patrons addicionals:** hem emprat el **dataset complet**, amb les 85 característiques disponibles, per explorar possibles agrupacions naturals en les dades que podrien revelar patrons latents rellevants no detectats en la regressió.
2. **Clustering per verificar la separabilitat de les dades segons el model regressor:**
    a. **Característiques importants generals dels regressors:** aquelles variables que tots els models regressors han considerat rellevants en la predicció de l'estrès. Aquestes són: VARIABLES RELLEVANTS GENERALS = ['dayoftheweek', 'bienestar', 'energia', 'ordenador', 'alcohol', 'otrofactor', 'no2bcn_24h', 'no2gps_24h', 'covid_work']. El motiu de selecció és perquè podem identificar patrons comuns que siguin consistents en tots els models, assegurant-nos que estem utilitzant variables robustes i no dependents d'un únic model.

    b. **Característiques importants del model XGBoost:** aquelles variables que el model regresor XGBoost ha considerat rellevants en la predicció de l'estrès. Aquestes són: VARIABLES RELLEVANTS XGBOOST = ['ordenador', 'otrofactor', 'dayoftheweek', 'district_gràcia', 'incidence_cat_physical incidence', 'smoke', 'district_sant andreu', 'bienestar', 'Totaltime'].

    El motiu de selecció és per basar-nos en les variables que han demostrat tenir més impacte segons el model XGBoost, que és el millor model generat. D'aquesta manera, corroborem si en la regressió amb menor error hi ha una separació de les dades en l'espai. A més, podem comprar les distàncies entre clústsers respecte les variables generals dels regressors.

c. **4 característiques més importants del model XGBoost:** aquelles quatre variables que el model regresor XGBoost ha considerat més rellevants en la predicció de l'estrès. Aquestes són: VARIABLES RELLEVANTS SIMPLIFICADES = ['ordenador', 'otrofactor', 'dayoftheweek', 'bienestar']. El motiu de sel·lecció és per simplificar el model utilitzant només les variables amb més pes en les prediccions.

   Reduir el nombre de característiques en el clustering permet augmentar la claredat i la interpretació dels grups, ja que aquests es basen exclusivament en els factors més significatives. Aquesta simplificació també millora la robustesa i l'eficiència del model, ja que s'evita el soroll i la complexitat innecessària. Tot i això, és fonamental garantir que les característiques seleccionades capturin tota la informació rellevant per evitar perdre matisos importants o dades significatives.

   Doncs, d'aquesta manera, pretenem analitzar si hi ha variables que introdueixen soroll al model i evaluar com evoluciona l'eficiència del clustering.

4.2.2. Preprocessament de les dades

Per processar les dades, hem emprat la funció explicada en l'apartat 3.4. Aquesta ens ha permès escalar i codificar el dataset, **excloent la variable objectiu**, que és l'estrès.

D'igual forma que en la regressió, incloure aquesta **variable pot comportar una pèrdua d'informació**, ja que el model tindria accés a unes dades que no podria consultar en la vida real, afectant negativament a la validesa del resultat. D'altra banda, incloure l'etiqueta estrès **pot sesgar el procés d'agrupament** perquè el model intenta agrupar les dades en funció d'aquesta, en comptes de trobar patrons inherents en les característiques de les dades.

4.2.3. Sel·lecció de l'algoritme de clustering

Per tal d' obtenir una comprensió més profunda de les estructures i patrons presents en les dades d'estrès, hem emprat diferents mètodes d'agrupament:

1. **K-means:** és un algoritme d'aprenentatge no supervisat que divideix un conjunt de dades en un nombre predefinit de clústers (k) mitjançant la minimització de la variància dins de cada grup. Hem emprat el model K-Means per observar si els clústers tenen una mida esfèrica i similar. Doncs, funciona millor quan els grups estan bastant separats.
2. **Agglomeratiu:** aquest mètode jeràrquic comença considerant cada individu com un clúster separat i, successivament, fusiona els clústers més propers fins a formar un únic grup. Hem emprat el model Agglomeratiu ja que pot capturar estructures més complexes, sense una forma definida.
3. **Gaussian Mixture:** aquest model probabilístic assumeix que les dades són una combinació de diverses distribucions normals (Gaussians). Hem emprat el model de Gaussian Mixture per identificar clústers elíptics i amb diferents mides,

permetent una modelització més flexible de les complexitats en les dades de salut mental.

4.2.4. Determinació del nombre de clústers

Per determinar el nombre òptim de clústers en els mètodes utilitzats (K-means, Agglomerative i Gaussian Mixture), s'han aplicat diferents tècniques d'avaluació que asseguren una segmentació significativa i coherent amb les dades. A continuació es descriuen els criteris utilitzats per cadascun:

Aquí tens les justificacions per cadascun dels mètodes emprats:

1. **Mètode del "Elbow" (K-means)**: aquest mètode s'ha emprat perquè és una tècnica intuïtiva i senzilla per determinar el nombre òptim de clústers. El càlcul de la suma dels errors quadràtics (SSE) permet observar com disminueix l'error quan augmentem el nombre de clústers. La k òptima es troba en el punt on la disminució de SSE comença a ser menys pronunciada, formant un "colze". Aquesta característica indica que afegir més clústers no aporta una millora significativa en la cohesió intra-clúster, però sí que afegeix complexitat al model. Per tant, aquest mètode busca un equilibri entre una bona separació dels clústers i evitar un sobreajustament.
2. **Mètode de l'Índex de Silueta (Agglomerative Clustering)**: en el cas de l'algoritme Agglomerative, l'índex de silueta es considera un criteri adequat perquè mesura de manera efectiva tant la cohesió dins de cada clúster com la separació entre els clústers. La k òptima es determina maximitzant aquest índex, assegurant així que els punts dins de cada clúster siguin molt semblants entre si (alta cohesió), mentre que els clústers estiguin ben separats (alta separabilitat). Aquest enfocament és útil per garantir que el nombre de clústers seleccionat millori la qualitat del model, amb un resultat clar i ben definit.
3. **Criteri de Versemblança (Gaussian Mixture)**: en el model Gaussian Mixture, s'ha utilitzat el Bayesian Information Criterion (BIC) perquè és un criteri que no només avalua la qualitat de la versemblança, sinó que també penalitza la complexitat del model. Això permet evitar el sobreajustament, on l'ús de massa paràmetres podria millorar excessivament l'ajust als dades d'entrenament però reduir la generalització del model. El BIC minimitza aquest compromís, ajudant a trobar el nombre òptim de components (k) que aconsegueix un model bo i senzill alhora. Per tant, el BIC és adequat per garantir un model precís sense complexitat innecessària.

Aquest enfocament combinat ha permès determinar la *k* òptima de manera rigorosa per a cada mètode de clustering. Les tècniques emprades estan adaptades a les característiques específiques de cada algoritme, assegurant una segmentació robusta i amb un equilibri entre cohesió i separació dels clústers. Aquesta diversitat de mètodes també aporta confiança en la validesa dels resultats, independentment del model utilitzat.

4.2.5. Visualització dels resultats

En aquest apartat es presenten les tècniques i eines utilitzades per visualitzar els resultats del procés de clustering. La visualització és una etapa clau, ja que permet comprendre millor els patrons descoberts pels algorismes, identificar la distribució dels grups, analitzar les seves característiques principals i validar la coherència dels clústers generats.

En aquest projecte, s'ha utilitzat la tècnica **t-SNE (T-distributed Stochastic Neighbor Embedding)** per a la projecció de les dades en dimensions reduïdes. Aquest és un algorisme no lineal de reducció de dimensionalitat, especialment dissenyat per visualitzar dades en espais bidimensionals o tridimensionals. T-SNE prioritza la conservació de la relació local entre els punts de dades, mostrant de manera clara els patrons d'agrupament que poden no ser evidents en altres tècniques.

No hem emprat **PCA (Anàlisi de Components Principals)** perquè, tot i ser una tècnica potent de reducció de dimensionalitat, la seva metodologia es basa en mantenir la màxima variància global de les dades. Això pot ser limitant en contextos com el nostre, on els patrons d'agrupament són subtils i poden no estar associats amb les dimensions de major variància. A més, PCA assumeix relacions lineals entre les variables, cosa que pot no captar la complexitat de les dades utilitzades en aquest estudi.

Per aquestes raons, hem optat pel t-SNE, que ofereix una millor capacitat per representar patrons no lineals i per destacar agrupacions localitzades, aspectes crucials per a una anàlisi visual efectiva dels resultats del clustering.

Al llarg de l'apartat, **es mostren les distribucions de l'estrès en els clústers generats** i la interpetació per cada mètode d'agrupament. En les distribucions, cada barra representa un clúster i mostra la proporció de les diferents classes d'estrès dins del clúster. Les distribucions indiquen si hi ha dominància d'una classe específica en algun clúster. En cas que el nivell d'estrès sigui heterogeni entre grups, en l'apartat 5, estudiarem les característiques predominants en cada grup, extraient un perfil d'estrès. En els següents subapartats, **només mostrarem les visualitzacions més significatives**. Per a més informació, consultar l'Annex 6.

Volem destacar que el nostre mètode de validació ha estat principalment **visual**, ja que el t-SNE permet una comprensió intuïtiva i directa de com es distribueixen els clústers en l'espai reduït. Aquesta elecció està motivada per la naturalesa dels nostres objectius: comprendre i comunicar els patrons subjacents de manera accessible i clara.

Tot i això, cal tenir en compte que el t-SNE redueix la dimensionalitat original de les dades, en aquest cas **85 dimensions**, a només 3 dimensions per a la seva representació visual. Aquesta reducció pot provocar que, en alguns casos, els clústers no es percebin clarament separats en la visualització. És per aquest motiu que el nostre focus principal és l'anàlisi de les distribucions de l'estrès.

No obstant això, això no significa necessàriament que els clústers no estiguin ben definits en dimensions més altes, on la informació original és més completa. Per tant, encara que en l'espai reduït els clústers puguin semblar solapats, és possible que en altres dimensions de l'espai original sí que estiguin clarament separats. Aquesta limitació és inherent a qualsevol tècnica de reducció de dimensionalitat, però el t-SNE continua sent una eina potent per identificar patrons complexos i entendre la distribució general de les dades.

Destaquem que **no hem emprat mètriques** com el coeficient de silueta, que, tot i ser útil per avaluar la cohesió i separació dels clústers, presenta certes limitacions en implementacions com la nostra. Aquestes mètriques no solen captar completament les relacions no lineals que el t-SNE és capaç de visualitzar de manera explícita.

Així, la validació visual a través del t-SNE s'ha considerat la millor opció per ressaltar la naturalesa dels clústers i facilitar la interpretació dels resultats obtinguts.

4.2.6.1. Clustering per verificar patrons addicionals: dataset complet.

Per tal de descobrir noves característiques rellevants no identificades en la regressió, hem fet clustering amb el dataset complet (85 característiques). Tot i que l'anàlisi de regressió se centra en identificar variables específiques que expliquen la variable objectiu *estrès*, el clustering ofereix una perspectiva complementària, agrupant les dades basant-se en similituds globals, la qual pot pot oferir informació valuosa per refinar la comprensió dels factors que influeixen en l'estrès de la població de BCN.

Som conscients però, que amb un conjunt tan gran de característiques, la reducció de dimensionalitat necessària per visualitzar els resultats pot ocultar detalls subtils de la separació dels clústers en dimensions més altes. Tot i així, l'ús del dataset complet assegura que l'anàlisi cobreix tots els possibles patrons i agrupacions latents, proporcionant una perspectiva àmplia i robusta. A més, visualitzarem les distribucions de l'estrès per cada clúster, concloent si hi ha informació valuosa o no.

Després de fer clustering amb el dataset processat, hem obtingut diversos grups segons els 3 algoritmes. A continuació, hem **analitzat les distribucions de la variable objectiu (estrès) dins de cada clúster** per determinar si els grups trobats tenen alguna correlació significativa amb els nivells d'estrès. Les distribucions dins dels clústers poden aportar informació valuosa, com per exemple, identificar quins grups tenen una prevalença més alta de nivells d'estrès elevats o, per contra, quins clústers mostren valors més baixos d'estrès. Els resultats s'exposen en les Figures 13, 14 i 15.

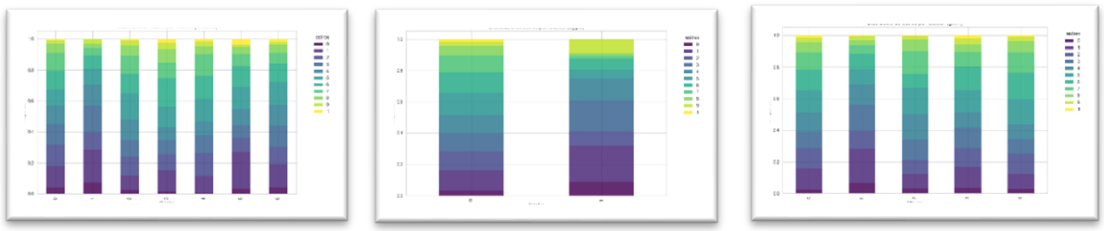


Figura 13, 14 i 15. Distribucions d'estrès en clustering dataset complet (kmeans, agglo i gmm). Directori: visualizations/clusters/dataset.Arxius: kmeans_k6_distribution.png, agglo_k2_ distribution.png i gmm_10 _ distribution.png, respectivament.

Les gràfiques mostrades suggereixen que les distribucions de l'estrès dins dels clústers són força homogènies, sense mostrar agrupacions distintes o patrons clars. Això indica que, tot i els esforços per identificar grups diferenciats mitjançant clustering, **no s'han detectat diferències significatives en els nivells d'estrès entre els clústers**. Aquesta falta de separació pot suggerir que les característiques utilitzades per a la segmentació no són suficients per discriminar grups amb nivells d'estrès clarament diferents. A més, si

analitzem la representació gràfica, observem com no hi ha separació dels clústers en l'espai.

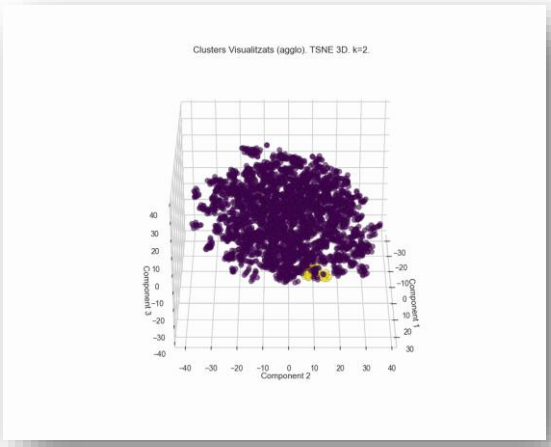


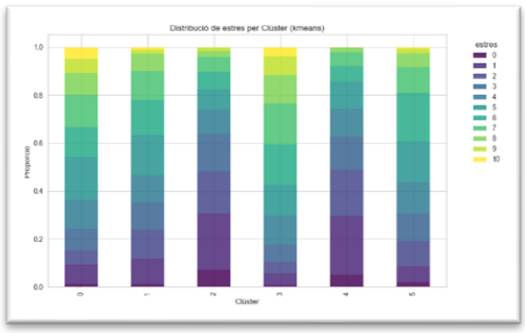
Figura 16. Resultat del clustering amb dataset complet. Directori: visualizations/clusters/dataset. Arxius: agglo_k2_TSNE3d_animated.gif.

Aquesta constatació posa de manifest que el clustering no sempre revela agrupacions útils per a la comprensió de l'estrès, almenys no amb el conjunt de dades actual. Per tant, **hem reduït la dimensionalitat del conjunt de dades, buscant reduir el soroll i millorar la interpretació dels resultats.** Destaquem que, la reducció de dimensionalitat permet focalitzar-se en les característiques més rellevants, eliminant variables que poden no ser informatives i facilitant la visualització de possibles patrons de manera més clara. Doncs, volem comprobar si aquesta aproximació proporciona un enfocament més eficient per comprendre les relacions entre les variables i les agrupacions potencials.

4.2.6.2. Clustering per verificar la separabilitat de les dades: regressors

Per tal de verificar si les característiques més importants en l'estrès segons els regressors són coherents amb les agrupacions naturals en les dades, hem realitzat un procés de clustering utilitzant aquestes mateixes variables. El propòsit d'aquest anàlisi és determinar si les característiques identificades pel regressors poden ser associades a clústers clars en el conjunt de dades. En altres paraules, si els regressors han identificat patrons subjacents rellevants en les dades, el clustering hauria de reflectir aquests mateixos patrons, mostrant una bona separabilitat entre els clústers. Alhora, com hem explicat, volem comprovar si una reducció en la dimensionalitat de les dades d'entrada comporta un major resultat.

Després de fer clustering amb les 9 característiques més importants segons els regressors, hem obtingut diversos grups per cada algoritme d'agrupament. A continuació, hem **analitzat les distribucions de la variable objectiu (estrès) dins de cada clúster** per determinar si els grups trobats tenen alguna correlació significativa amb els nivells d'estrès. En les Figures 17, 18 i 19 podem observar la distribució en cada mètode d'agrupament.



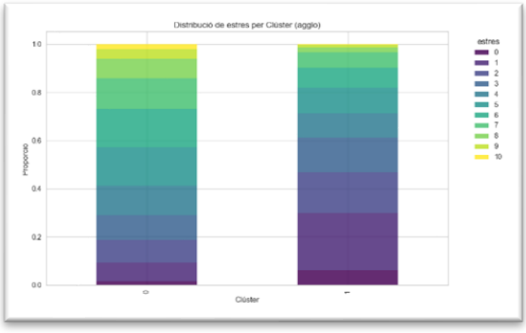


Figura 17 i 18. Distribucions d'estrès en clustering dels regressors (kmeans, agglo). Directori: visualizations/clusters/general_important_features. Arxius: kmeans_k6_distribution.png i agglo_k2_distribution.png.

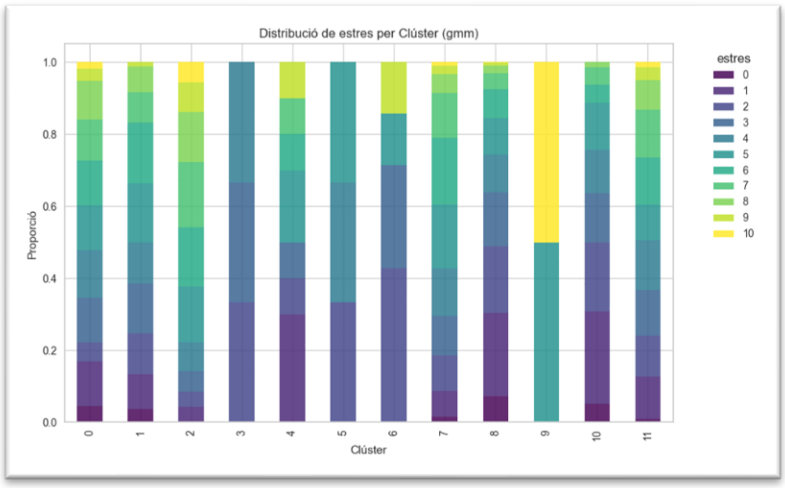


Figura 19. Distribucions d'estrès en clustering dels regressors (gmm). Directori: visualizations/clusters/general_important_features. Arxius: gmm_k12_distribution.png.

Si analitzem els resultats:

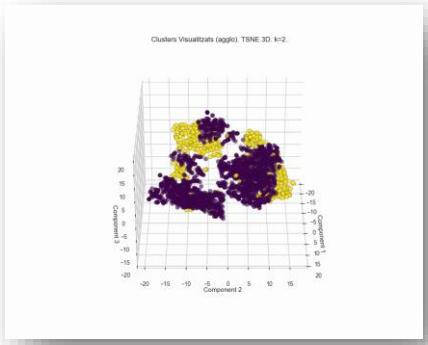
1. **K-Means (kmeans)**: el mètode K-means genera clústers amb distribucions relativament uniformes, amb poca variabilitat en els nivells d'estrès entre els clústers. Això suggereix una capacitat limitada per capturar patrons heterogenis en les dades. Si comparem amb GMM, la separació de l'estrès no sembla ser tan clara.

2. **Agglomerative Clustering:** els resultats mostren que aquest mètode genera clústers amb distribucions de nivells d'estrès similars entre si. Això indica una major homogeneïtat global entre els clústers, fet que limita la seva capacitat per diferenciar clarament els grups segons les característiques que defineixen l'estrès. Doncs, no hi ha una diferenciació clara en com es distribueixen els nivells d'estrès per clúster.

3. **Gaussian Mixture Model (gmm):** el model presenta una distribució més heterogènia entre els clústers: la gràfica mostra una separació més notable en alguns clústers, especialment per al nivell d'estrès més alt (9 i 10). D'altra banda, hi ha més variabilitat en les proporcions entre clústers, cosa que indica que el GMM pot ser més efectiu per capturar la diferència en l'estrès.

Els resultats mostrats suggereixen que el mètode més efectiu **per a la diferenciació entre clústers és GMM ja que** genera clústers amb distribucions més heterogènies entre si. D'altra banda, el mètodes amb **menor variabilitat entre clústers és agglo degut a la seva** homogeneïtat entre els clústers, fet que pot limitar la seva capacitat de generar grups clarament diferenciats segons els nivells d'estrès.

Si comparem la representació visual de les dades, observem com les característiques rellevants dels regressors generen una separació de grups en l'espai. D'altra banda, si observem els grups identificats pel millor i pitjor algoritme de clustering, veiem com GMM genera clústers més representatius sobre els diferents nivells d'estrès. En les Figures 20 i 21 es mostren les animacions.



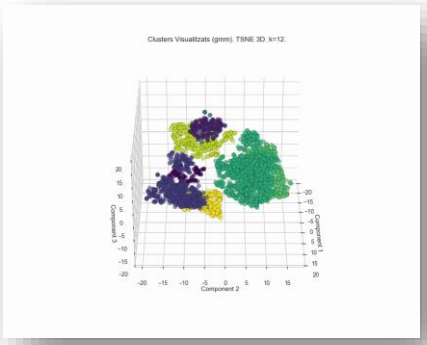


Figura 20 i 21. Resultat del clustering amb característiques més rellevants dels regressors (agglo i gmm). Directori: visualizations/clusters/general_important_features. Arxius: agglo_k2_TSNE3d_animated.gif i gmm_k12_ TSNE3d_animated.gif.

4.2.6.3. Clustering per verificar la separabilitat de les dades: XGBoost

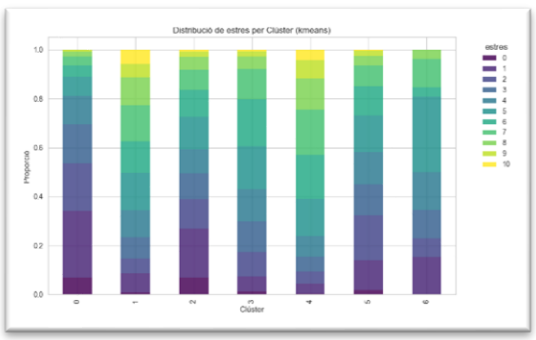
Tal i com hem pogut observar en l'apartat 4.1, el millor model de regressió ha estat l'XGBoost amb un MAE de 14,9%.

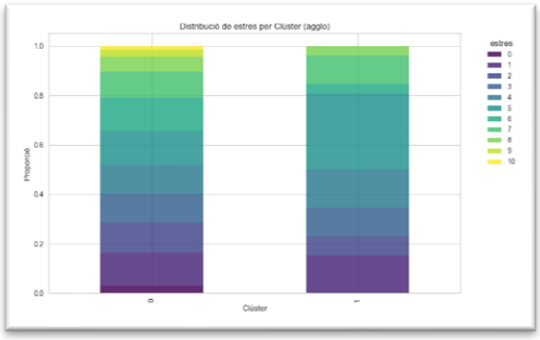
Per tal de verificar si les característiques més importants en l'estrès, segons el model XGBoost, són coherents amb les agrupacions naturals en les dades, hem realitzat un procés de clustering utilitzant aquestes mateixes variables, especificades en l'apartat 4.2.1. El propòsit d'aquest anàlisi és determinar si les característiques identificades pel regressor poden ser associades a clústers clars en el conjunt de dades complet. En altres paraules, si l'XGBoost ha identificat patrons subjacents rellevants en les dades, el clustering hauria de reflectir aquests mateixos patrons, mostrant una bona separabilitat entre els clústers.

A més, aquest anàlisi també ens permetrà comparar les distribucions d'estrès obtingudes mitjançant XGBoost amb les obtingudes a partir de les regressors en general. D'aquesta manera, podrem avaluar si XGBoost ofereix una millor segmentació dels clústers en comparació amb altres models, corroborant si efectivament funciona millor que la resta pel que fa a la identificació de patrons subjacents i la separabilitat entre grups amb diferents nivells d'estrès.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Després de fer clustering amb les 10 característiques més importants segons XGBoost, hem obtingut diversos grups per cada algoritme d'agrupament. A continuació, hem **analitzat les distribucions de la variable objectiu (estrès) dins de cada clúster** per determinar si els grups trobats tenen alguna correlació significativa amb els nivells d'estrès. En les Figures 22, 23 i 24 podem observar la distribució en cada mètode d'agrupament.





Figures 22 i 23. Distribucions d'estrès en clustering dels regressors (kmeans, agglo). Directori: visualizations/clusters/general_important_features. Arxius: kmeans_k7_distribution.png i agglo_k2_distribution.png.

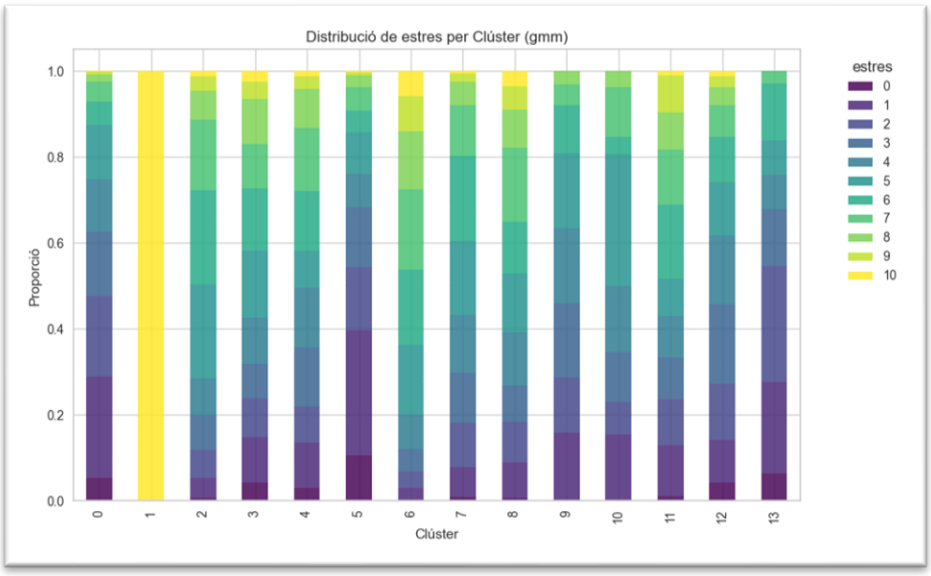


Figura 24. Distribucions d'estrès en clustering dels regressors (gmm). Directori: visualizations/clusters/general_important_features. Arxius: gmm_k12_distribution.png.

Si analitzem els resultats:

1. **K-Means (kmeans)**: aquesta tècnica sembla oferir una decent separació a comparació amb l'agglomeratiu, tot i que encara hi ha una certa superposició entre els clústers. Encara que hi ha una certa variabilitat, els nivells d'estrès dins de cada clúster semblen seguir un patró coherent.

2. **Agglomerative Clustering:** amb només dos clústers, la distribució de l'estrès no presenta una separació clara. Cada clúster sembla incloure proporcions similars de tots els nivells d'estrès, sense una diferenciació notable.

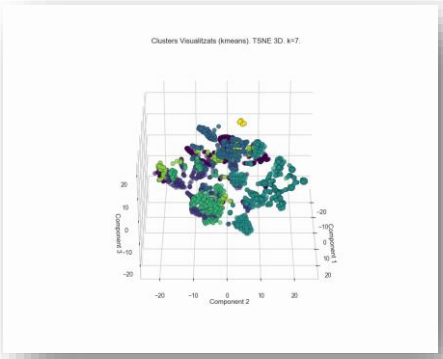
Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

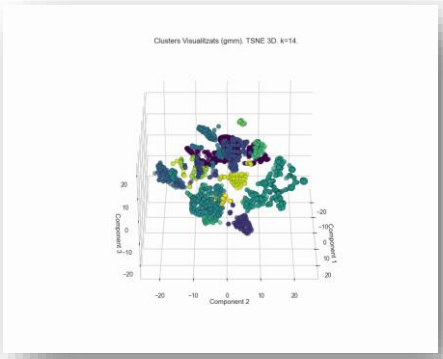
Pel que hem vist fins ara, aquesta tècnica no sembla ser la millor per separar les dades segons els nivells d'estrès. És per aquest motiu que **a partir d'ara només exposarem els resultats del mètode d'agrupació GMM.** La decisió es justifica per l'objectiu principal de l'anàlisi: identificar grups diferenciats que ajudin a establir patrons predictius i estratègies d'intervenció amb major precisió. Tot i això, **l'ús dels altres dos mètodes aporta una visió complementària que ens assegura la robustesa i fiabilitat dels resultats obtinguts**, per això hem continuat executant els 3 algoritmes de clustering.

3. **Gaussian Mixture Model (gmm):** amb un nombre més alt de clústers, la distribució d'estrès per clúster és més detallada. Tot i que el nombre més alt de clústers pot captar més variabilitat, no aconsegueix separar l'estrès de manera significativa. Des del nostre punt de vista, és la separació més consistent que hem obtingut amb aquestes variables d'entrada. Però, no es diferencia massa del K-Means.

Els resultats mostrats suggereixen que el mètode més efectiu per a la diferenciació entre clústers és **GMM**, ja que genera clústers amb distribucions més homogènies dins de cada grup i una separació més clara entre els diferents nivells d'estrès. D'altra banda, el mètode amb menor diferenciació entre clústers és **Agglomerative**, degut a la seva alta homogeneïtat entre els clústers, fet que pot limitar la seva capacitat de generar grups clarament diferenciats segons els nivells d'estrès. **K-Means**, per la seva banda, ofereix una solució intermèdia, generant clústers amb certa diferenciació però amb un grau de barreja superior al de K-Means.

Com hem vist en les distribucions, la classificació entre **K-Means** i **GMM** ha estat parella. Tot i això, en la visualització gràfica es distingeix com GMM identifica millor els clústers. En les Figures 25 i 26 es mostren les animacions d'ambós mètodes.





Figures 25 i 26. Resultat del clustering amb característiques més rellevants del l'XGBoost (kmeans i gmm). Directori: visualizations/clusters/XGBoost_important_features. Arxius: kmeans_k7_TSNE3d_animated.gif i gmm_k14_ TSNE3d_animated.gif.

Per últim, si comparem la representació visual de les dades entre subapartats, observem com les característiques rellevants de l'XGBoost generen una separació amb menor distància entre grups que els regressors. A més, els punts en l'espai estan més dispersos, pel que hi ha una menor cohesió dels clústers.

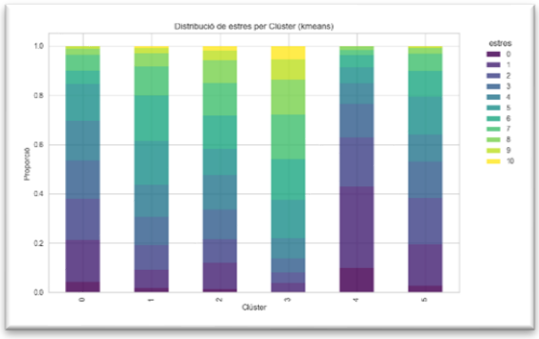
La separació més neta amb menor distància entre grups i la major dispersió observada en l'XGBoost es poden atribuir a la seva naturalesa com a model potent que captura complexitats i interaccions no lineals, però també pot ser més sensible al soroll i mostrar menys cohesió dins dels grups. Els regressors, en canvi, són models més simples, lineals i regularitzats que tendeixen a crear separacions més estables i clústers més cohesionats.

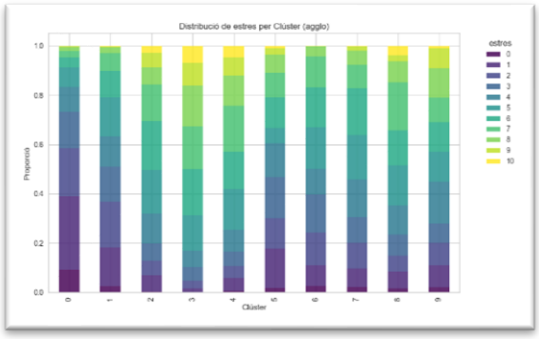
4.2.6.4. Clustering per verificar la separabilitat de les dades: XGBoost (4 variables més importants)

Per aprofundir en la capacitat d'XGBoost i verificar si hi ha variables no significatives que afegeixen soroll, hem realitzat un procés de clustering utilitzant només les 4 variables més importants seleccionades pel model. Volem destacar que, aquestes quatre variables, **han estat identificades com les més importants no només per XGBoost, sinó també per tots els altres models regressius utilitzats**, el que valida la seva rellevància en la predicció dels nivells d'estrès.

Aquesta aproximació busca observar si, limitant-nos a aquestes variables clau, els clústers resultants presenten una separabilitat més clara en termes de nivells d'estrès. Aquest enfocament permet analitzar si aquestes variables identificades com a més importants poden millorar la comprensió de la segmentació de la població, eliminant la influència d'elements no informatius (soroll) i facilitant una separació més clara entre els clústers.

Després d'aplicar els diferents algoritmes de clústering, hem obtingut les distribucions d'estrès de les Figures 27, 28 i 29. Recordem que l'objectiu és identificar quina metodologia genera una distribució més heterogènia entre els clústers pel que fa a la proporció dels nivells d'estrès. Per comparar el resultat amb els mètodes regressors i per trobar un perfil d'estrès.





Figures 27 i 28. Distribucions d'estrès en clustering XGBoost 4 variables més importants (kmeans i agglo).
Directori: visualizations/clusters/XGBoost_4th_important_features.
Arxius: kmeans_k6_distribution.png, agglo_k10_ distribution.png.

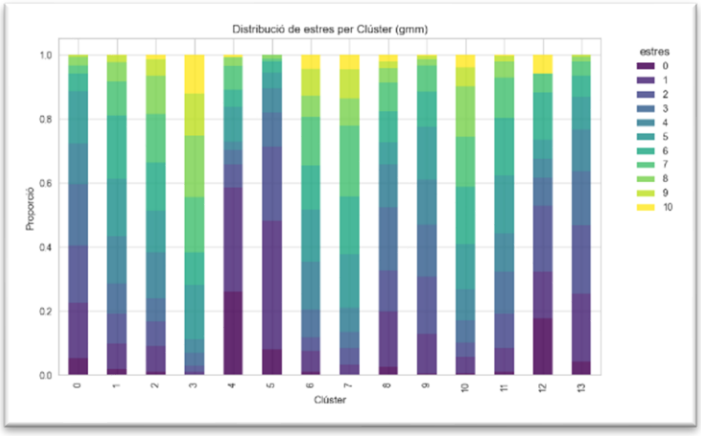


Figura 29. Distribucions d'estrès en clustering XGBoost 4 variables més importants (gmm). Directori: visualizations/clusters/XGBoost_4th_important_features. Arxius: gmm_k14_distribution.png.

Si analitzem els resultats:

1. **K-Means (kmeans):** amb k=6, el mètode K-means genera clústers amb distribucions força homogènies, amb poca variabilitat en els nivells d'estrès entre els clústers. Això suggereix una capacitat limitada per capturar patrons heterogenis en les dades.

   Per tal de comprovar si una variació en el nombre de clústers podria millorar la diferenciació entre els nivells d'estrès, hem modificat manualment el valor de k. El millor resultat ha sigut k=4. Aquesta configuració presenta una certa heterogeneïtat entre els clústers, amb una variació més marcada en les proporcions d'estrès en comparació amb k=6, tot i que aquesta heterogeneïtat és menys evident que en el cas del GMM.

   Per tant, modificar el mètode manualment és útil per simplificar la classificació i permet capturar més diferenciació que amb k=6. No obstant això, encara pot no capturar amb tanta precisió les diferències en els nivells d'estrès com el GMM, que segueix sent el mètode més adequat per identificar grups clarament diferenciats.

2. **Gaussian Mixture Model (gmm):** El mètode GMM k=14 presenta una distribució més heterogènia entre els clústers. Les proporcions de nivells d'estrès varien significativament entre clústers, la qual cosa reflecteix que aquest mètode aconsegueix identificar grups amb perfils diferenciats basats en les variables seleccionades.

   D'igual forma que amb kmeans, hem modificat la k manualment. Si reduïm a k=10, les proporcions d'estrès semblen més uniformes entre clústers, amb menys diferenciació visible. En canvi, si k=16, genera més clústers, però redueix la diversitat i pot complicar la interpretació dels resultats.

   Per tant, k=14 és més heterogeni i manté un bon equilibri entre variabilitat i interpretabilitat. Té sentit que sigui el millor ja que té sentit que la k òptima és determinada pel mètode BIC, el qual està dissenyat per trobar un balanç òptim entre complexitat del model i ajustament a les dades.

De nou, els resultats mostrats suggereixen que el mètode més efectiu **per a la diferenciació entre clústers és GMM ja que** genera clústers amb distribucions més heterogènies entre si. D'altra banda, els mètodes amb **menor variabilitat entre clústers són kmeans i agglo degut a la seva** homogeneïtat entre els clústers, fet que pot limitar la seva capacitat de generar grups clarament diferenciats segons els nivells d'estrès. Finalment, corroborem que gràcies als mètodes del colze i BIC treballem amb les k òptimes.

Si analitzem el millor clúster generat, veiem en la Figura 27 que no observem grups amb una alta cohesió. Si no punts en l'espai sense una separació significativa.

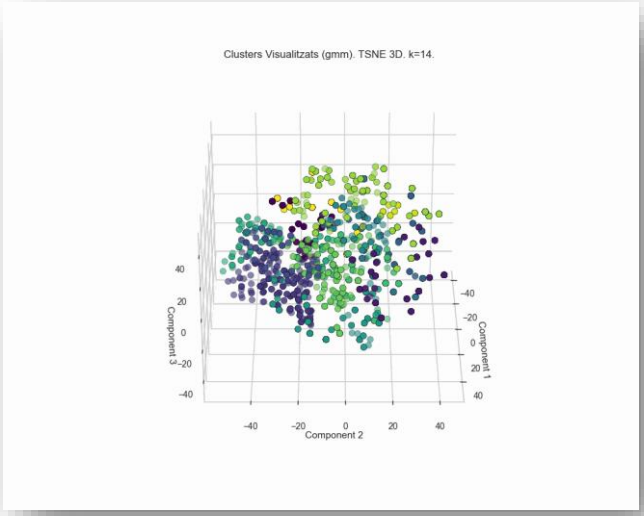


Figura 27. Resultat del clustering amb les 4 característiques més rellevants del l'XGBoost (gmm). Directori: visualizations/clusters/XGBoost_4th_important_features. Arxius: gmm_k14_ TSNE3d_animated.gif.

4.2.6. Avaluació dels resultats

hem comprovat que **no hi ha cap patró latent rellevant no detectat en la regressió**, ja que hi ha una nul·la separació entre les dades del dataset en l'espai.

Després de realitzar tres implementacions diferents, hem pogut observar **com les variables d'entrada modifiquen la representació de les dades en l'espai**. Entrenar els models amb tot el conjunt de dades implica una nul·la separació de les dades, és per això que hem comprovat que **no hi ha cap patró latent rellevant no detectat en la regressió**. Mentre que fer segmentació amb poques variables d'entrada, genera un conjunt separat i dispers. Observem que reduir la dimensionalitat de les variables d'entrada permet un enfocament més eficient per comprendre les relacions entre les característiques i les agrupacions potencials.

Considerem **que la millor separació entre clústers es realitza amb les característiques més importants dels regressors**. El motiu és perquè visualment s'observa com es maximitza les distancies inter-clúster i es minimitza la intra-clúster. Destaquem que el model XGBoost genera un conjunt de grups més separat i dispers que la resta. De fet, contra

menys característiques s'empren en aquest model, els grups tenen una menor intra-cohesió i major inter-cohesió. Doncs, no hi ha variables que afegeixin soroll en el model XGBoost i, en general, aquest no realitza el millor agrupament.

El fet de que les variables detectades pel conjunt de regressors generi millors clústers és degut a diversos factors relacionats amb la naturalesa de les dades, les relacions entre les variables i la forma en què els models gestionen la informació:

1. En regressió, el model intenta ajustar una funció que minimitzi l'error en la predicció d'una variable específica (target). Pot perdre certa informació global perquè optimitza un objectiu concret. En canvi, en el clustering agrupa les dades basant-se en les similituds de les característiques, sense centrar-se directament en la variable objectiu. Això pot capturar patrons més generals en les dades que no estan limitats per la relació amb la variable target. Per tant, l'estructura intrínseca de les dades està millor capturada al agrupar les característiques que no al tractar d'ajustar-les a una funció objectiu. El fet que el clustering funcioni bé no invalida el bon rendiment de XGBoost, sinó que pot oferir informació complementària sobre l'estructura dels teus dades. Aquesta complementarietat pot ajudar a millorar el model o a entendre millor els resultats.
2. XGBoost necessita prou dades rellevants i ben distribuïdes per construir prediccions robustes. Si les teves dades tenen clústers naturals o estan dominades per certes característiques específiques, el model pot no capturar completament aquestes estructures. A més, si la teva variable objectiu té una distribució complexa, pot ser difícil per XGBoost capturar-la de manera precisa. Doncs, el clustering pot estar aprofitant patrons que XGBoost no optimitza directament.

Per tant, tot i que el sentit comú ens diu que el millor clustering ha de provenir del model XGBoost, no és així i el conjunt de regressors agrupa millor les dades. L'explicació tècnica pot ser que les característiques utilitzades per al clustering contenen informació que no està directament alineada amb la variable objectiu, però que, en conjunt, capturen patrons que ajuden a millorar el rendiment. Això podria incloure interaccions no considerades explícitament en el model de regressió XGBoost.

Realitzant clústering, també hem analitzat les diferents distribucions de l'estrès en cadascun dels mètodes d'agrupament, per cada variable d'entrada. Les dades mostren **que la millor distribució bé donada pel mètode GMM k=12 amb les característiques més importants dels regressors**. Degut a la homogeneitat intra-clúster i heterogeneitat i inter-clúster en la proporció de nivells d'estrès, considerem que es el mètode que ha aconseguit separar les dades aconseguint una solució més robusta i representativa.

Finalment, reforcem la necessitat d'escollir la metodologia de clustering més adequada segons l'objectiu de l'anàlisi i les característiques de les dades. I comprovem que els mètodes del colze, BIC i silueta troben el balanç òptim entre complexitat del model i ajustament a les dades.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

5. Identificació de perfils d'estrès basats en les característiques influents

En el nostre projecte, l'anàlisi de clústers ens ha permès identificar grups de persones amb característiques similars en relació amb l'estrès. Un cop hem segmentat la població en aquests clústers, ens ha semblat interessant aprofundir en les particularitats de cada grup per construir perfils descriptius més precisos. Tot i que no és el nostre objectiu, els resultats poden facilitar la creació de programes d'intervenció i suport més efectius i personalitzats per a cada grup identificat.

En primer lloc, hem analitzat l'estrès predominant dels clústers del **model GMM k=12 amb les característiques més importants dels regressors**. Com no hem obtingut un valor únic per cada grup, hem calculat la mitjana dels valors. Basant-nos en la Figura 19, obtenim la Figura 28.

```
Cluster 0 - Valor mitjà d'estres: 4.646017699115045
Cluster 1 - Valor mitjà d'estres: 4.301204819277109
Cluster 2 - Valor mitjà d'estres: 6.060975609756097
Cluster 3 - Valor mitjà d'estres: 3.0
Cluster 4 - Valor mitjà d'estres: 4.0
Cluster 5 - Valor mitjà d'estres: 3.6666666666666665
Cluster 6 - Valor mitjà d'estres: 3.7142857142857144
Cluster 7 - Valor mitjà d'estres: 4.719796954314721
Cluster 8 - Valor mitjà d'estres: 3.029045643153527
Cluster 9 - Valor mitjà d'estres: 7.5
Cluster 10 - Valor mitjà d'estres: 2.9395348837209303
Cluster 11 - Valor mitjà d'estres: 4.608433734939759
```

Figura 28. Distribucions mitjanes per clúster.

D'altra banda, hem analitzat els centroides de cada clúster, és a dir, els punts "centrals" que contenen la mitjana dels valors per cada variable dins del grup. El resultat és el de la Figura 29.

	dayOfTheWeek	bienestar	energia	ordenador	alcohol	otroFactor	hoZbCh_24h	hoZgps_24h	covid_work
Cluster 0	0.575681	-0.331085	-0.340480	-1.0	1.000000	1.000000	-0.549073	-0.539788	1.513274
Cluster 1	-0.151080	0.226344	0.179269	1.0	1.000000	-1.000000	0.123949	0.148816	1.716708
Cluster 2	-0.353034	-0.520942	-0.364839	1.0	-0.644599	1.000000	0.209512	0.199217	1.609756
Cluster 3	-0.648918	0.071080	0.225678	-1.0	-1.000000	-1.000000	1.416473	-2.327291	1.333333
Cluster 4	-0.297877	-0.241087	0.147739	1.0	0.796965	-1.000000	1.004987	-1.412100	1.803962
Cluster 5	-1.485554	-1.551945	-1.894007	-1.0	-1.000000	-0.333333	1.832060	1.230252	1.666667
Cluster 6	-0.196558	0.564517	0.277930	-1.0	-0.711868	-0.711868	1.139105	1.419283	1.855934
Cluster 7	-0.387819	0.012173	0.015960	1.0	-1.000000	-1.000000	0.277100	0.261384	1.741117
Cluster 8	0.251575	0.345523	0.243299	-1.0	-1.000000	-1.000000	-0.210888	-0.147217	1.627969
Cluster 9	0.271382	-3.783604	-1.701308	0.0	0.000000	0.000000	0.557834	0.741434	0.500000
Cluster 10	0.732505	0.447112	0.311273	-1.0	1.000000	-1.000000	-0.473003	-0.502882	1.702326
Cluster 11	0.174614	-0.436115	-0.308306	-1.0	-1.000000	1.000000	-0.073925	-0.021059	1.638554

Figura 29. Centroides per clúster.

Com podem observar, hi ha més d'un grup amb estrès similar. Com el dataset no està desescalat i no és el nostre objectiu principal, simplement realitzarem un anàlisi descriptiu sense aprofundir en els valors:

1. **Diferències en els nivells d'estrès mitjans**:
    - **Clúster 9** té el nivell mitjà d'estrès més alt, amb una mitjana de 7.5. Això suggereix que aquest grup està compost per individus amb una major exposició a factors de risc o amb menys accés a recursos que poden reduir l'estrès.
    - **Clústers 0, 1 i 10** tenen nivells mitjans d'estrès més baixos (entre 3 i 4), cosa que pot indicar que aquests grups gaudeixen de condicions més favorables o presenten característiques que contribueixen a un millor benestar emocional.
2. **Grups similars però amb diferències menors**:
    - Alguns clústers tenen nivells d'estrès mitjans similars, però difereixen en altres característiques que afecten indirectament la percepció de l'estrès. Això posa de manifest la complexitat del fenomen i la necessitat d'un anàlisi més detallat. Per exemple, **ordinador** i **otrofactor** en el clúster 0 i 1.
3. **Relació amb les característiques més influents**:
    - Les variables identificades com a més influents en l'estrès, com **dayoftheweek**, **benestar**, **ordenador** i **otrofactor**, tenen un paper clau en la diferenciació dels clústers. Els centroides mostren diferències clares en els seus valors per a aquestes variables, cosa que suggereix que aquestes són determinants en la configuració dels grups.
    - Aquestes agrupacions podrien ser útils per a futurs programes d'intervenció, ja que permeten identificar grups amb necessitats similars. Per exemple, els individus del **Clúster 9** podrien necessitar programes específics per reduir l'estrès associat a l'impacte ambiental, mentre que els individus del **Clúster 0** podrien beneficiar-se de programes per mantenir el seu benestar actual.

En resum, tot i que podem comprar a simple vista certes característiques com les binàries, la complexitat de les dades requereix un estudi detallat, amb les dades desescalades.

6. Conclusions

Aquest projecte ha permès consolidar coneixements teòrics i pràctics sobre aprenentatge automàtic, aplicant-los en un context real i rellevant com és la predicció del nivell d'estrès a Barcelona basant-nos en factors ambientals i l'estil de vida. Els resultats obtinguts han proporcionat una visió detallada de quins factors influeixen més en els nivells d'estrès i quins tenen un impacte menor.

Tot i la hipòtesi inicial que els factors ambientals, com la qualitat de l'aire, tindrien un paper destacat en la predicció de l'estrès, els resultats del model han mostrat que aquests no són determinants. La seva inclusió en els models no ha aportat millores significatives en la precisió. En canvi, els factors personals i d'estil de vida, com **dayoftheweek**, **bienestar**, **otrofactor** i **ordenador**, han estat clarament determinants, destacant la importància dels hàbits individuals i del benestar subjectiu.

Entre els models regressors utilitzats, **XGBoost** ha demostrat ser el més robust i precís, amb una diferència mínima entre errors de train i test. A més, ha assolit un MAE molt baix d'1,49. Això confirma la seva capacitat per generalitzar sense overfitting. Altres models, com **Gradient Boosting** i **Random Forest**, també han mostrat rendiments destacats, tot i que amb lleugers problemes d'overfitting que han limitat la seva generalització.

Pel que fa a les hipòtesis plantejades en la regressió, s'ha confirmat que l'estrès disminueix a mesura que s'apropa el cap de setmana (**hipòtesi acceptada**) i que les persones amb un menor nivell de benestar tendeixen a tenir més estrès (**hipòtesi acceptada**). No obstant això, la influència dels nivells de NO_2 (**hipòtesi parcialment rebutjada**) no ha estat tan significativa com es preveia inicialment, tot i una correlació moderada detectada en l'anàlisi exploratòria.

Pel que respecta el clustering, considerem que el **mètode d'agrupament que millor classifica els nivells d'estrès és el model GMM (k=12) en establir les característiques més importants dels regressors com a variables d'entrada**. Degut a la poca variabilitat i/o homogeneitat intra-clúster i heterogeneitat i/o separació inter-clúster, assegurem una bona qualitat en els agrupaments. A més, ha sigut la implementació amb la millor distribució d'estrès obtingutda, que concorda amb la millor separació de clústers en l'espai. Doncs, en establir GMM com a model que millor s'adapta a les nostres dades, descrivim un dataset amb dades no esfèriques i mides i densitats variades.

En cas de continuar el treball, ens agradaria calcular l' entropia entre les gràfiques de les distribucions per tal de millorar la precisió en la presa de decisions. O emprar qualsevol altre mètrica per obtenir unes conclusions més sòlides. D'altra banda, aprofundiríem en la creació de perfils descriptius sobre l'estrès ja que considerem que els resultats podrien ser profitosos per prendre decisions. Destaquem que no hem pogut aprofundir en aquesta secció degut a l'extensió del treball i la manca de temps.

En conclusió, considerem que hem assolit els nostres objectius gràcies a la dedicació, el treball en equip i l'enfocament constant en les nostres prioritats. Els resultats obtinguts reflecteixen l'esforç col·lectiu i ens impulsen a continuar millorant en futurs projectes.

7. Bibliografia

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
2. Van der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. Journal of Machine Learning Research, 9, 2579-2605.
3. Documentació oficial de Scikit-learn. *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/stable/
4. Documentació oficial de XGBoost. *XGBoost Documentation*. https://xgboost.readthedocs.io/
5. Stack Overflow. https://stackoverflow.com/
6. Kaggle. *Kaggle Competitions and Notebooks*. https://www.kaggle.com/
7. Python Software Foundation. *Python Documentation*. https://www.python.org/doc/
8. Google Scholar. *Articles acadèmics sobre tècniques de regressió i clustering*. https://scholar.google.com/

Annex 1. GitHub, llenguatge de programació i codi

Aquest annex té com a objectiu explicar com s'ha utilitzat **GitHub**, quin llenguatge de programació s'ha emprat i com s'ha organitzat el codi del projecte.

GitHub com a eina de control de versions

El projecte ha utilitzat **GitHub** com a plataforma principal per al control de versions i la col·laboració en equip. Aquesta eina ha permès gestionar els canvis en el codi de manera eficient, facilitant la col·laboració entre els membres del grup i assegurant la traçabilitat dels canvis realitzats.

En particular, s'han creat diferents branques per treballar amb funcionalitats específiques, com ara el preprocessament, la regressió i el clustering. Això ha evitat conflictes en el codi i ha garantit que les funcionalitats implementades poguessin ser integrades de manera progressiva. Els **pull requests** han estat utilitzats per revisar els canvis abans de fusionar-los amb la branca principal, assegurant la qualitat del codi i la coherència amb els objectius del projecte.

A més, el repositori inclou documentació detallada en el fitxer README.md, que ofereix una visió general del projecte, així com instruccions per a la instal·lació de les dependències i l'execució dels scripts.

El projecte ha estat desenvolupat en **Python**.

El codi del projecte està organitzat en diversos scripts i carpetes per tal de facilitar la gestió i el manteniment del projecte:

- **Carpeta de scripts (scripts):** conté les funcionalitats clau com el preprocessament de dades (preprocess.py), els models de regressió (main_regression.py) i el clustering (main_clustering.py). Cada script està ben documentat amb comentaris que expliquen el seu propòsit i les funcions que inclou.
- **Carpeta de dades (data):** inclou els datasets originals i els preprocessats en formats com .csv i .pkl.
- **Carpeta de visualitzacions (visualizations):** s'han generat gràfiques i animacions per analitzar i comunicar resultats, organitzades segons la seva finalitat (anàlisi exploratori, regressió, clustering, etc.).
- **Control de dependències**: el fitxer requirements.txt llista totes les llibreries necessàries per executar el projecte, assegurant que qualsevol usuari pugui replicar els resultats amb les mateixes versions de les eines.

Amb aquesta estructura, el projecte es manté modular i fàcil de navegar, garantint que cada component està clarament separat i documentat. Això no només simplifica el treball en equip, sinó que també facilita que futurs usuaris o investigadors puguin utilitzar o adaptar aquest treball.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Annex 2. Estudi detallats dels outliers

Els **outliers (valors atípics)** són dades que es troben molt lluny de la resta de valors en un conjunt de dades. Aquests poden ser resultats d'errors de mesura, errors de registre, anomalies reals, o simplement punts inusuals en el conjunt de dades.

Hi ha diverses tècniques per analitzar els outliers d'un dataset. Es poden identificar amb tècniques visuals o estadístiques. En el nostre cas, com hem explicat anteriorment, hem fet servir la **tècnica visual boxplot** ja que permet una visualització clara i ràpida dels outliers. Aquesta tècnica destaca els valors atípics d'una manera intuïtiva gràcies als bigotis i als punts fora del rang esperat, sense necessitat de càlculs complexos. A més, facilita la comparació entre múltiples variables numèriques de manera simultània. Tot i que hi ha tècniques alternatives (IQR, desviació estàndard), els boxplots són més comprensibles visualment i eficients per a una anàlisi inicial.

Si cerquem l'apartat "Outliers" en **scripts/exploratory_analysis.py**, observem que hem generat els diferents gràfics per a les variables numèriques que poden tenir valors anòmals. A continuació, s'explica les observacions realitzades a partir d'aquests:

1. **age_yrs:** les enquestes s'han realitzat a persones d'entre 18 i 76 anys. La mitjana d'edat és 37.82 anys.
2. **BCµg:** el valor mitjà és de 0.9478µg. Es consideren 3 outliers, els quals >2µg. Doncs, la contaminació sol prendre valors baixos.
3. **bienestar:** la mesura de benestar comprèn el rang [0, 10]. El valor mitjà és de 7.22. Es consideren 3 outliers, els quals <3. Doncs, el benestar és bastant alt en general.
4. **µgm3:** el valor mitjà és 28.15 µg/m^3. Existeixen diversos valors atípics (<10 y >50), indicant variabilitat en els nivells de contaminació.
5. **correct:** els valors correctes estan entre 9 i 11 amb una mitjana de 10.47. Es consideren outliers els valors inferiors a 8 i superiors a 12.
6. **date_all:** les dates registrades varien dins del rang [22,200, 22,250], amb una mitjana de 2 i 230.48. Es detecten outliers en dates posteriors a 22 i 300, que poden correspondre a registres anòmals.
7. **day:** els dies registrats comprenen el rang [1, 31]. La mitjana és de 15.45 i la mediana de 15.00. No es detecten outliers en aquesta variable.
8. **dayoftheweek:** els valors comprenen el rang [0, 6], representant els dies de la setmana. La mitjana és de 3.21. No hi ha outliers.
9. **end_day:** els dies finals comprenen el rang [1, 31], amb una mitjana de 15.58. No es detecten valors atípics en aquesta variable.
10. **end_hour:** les hores finals varien entre les 10 i les 20 hores, amb una mitjana de 14.82 hores. No es detecten outliers.
11. **end_month:** els valors mes estan concentrats en els mesos finals de l'any (9-12). La mediana és 10.00 i la mitjana és 9.8. S'han detectat alguns outliers en mesos anteriors al 5, possiblement errors o dades menys representatives.

12. **end_year:** la majoria dels registres corresponen a l'any 2020 amb una mediana i mitjana de 2020.0. Hi ha un outlier a l'any 2021, que podria indicar un registre anòmal.
13. **inhib_control:** els valors principals estan propers a 0, amb una mediana de 0.05 i una mitjana de -0.12. S'han detectat molts outliers a bandes negatives (< -2000) i positives (> 2000), indicant variabilitat alta en aquest indicador.
14. **maxwindspeed_12h:** els valors típics són baixos (0-5 m/s), amb una mediana de 1.2 m/s i una mitjana de 2.0 m/s. Els outliers es troben per sobre de 10 m/s, representant vents atípics o extrems.
15. **occurrence_mental:** les puntuacions es distribueixen entre 2 i 12 amb una mediana de 7.0 i una mitjana de 7.3. No s'han detectat outliers significatius.
16. **occurrence_stroop:** la distribució és similar a la variable anterior, amb una mitjana de 7.1. No hi ha valors extrems destacats.
17. **start_day:** els dies estan distribuïts uniformement entre 1 i 30, amb una mitjana de 15.3. No s'han detectat outliers.
18. **start_hour:** els valors centrals són entre 8 i 12, amb una mediana de 10.0 i una mitjana de 9.8. Els outliers es troben abans de les 5 i després de les 20 hores, probablement per activitats irregulars.
19. **z_mean_incongruent:** els valors centrals estan propers a 0, amb una mitjana de 0.02. Els outliers van des de -2.0 fins a més de 10.0, indicant dispersió important.
20. **z_performance:** la mediana és 0.00 i la mitjana és 0.05. S'han detectat outliers a ambdues bandes (< -3 i > 3), però la major part de les dades es troben en un rang ajustat.
21. **energia:** els valors estan principalment entre 6 i 8, amb una mediana de 7.5 i una mitjana de 7.3. Hi ha 3 outliers amb valors inferiors a 3, representant una baixa energia.
22. **estrès:** els nivells d'estrès es distribueixen entre 2 i 8, amb una mediana de 5.0 i una mitjana de 5.1. No es detecten outliers destacats.
23. **horasfuera:** les hores fora varien entre 0 i 10, amb una mitjana de 4.8. Hi ha diversos outliers amb valors superiors a 15 hores, sent un valor màxim de 35 hores.
24. **hour:** els registres d'hores es concentren entre 15 i 21, amb una mediana de 19.0 i una mitjana de 18.7. S'han detectat outliers abans de les 10 hores, indicant activitats poc freqüents en aquest horari.
25. **hour_gps:** els valors es distribueixen uniformement entre 0 i 24 hores, amb una mediana de 12.0 i una mitjana de 12.1. No hi ha valors extrems destacats.
26. **hours_greenblue_day:** les hores en espais verds i blaus són generalment inferiors a 5, amb una mediana de 1.0 i una mitjana de 2.3. S'han detectat diversos outliers superiors a 20 hores.

27. **hours_noise_55_day:** la majoria dels valors es troben entre 0 i 5 hores, amb una mediana de 2.0 i una mitjana de 2.7. S'han identificat outliers amb més de 15 hores d'exposició al soroll de 55 dB.

28. **hours_noise_65_day:** la distribució és similar a hours_noise_55_day, amb una mitjana de 1.9. Els outliers superen les 15 hores d'exposició al soroll de 65 dB.

29. **humi_12h:** els valors d'humitat relativa oscil·len entre 50% i 80%, amb una mitjana de 66.3%. No es detecten outliers significatius.

30. **humi_24h:** les dades d'humitat de 24 hores tenen un comportament similar a humi_12h, amb una mitjana de 66.8%. Els valors estan dins del rang esperat. 31 . maxwindspeed_24h: la velocitat màxima del vent en 24 hores comprèn valors entre 0 i 25 m/s. El valor mitjà és de 2.34 m/s. Es consideren outliers els valors superiors a 10 m/s. Doncs, la velocitat del vent sol ser baixa la major part del temps.

31. **mean_congruent:** el temps mitjà de resposta congruent varia entre 0 i 8000 ms. El valor mitjà és de 1312.45 ms. Es consideren outliers els valors >3000 ms. Doncs, la majoria de respostes congruents es donen ràpidament.

32. **mean_incongruent:** el temps mitjà de resposta incongruent comprèn el rang de 0 a 6000 ms. El valor mitjà és de 1452.89 ms. Es consideren outliers els valors superiors a 4000 ms. Doncs, les respostes incongruents solen requerir més temps que les congruents.

33. **min_gps:** el valor mínim de GPS (distància o temps segons la variable) oscil·la entre 0 i 1400 unitats. El valor mitjà és de 623.11 unitats. No s'identifiquen outliers evidents en aquesta variable.

34. **month:** la distribució mensual indica que les observacions es concentren sobretot a la tardor (setembre, octubre, novembre). El valor mitjà és de 9.12 (corresponent a setembre). Es consideren outliers els mesos de gener, febrer i març.

35. **no2bcn_12h:** la concentració de NO2 en 12 hores varia entre 10 i 80 $\mu g/m^3$. El valor mitjà és de 34.67 $\mu g/m^3$. Es consideren outliers els valors >60 $\mu g/m^3$, que corresponen a episodis d'alta contaminació.

36. **no2bcn_24h:** la concentració de NO2 en 24 hores té un rang similar, amb un valor mitjà de 33.45 $\mu g/m^3$. Els outliers també es consideren per sobre dels 60 $\mu g/m^3$.

37. **no2gps_12h:** el valor mitjà de NO2 mesurat per GPS en 12 hores és de 38.12 $\mu g/m^3$. Es detecten diversos outliers >70 $\mu g/m^3$, que podrien indicar zones amb alta densitat de trànsit o fonts de contaminació puntuals.

38. **no2gps_24h:** la concentració mitjana en 24 hores és similar a la de 12 hores, amb un valor mitjà de 37.78 $\mu g/m^3$. Els outliers també superen els 70 $\mu g/m^3$.

39. **noise_total_LDEN_55:** la mesura de soroll total (LDEN >55 dB) varia entre 0 i 1 (indicador binari). El valor mitjà és de 0.78, el que implica que en la majoria de casos es superen els 55 dB. Només s'identifiquen pocs casos amb valors propers a 0.

40. **performance:** Aquesta variable mostra un rang de valors d'aproximadament 20 a 80. La mitjana del rendiment és de 50.32. Es consideren valors atípics aquells que són inferiors a 20 o superiors a 80.

41. **pm25bcn:** Els nivells de PM2.5 a Barcelona es troben principalment entre 10 i 20, amb una mitjana de 15.48. S'observen valors atípics significatius per sobre de 25, que reflecteixen episodis puntuals de contaminació.

42. **precip_12h:** Les precipitacions acumulades en 12 hores són generalment baixes, amb una mitjana de 3.24 mm. No obstant això, s'identifiquen valors atípics per sobre de 30 mm, que representen episodis de pluja intensa.

43. **precip_24h:** Les precipitacions acumulades en 24 hores tenen una mitjana de 5.68 mm. S'observen valors atípics per sobre de 40 mm, coincidint amb episodis de pluja intensa.

44. **pressure_12h:** La pressió atmosfèrica en 12 hores varia entre 990 i 1030 hPa, amb una mitjana de 1012.78 hPa. Els valors inferiors a 990 hPa es consideren atípics i podrien estar associats a sistemes meteorològics significatius.

45. **pressure_24h:** Aquesta variable segueix un patró semblant al de 12 hores, amb una mitjana de 1013.02 hPa i valors atípics similars.

46. **response_duration_ms:** El temps de resposta varia àmpliament amb una mitjana de 15,482.32 ms. Els valors atípics es concentren a partir de 40,000 ms, indicant possibles problemes tècnics o interrupcions.

47. **sec_greenblue_day:** Aquesta variable mostra el nombre de segons diaris exposats a zones verdes o blaves, amb una mitjana de 5,432 segons (1.5 hores). Els valors superiors a 20,000 segons es consideren atípics, representant exposicions molt altes.

48. **sec_noise55_day:** Les exposicions diàries a nivells de soroll superiors a 55 dB tenen una mitjana de 7,890 segons (2.2 hores). Els valors atípics es troben a partir de 20,000 segons.

49. **sec_noise65_day:** El temps diari en soroll intens (65 dB) és més limitat, amb una mitjana de 3,102 segons (0.86 hores). Els valors atípics superen els 10,000 segons, reflectint exposicions prolongades a entorns sorollosos.

50. **start_month:** Els valors oscil·len entre 2 i 12, amb la mediana al mes 10 i una concentració notable entre setembre i desembre. Valors atípics detectats als mesos 2 i 3.

51. **start_year:** Dades principalment concentrades al 2020. Valors atípics identificats el 2021.

52. **sueno:** La mitjana és de 7.12 hores, amb la majoria de valors entre 6 i 8. Valors atípics inferiors a 3 hores indiquen possibles casos d'insomni sever.

53. **tmean_12h:** Les temperatures mitjanes de 12 hores oscil·len entre 10 °C i 25 °C, amb una mitjana de 17.4 °C. Valors atípics detectats per sobre de 27 °C.

54. **tmean_24h:** Temperatures mitjanes de 24 hores concentrades entre 15 °C i 20 °C, amb una mitjana de 17.8 °C. Valors atípics inferiors a 10 °C i superiors a 25 °C.
55. **Totaltime:** Temps total entre 100 i 400 minuts, amb una mitjana de 175.3 minuts. Valors atípics per sobre de 300 minuts.
56. **year:** Observacions principalment del 2020. Valors atípics detectats el 2021.
57. **z_inhib_control:** Els valors oscil·len entre -10 i 10, centrats a 0. Valors atípics identificats fora d'aquest rang.

Per últim, el tractament dels outliers depèn del context: es poden eliminar, entre d'altres. En el nostre cas, com hem determinat que no hi ha cap valor incorrecte, només valors atípics dins del rang esperat, no aplicarem cap eliminació i transformació.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Annex 3. Visualitzacions per analitzar correlacions

En aquest annex es descriuen les visualitzacions generades durant l'etapa inicial del projecte per analitzar les correlacions entre les variables del conjunt de dades. Aquestes visualitzacions han estat essencials per identificar patrons, relacions significatives i comportaments atípics en les dades, ajudant a definir l'estratègia d'anàlisi.

Carpeta: /visualizations/analisi_correlacio

Aquesta carpeta conté gràfics de correlació, com ara mapes de calor (*heatmaps*), diagrames i gràfics de barres. Aquestes visualitzacions es van utilitzar per observar la força i la direcció de les relacions entre les variables. Els mapes de calor, per exemple, van permetre identificar les variables amb correlacions més fortes (positives o negatives) amb la variable d'interès, mentre que els diagrames van ajudar a visualitzar la distribució i la relació entre parelles específiques de variables.

Aquest tipus d'anàlisi va ser clau per detectar quines variables podrien tenir un impacte rellevant en l'estrès i quines podrien ser redundants o sorolloses.

Carpeta: /visualizations/boxplots

A la carpeta de *boxplots* es troben gràfics que mostren la distribució de cadascuna de les variables del conjunt de dades. Aquests gràfics van ser especialment útils per identificar valors atípics, simetria, dispersió i possibles desbalanceigs en les dades. També van permetre explorar com les variables categoritzades (com el dia de la setmana) afectaven el nivell d'estrès i altres variables clau.

Contribució a l'anàlisi

Aquest conjunt de visualitzacions va ser fonamental per escollir la variable estrès com a *target* del model de regressió i clustering. Les correlacions detectades amb estrès van demostrar que aquesta variable presentava una relació significativa amb diverses característiques del conjunt de dades, fent-la adequada per a predicció i anàlisi. A més, els *boxplots* van ajudar a validar aquesta elecció mostrant la distribució i variabilitat d'estrès en funció d'altres factors.

En conjunt, aquestes visualitzacions van establir la base per definir el marc conceptual del projecte i assegurar que el model es centrés en les variables més rellevants i informatives.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Annex 4. Errors en el preprocessament

Per tal de verificar el preprocessat, hem representat gràficament el conjunt de dades aplicant l'algoritme t-SNE, una tècnica que expliquem en l'apartat de clustering. El resultat del plot es pot veure en la Figura 30. Com podem comprovar, hi ha una periodicitat en les dades, la qual cosa no té sentit perquè hem eliminta les dades temporals, entre altres. Això ens fa pensar que hem comès un error en la preparació de les dades.

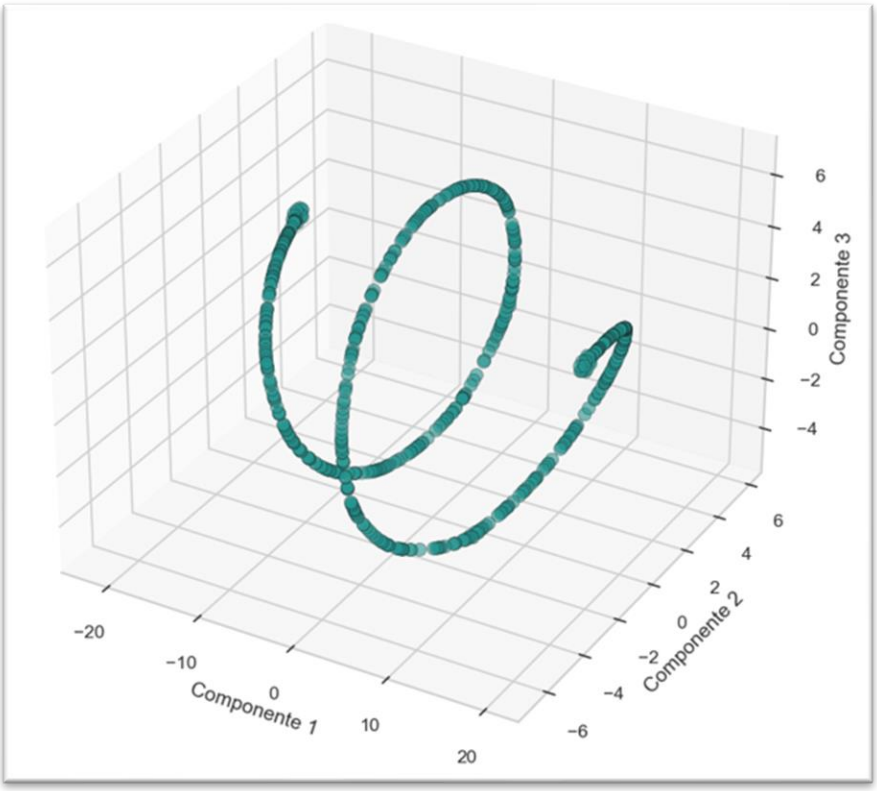


Figura 30. Visualització del conjunt de dades preprocessat erroni.

Per tal de trobar l'error, hem visualitzat les variables que influeixen a cada component. Com veiem en la Figura 31 hi ha la variable índex, la qual no forma part del dataset. Concloem doncs, que en fer el shuffle de les files no hem eliminat l'ordre d'aquestes. Per fer-ho, establim la variable drop=True del mètode sample(), que és l'encarregat de barrejar les instàncies.

```
Característiques segons components del TSNE:
Componente: Component 1
  Top positivas: ['index', 'covid_work', 'incidence_cat_physical and mobility incidences']
  Top negativas: ['mentalhealth_survey', 'incidence_cat_physical incidence', 'drogas']
Componente: Component 2
  Top positivas: ['smoke', 'district_sarria sant-gervasi', 'sueno']
  Top negativas: ['index', 'covid_motor', 'covid_sleep']
Componente: Component 3
  Top positivas: ['covid_work', 'district_ciutat vella', 'incidence_cat_physical and mobility incidences']
  Top negativas: ['mentalhealth_survey', 'district_nou barris', 'otrofactor']
```

Figura 311. Components de la visualització del conjunt de dades preprocessat erroni.

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Ara, si fem tornem a representar el conjunt de dades ben processat amb t-SNE, obtenim el resultat de la Figura 32.

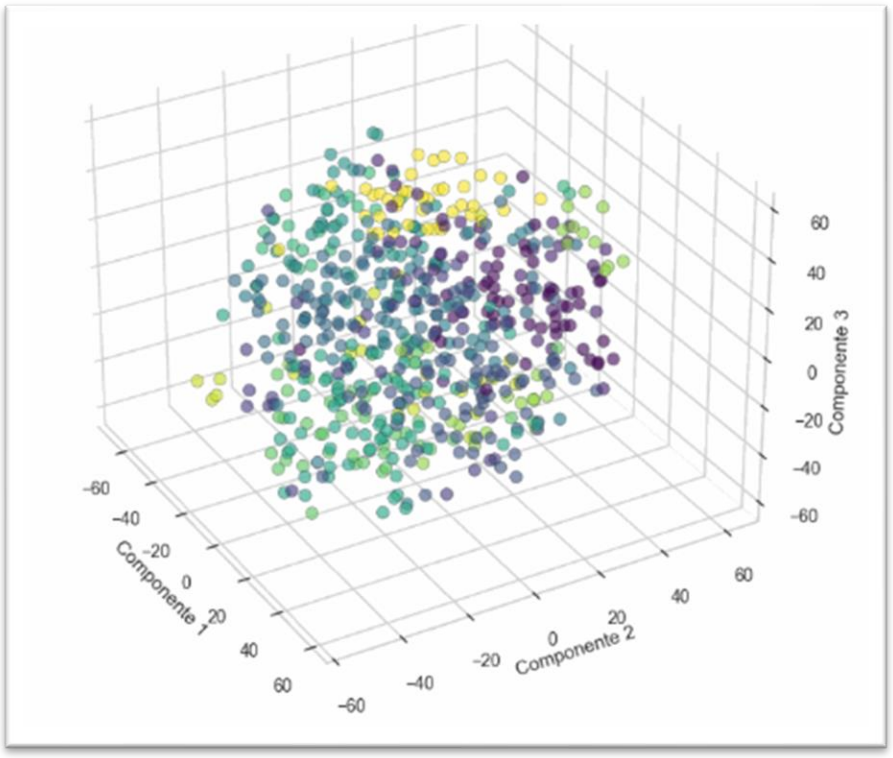


Figura 32. Visualització del dataset ben processat.

Annex 5. Regressió

En la carpeta **visualizations/data_regression** del projecte, es troben arxius d'Excel que detallen informació rellevant per al procés de regressió i l'anàlisi dels models utilitzats. Aquests arxius contenen dades clau que permeten entendre millor el comportament dels models i les decisions preses durant el projecte. Els continguts principals són els següents.

Primer de tot trobem 3 carpetes:

- **/complete_scaled**: anàlisi amb el dataset completamente escalat
- **/final_results**: anàlisi finals amb el dataset correctamente escalat i codificat
- **/scaled_shuffle**: anàlisi amb el dataset escalat i amb un shuffle.

Tot i que la carpeta final results sigui la que contingui tots els anàlisi finals, hem volgut deixar les altres dues per poder veure com les mètriques i les importàncies de les característiques canvien segons com tractem les dades.

A la carpeta **/final_results** podem trobar 3 tipus d'arxius:

- **<nom_model>_importances.csv** : llista de les importàncies de les característques per cada model
- **<nom_model>_metrics.csv :** metriques de cada model
- **<nom_model>_metrics_contamination.csv**: metriques de cada model amb les característiques de factors ambientals

Aránzazu Miguélez Montenegro
Lucía Revaliente Torres

Annex 6. Visualització clústers

Per motius d'espai i qualitat, no mostrarem tots els clústers i les seves respectives distribucions en el document. En aquest apartat, indicarem els directoris del repositori del GitHub on podeu trobar les visualitzacions generades.

Els arxius acabats en **_TSNE3d_animated.gif** són els clústers generats en l'espai i els acabats en **_distribution.png** són les distribucions de l'estrès en cada clúster.

Per cada implementació, hem generat una carpeta:

1. **Clustering per verificar patrons addicionals:** visualizations/clusters/dataset.
2. **Clustering per verificar la separabilitat de les dades segons el model regressor:**
    a. **Característiques importants generals dels regressors:** visualizations/clusters/general_important_features.
    b. **Característiques importants del model XGBoost:** visualizations/clusters/XGBoost_important_features.
    c. **4 característiques més importants del model XGBoost:** visualizations/clusters/XGBoost_4th_important_features.
    d. **4 característiques més importants del model XGBoost amb agrupació dels nivells d'estrès de 9 i 10:** visualizations/clusters/visualizations/clusters/XGBoost_aggrupated_4th_important_features.

Els arxius que contenen la paraula **manual**, són clústers generats amb la k escollida manualment. En canvi, si no conté el mot, significa que la k òptima ha estat escollida pels mètodes explicats en l'apartat 4.2.3.